BLIDA 1 UNIVERSITY

Faculty of Sciences

Computer Science Department

Master's thesis

In Computer Science

Option : Natural Language Processing

THEME :

Evolutionary Algorithm for the Study of the Food Pairing Hypothesis in the Algerian cuisine

Realised by

KERBEDJ Tarek

CHAHBOUB Racha

Supervised by

Dr. BACHA Siham

July 2022

Acknowledgements

We want to express our gratitude to Mrs. BACHA, our teacher, and supervisor, who helped us finish this thesis with promising results with her sage counsel, constant constructive criticism, and priceless patience. We also want to express our gratitude to our instructors who have helped us get to this point in this academic journey, especially Mrs. MEZZI who provided us with encouragements and writing materials.

We are also very appreciative of the moral support and encouragement from our families, especially our parents, who helped us stay motivated throughout the entire work period.

Abstract

Cooking forms the core of our cultural identity, and culinary practices have always had great variations from region to region, and in order to study the relationships underlying the structures of these different cuisines, Computational gastronomy emerged, which is an interdisciplinary field that studies food using Data science.

One of the main questions in this field is the food pairing hypothesis, which states that combined ingredients with common flavor compounds taste better than their counterpart, this hypothesis has been studied in Western, European, and middle eastern cuisines, However, there are no available studies conducted in the North African region [1].

In this study, we used genetic algorithms (GAs) to test this hypothesis in the Algerian cuisine, by applying different computational methods to traditional Algerian recipes extracted from the book "la cuisine algérienne" by Mrs. Bouayad, which have been subsequently preprocessed using different NLP techniques. Our research showed that the pattern of ingredients constituting Algerian recipes has a negative food pairing tendency and that result is consistent with the south European cuisine which indicates a trend in the Mediterranian region.

keywords: Computational gastronomy, Food pairing hypothesis, Data science, Evolutionary algorithms, Algerian cuisine.

Resumé

La cuisine constitue le cœur de notre identité culturelle, et les pratiques culinaires ont toujours connu de grandes variations d'une région à l'autre. Afin d'étudier les relations sous-jacentes aux structures de ces différentes cuisines, la science de gastronomie computationnelle a émergée. Un domaine interdisciplinaire qui étudie la nourriture en utilisant la science des données.

L'une des principales questions dans ce domaine est l'hypothèse de l'appariement des ingrédients, qui stipule que les ingrédients combinés avec des composés de saveur communes ont un meilleur goût que leur contrepartie. cette hypothèse a été étudiée dans les cuisines occidentales, européennes et du Moyen-Orient, Cependant, il n'y a pas d'études disponibles menées dans la région de l'Afrique du Nord.

Dans ce travail, nous avons utilisé une approche basée sur les algorithmes génétiques (AGs) pour tester l'hypothèse d'appariement d'ingrédients dans la cuisine algérienne. Nous avons créé un corpus de recettes traditionnelles algériennes authentiques extraites à partir du livre de la cuisine alégrienne de Mme Bouayad. En utilisant des techniques de traitement automatiques de la langue ces recettes ont été prétaitées. L'application des algorithmes génétiques sur ces recettes néttyées a montré que l'hypothèse d'appariement d'ingrédients dans la cuisine algérienne a une tendance négative. Ce résultat est cohérent avec la cuisine du sud de l'Europe qui indique une tendance dans la région méditerranéenne.

Mots clés: gastronomie computationnelle, hypothèse du food pairing, cuisine Algérienne, algorithmes évolutionnaires

Contents

List of Figures

List of Tables

Acronyms

CM-C Copy-Mutate Category only. 31

CM-M Copy-mutate Mixture. 31

CM-R Copy-Mutate Random. 31

EA Evolutionary Algorithm. 9, 34–37

EC Evolutionary Computation. 34

EP Evolutionary Programming. 36

ES Evolutionary Strategy. 36

FV Fitness Value. 49, 74

GA Genetic Algorithm. 7, 36–38, 42, 43

GP Genetic Programming. 36

IFW Ingredient Frequency Weighting. 10–13, 49, 74, 75, 77, 79–81, 83–89, 91, 92, 94–98

NLP Natural Language Processing. 21, 22, 58

NLTK Natural Language ToolKit. 8, 56, 58

re Regular Expression. 57

TSP Travelling Salesman Problem. 43

VCF Volatile compounds in Food. 20, 27

General Introduction

Global context

"To prepare a dish is to write a coherent story, and this story must then find an echo in its audience," said chemist Raphael Haumont, author of the book "Chemist in the Kitchen." This is true for all cultures because food is an identity factor that has been anchored since the beginning of humanity, and it represented first and foremost an element of survival, which then evolved as a result of various circumstances. A nation's cuisine is intimately connected to its cultural, social, and historical dimensions, as it reflects many of its characteristics: we can see traces of all its ancient civilisations, it tells us about the properties of its land and climate, and it transmits the values of its people through its culinary traditions. Throughout the centuries, cooking has been a trial-and-error learning process, but once the available elements were mastered, it became an art that opened the door to creativity. In recent years, scientists have joined gastronomes to add a rational dimension to understanding the reason behind certain unlikely but delicious combinations, or what causes different preferences and associations across the world's populations. Food pairing, a field that combines gastronomy, chemistry, and computer science, is based on Blumental's hypothesis: "the more characteristic aroma compounds that two foods have in common, the better they taste together." Since this proposal was formulated, scientists from all over the world have been interested in verifying its veracity. If it is confirmed, it is said that food pairing is positive and if it is refuted it is said to be negative.

Problematic

The food pairing hypothesis holds true in Western cuisine, where chefs have brilliantly exploited it, such as the combination of chocolate and blue cheese in desserts, which share 73 aromatic molecules, or the kiwitre dish developed by chef Sang-Hoon Degeimbre, which features kiwi and oysters that share 14 aromatic molecules. East Asian and Southern European cuisine, on the other hand, have negative food pairings and thus use ingredients that do not share many flavor molecules. So, how about Algerian food? Do Algerians prefer flavors with common chemical components, or is it the other way around? What factors have influenced Algerian culinary culture? Is the colonial influence significant, or is it more influenced by its geographical location?

Objectives and challenges

Our research work aims to determine the nature of the food pairing of Algerian cuisine, by compiling traditional and authentic recipes and then constructing the flavor profile of each of the ingredients present in the corpus with their aromatic molecules, we used a genetic algorithm to optimize the search in space of these generated recipe components, The second objective of this work is to extract the most commonly used ingredients characteristic of Algeria, using Natural Language Processing techniques and visualization tools. We faced several obstacles during our research due to the lack of documentation on Algerian culinary culture, specifically a corpus of representative authentic recipes, which led us to create it manually; this step was also challenging because our reference book was in a scanned PDF format, which caused format problems. The lack of molecular components in some ingredients posed another challenge in our process.

Thesis plan

In this dissertation, we will first introduce the key concepts underlying our research, food pairing and NLP, followed by state of the art. In the following chapter, we will go over the

fundamentals of evolutionary algorithms, with an emphasis on genetic algorithms. The approach is summarized in Chapter 3 and the steps to our solution are detailed in it. Finally, chapter 4 presents the implementation of our solution, beginning with all the procedures we used to collect our corpus of recipes and the preprocessing methods we have used, and then we visualise the characteristic ingredients of Algerian cuisine. Finally, in the second part, we expose all the experiments we performed for the generation of recipes and their results, and then we discussed the latter and their implications to reach a conclusion.

Chapter 1

Food Pairing

1.1 Introduction

To integrate the important concepts of this subject, this chapter outlines some of the key components of the field of computational gastronomy, namely Food Pairing, Flavor Data, Recipes, and Natural Language Processing, and finally it overviews the most impactful works in the state of the art.

1.2 Computational Gastronomy

For centuries, people have been preparing food, either by following a set of predefined recipes obtained through tradition, books or by using crude methods such as: guesswork, intuition, trial and error. However, scientists have recently taken an interest in this subject and discovered that, despite the vast number of ingredients, we have only explored a small percentage of all the possible combinations. As a result, they began by studying the chemical compounds of various ingredients, and then introduced computational models to better understand the inner workings, thus creating a new field of research called computational gastronomy [1]. There are various dimensions of computational gastronomy covering: food paring, flavor data and recipes.

1.2.1 Food Pairing

Nowadays, the food industry needs to satisfy consumer demand for new products on a regular basis, which makes the research and development department an indispensable component in the evolution of the industry, since they will be collecting relevant information on the sector and analyzing it in order to innovate on a regular basis, but the research and the process of bringing the product to fruition does not keep up with the pace demanded by this field, so it is necessary to remedy this issue [6]. In this quest for constant efficiency and innovation, molecular gastronomy has imposed itself with its chemical approach to cooking, with the goal of presenting the best version of products by improving taste and flavour through various technologies.

Food pairing is one of the innovations pioneered by molecular gastronomists, it the practice of finding associations between different ingredients based on their gustatory properties, furthermore, it allows us to study different cuisines and identify their characteristics. One of the most important hypotheses in this field is presented by chef Huston Blumenthal in his book "The Fat Duck Restaurant" [7] : "foods that share a lot of flavor compounds taste delicious together."

Even though this field is relatively new, it has already found use cases in practice, with laboratory discoveries soon to be applied in restaurants. It is true that the application of computational gastronomy, and more specifically food pairing, is most prevalent in European restaurants, which are frequently ranked among the top in several rankings, such as the three Michelin starred "fat duck" of chef blumenthal in London [8], or the three star Italian restaurant of chef Massimo Bottura "Osteria Francescana" [9][10], However, interest in other parts of the world is beginning to emerge, as evidenced by a study conducted in Istanbul to investigate the impact of incorporating molecular cuisine into their restaurants[11], the results of which demonstrate that consumers are attracted to this impressive service, and they eventually concluded that it significantly increased tourism and business figures. In Africa a study conducted in Kenya looked at the role of traditional cuisine in regional restaurants[12].

1.2.2 Flavor Data

Because the flavor humans experience when eating depends on taste and smell, where aboutt 80% is due to the aroma [13], the first step to begin food pairing is the analysis of the food and the ingredients to work with. It is important to present these two concepts as well:

- Volatile chemical compounds / aroma compounds: are what confer to aliments their taste and smell, they're mostly volatile, and they can be identified by a series of chemical processes. Starting with Gas Chromatography in order to separate and identify the different components of a complex compound. Then using Mass Spectrometry to know the exact structure of each component.

- Flavor profile: a collection of volatile chemical compounds that humans can detect and smell. The high concentration of aromas present in them makes them perceptible to the nose; the selection process is determined by taking every compound present in concentrations greater than a certain threshold.

It is critical to have a structured collection of the flavour profiles of the ingredients used in the study, as these are way to quantify the flavours in our food. There are several resources that provide databases to meet this requirement, the Volatile compounds in Food (VCF) database has 624 ingredients (522 food products and 105 food product categories) and 7645 unique flavor compounds, whereas the Fenaroli's handbook has 1530 ingredients and 1107 flavor compounds. We can see that the ingredients in Fenaroli are much higher than in VCF, however it has fewer flavor compounds. The main difference between these two databases is the list of flavor compounds provided for each ingredient; some ingredients have compounds that are listed in Fenaroli and present in VCF, but VCF has a greater number of flavor compounds that are not listed in Fenaroli[14]. Aside from these two databases, which only provide the chemical aromatic components of the ingredients, there are other compilations of ingredients and their flavour profile, as well as other properties of interest in a particular area, such as natural origin, nutritional factors, or medicinal value (such as foodDB, FlavorNet, NutriChem, or the most comprehensive one FlavorDB)[15].

1.2.3 Recipes

A recipe describes how to prepare a dish and typically consists of two parts: the ingredients and the preparation. The object of study for a data-driven study must be well chosen according to the theme and, above all, well structured, so that the analysis is more effective. In our case, the recipes must be carefully chosen in order to accurately reflect the authenticity and roots of Algerian cuisine.

Algerian cuisine is a practice that carries thousands of years of history with its recipes that transmit the heritage of all the civilizations that have occupied it at any given time (Amazighs, Arabs, Turks, Andalusians, Spaniards, French...), their influences can be felt with the small differences that can be seen from one region to another, such as fish with honey, which has been traced back to Tnes centuries before its appearance in Europe [16], but the uniqueness is also very striking with the inescapable national dishes such as couscous, whose reputation precedes it, since it was listed as a world heritage site by UNESCO [17]. Even if ancestral preparation techniques have been well preserved through the generations, the concern of a very traditional culinary culture such as ours is the oral transmission of these, which is not in favour of computational study of recipes. Even so, there are an increasing number of resources devoted to Algerian recipes, such as websites or books that collect traditional recipes.

Although the books address the above - mentioned issue, they are still in formats that are difficult to process by computers (paper format), and the recipes are frequently written in a narrative style, which contradicts the structured principle of the data sought, making it difficult to process only the desired information.

The use of Natural Language Processing (NLP) techniques is the solution to this problem. NLP is used to preprocess and prepare recipes and ingredients in order to give them a more structured layout that would then be used as input to predictive or statistical. Using NLP allows researchers to study the underlying chemical relationships between ingredients in different cuisines and their impact on the perceived flavor which is the aim of food pairing.

1.2.4 Natural Language Processing

Language is one of the most important yet complex cognitive tasks performed by humans, with over 7000 different languages around the world [18]. It allows for communication and transmission of information through different mediums such as speech, writing, and signs. The emergence of both computers and various means of telecommunication in the 20th century, have led to the increasing demand for automatic, efficient methods of processing the human language, thus giving birth to the field of NLP.

Natural language processing or NLP for short is a subfield of artificial intelligence, which can be broadly defined as the application of computational techniques to the analysis and synthesis of natural language as well as speech[19]. NLP pipeline is constitued of four major steps [20]:

1. **Data Collection**: the first step in any NLP task is to acquire the raw textual data
2. **Data wrangling**: this includes preprocessing the text and preparing it for the model, and that can consist of a multitude of processes, such as:
   (a) **Cleaning**: this includes the removal of non-alphanumerical characters, HTML tags, and punctuation.
   (b) **Tokenization**: which is splitting chunks of text, such as sentences, into atomic entities like words.
   (c) **Stop word removal**: this includes the removal of words that are grammatically correct but carry no significance, such as or, to.
   (d) **Stemming**: it is defined as reducing words to their stem or root by removing prefixes and suffixes. for e.g.:computers $\rightarrow$ comput
   (e) **Lemmatization**: it differs from stemming in that it applies the morphological analysis of the word,rathern than a heuristic to reduce it to its lemma. For e.g.: computers $\rightarrow$ computer
3. **Feature engineering**: this step is vital in any NLP task, and in short it implies transforming the textual data into a numerical format such as vectors, that can then be used by the model, there is a variety of methods that can achieve this, including:

(a) **bag of words**: it describes the occurrence of words within a document, with complete disregard to order.

(b) **TF IDF**: it's a measure that is similar to bag of words, except that it accounts for the importance /relevance of the terms in a collection of documents.

(c) **Word embeddings**: word2vec is a notable example. It represents words through a real-valued vector. This method differs from the techniques mentioned above, in that it can capture the semantics of a word, using the notion of proximity which means words that are closer in meaning are represented closer to each other in the vector space.

4. **Model training and testing**: after choosing a model suitable for the task, it would be trained and tested using the clean data obtained through the steps above and it would finally be deployed.

1.3 Related Works

Heston Blumenthal, the famous chef of the fat duck restaurant, was seeking a replacement for salt in white chocolate recipes in 1992 when he discovered that caviar combined with white chocolate was a divine combination. He decided to consult François Benzi, a scientist who works in a flavorings and perfumes company, to understand what was behind this union's success. They discovered that both chocolate and caviar contain high levels of amines (amines contribute to the flavors we like in cooked meat, cheese, and other foods), and the fact that they shared this protein was most likely why they tasted good together [21]. From this experience emerged the food pairing hypothesis, which stated that recipes tend to be composed of ingredients that share chemical compounds, attempting to prove or refute the hypothesis for many cuisines around the world. One of the most recent and influential studies [2] in this field focused on taking a topological approach to the problem, creating a bipartite flavor network that showed the chemical components shared by ingredients, in order to investigate the hypothesis and discover patterns that transcend specific dishes. To avoid a western interpretation of the world's cuisine, they used a massive dataset of 56498 recipes from American repositories and added a Korean repository to

them, and the average number of ingredients used in the recipes is around 8. The recipes were divided into 5 geographical categories (North American, Western European, Southern European, Latin American, East Asian) to avoid a western bias and have a detailed vision of the world's cuisine. Those geographical categories are presented in figure 1.1 part C. To make the dataset more appropriate for the research, they replaced ingredients like essential oils or extracts with the flavor compounds of the original ingredient, because they are physically extracted from them, and the second post-processing step they did was to include the flavor compounds of more general ingredients into a more specific one (the flavor compound in "meat" is included in "beef" and "pork" as well). To conduct this research, they began by constructing a bipartite network, illustrated in figure 1.1 part A, composed of two types of nodes:

1. The first type represents 381 ingredients that are used in recipes all over the world.

2. On the other side of the network, the second type represents 1021 flavor compounds known to contribute to the flavor of each ingredient, and each ingredient is linked to all of the flavor compounds that contribute to its taste.

They constructed a flavor network, which is a projection of the resulting bipartite network; the projection can be resumed as follows:

1. Two ingredients are linked if they share at least one flavor compound.

2. The weight of each link varies according to the number of flavor compounds shared by the ingredients.

3. Due to a lack of systematic data, they did not consider the impact of concentration in each ingredient.

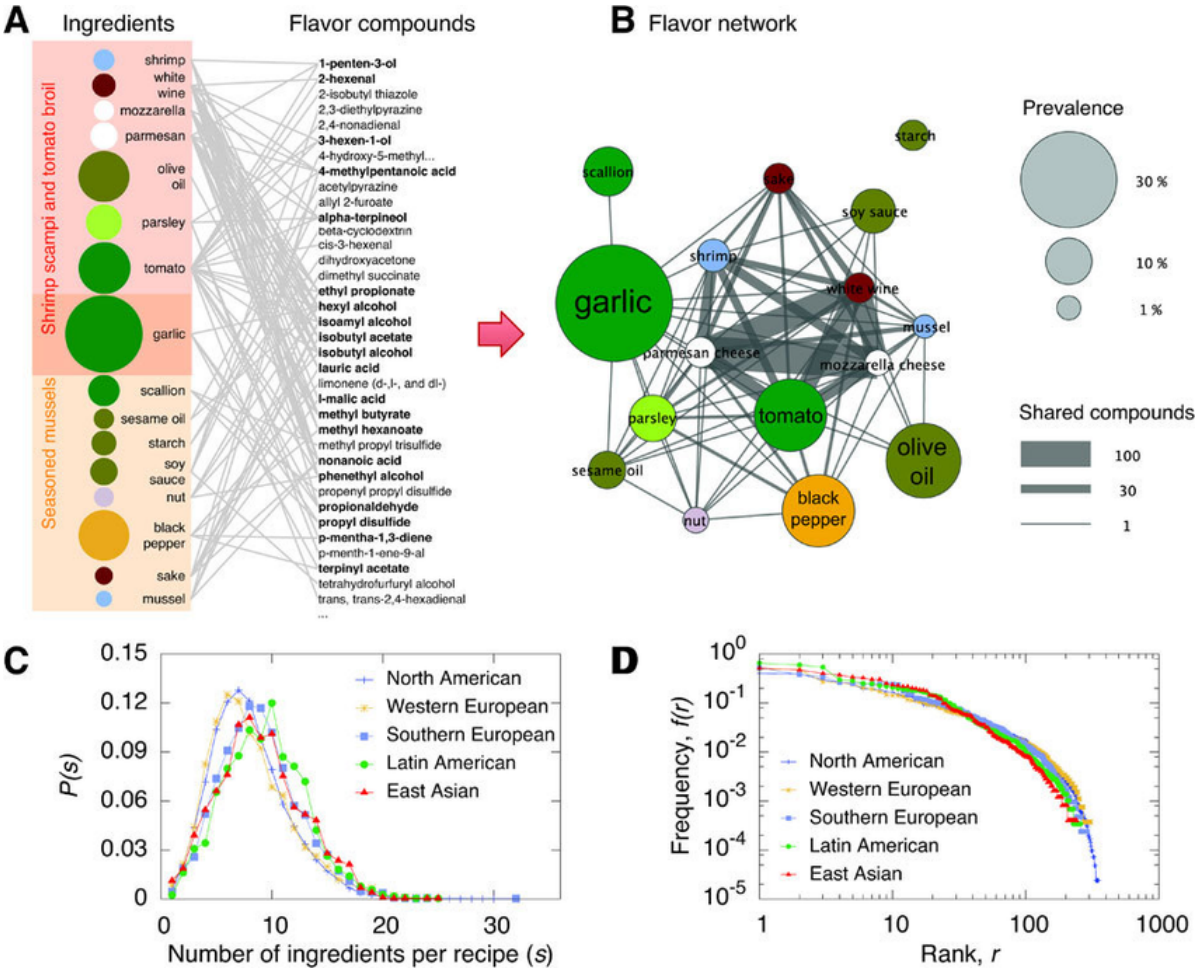


Figure 1.1: Flavor network. (A) bipartite network: The ingredients contained in two recipes (left column), together with the flavor compounds that are known to be present in the ingredients (right column). (B) flavor network: a projection of the bipartite network, nodes are ingredients, and the thickness of the links represents the number of flavor compounds two ingredients share and the size of each circle corresponds to the prevalence of the ingredients in recipes. (C) The distribution of recipe size, capturing the number of ingredients per recipe, across the five cuisines explored in our study. (D) The frequency-rank plot of ingredients across the five cuisines shows an approximately invariant distribution across cuisines [2]

The network is too dense to extract direct visualization from it, as shown in Figure 1.1 part A, so they only kept the statistically significant links for each ingredient by identifying them using a backbone extraction method, and for each node they kept the edges whose weight is statistically relevant given the strength of the node, resulting in a network where ingredients are grouped in categories.

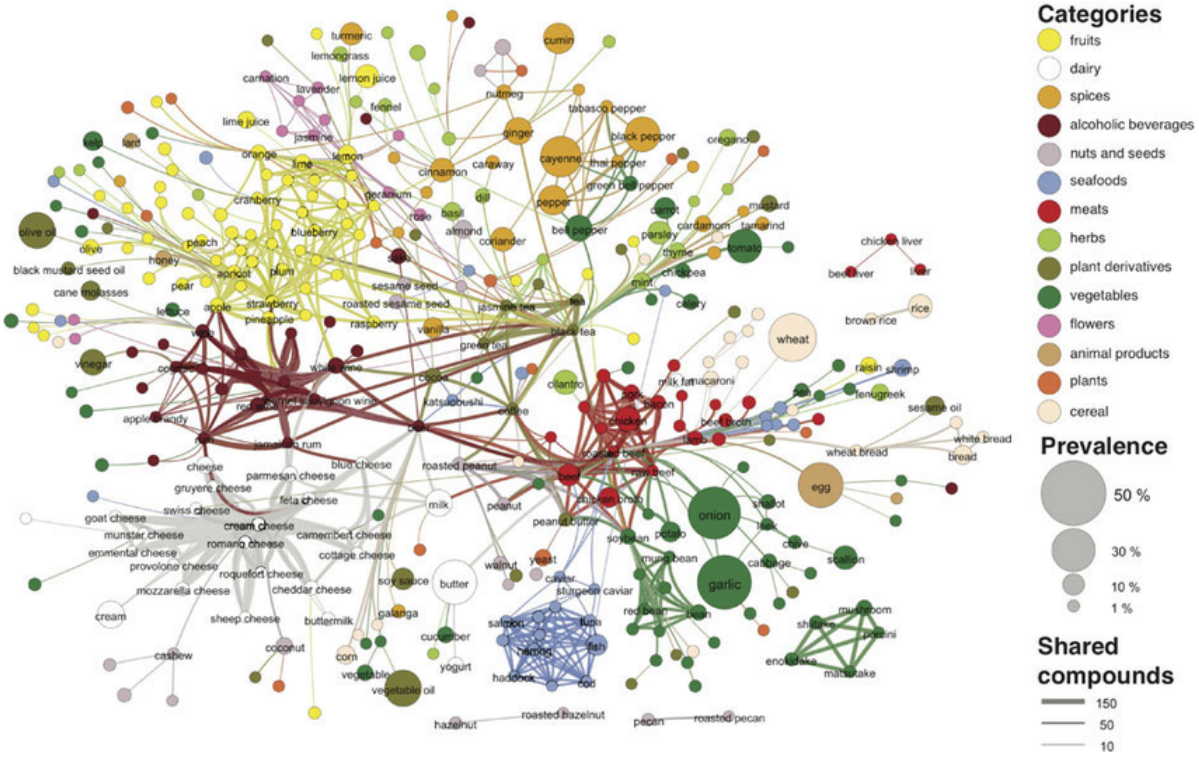


Figure 1.2: The backbone off flavor network[2]

This network representation allowed us to rephrase the hypothesis as a topological question: "Do we use ingredients that are strongly linked in the flavor network more frequently or avoid them?" [2]. This study resulted in two major conclusions. The first is that North American and western European cuisines tend toward recipes whose ingredients share a lot of flavor compounds, which confirms the starting hypothesis, while east Asian and southern European cuisines have negative tendencies with recipes that do not share many flavor compounds, which contradicts the starting hypothesis.Moreover, They concluded that ethnic and cultural aspects of a cuisine play a role in determining the ingredients of a recipe, rather than just the actual chemistry of ingredients.

A study on medieval European cuisine was conducted by Kush et al. [14] . To investigate its food pairing tendency, they created a dataset largely composed of medieval European recipes. They primarily compiled recipes from 1300 to 1615 from 25 text sources of England, Germany, France, and Italy. These recipes were used to extract the ingredients. To have a category of ingredients, they manually create a table of ingredients based on popularity, synonyms, and spelling variations. They used the following table to double-check the recipes' ingredients and discard any that were missing. They began with 4133 medieval recipes and ended up with 41

blank recipes after preprocessing manipulations. The goal of this research was to determine the impact of this difference in food pairing. After gathering medieval recipes on one hand and flavor compound datasets on the other, they proceeded to match the ingredients of recipes to the flavor compounds in each database, first using Fenaroli to match the strings simply, but they could only find 157 ingredients, leaving 229 unmatched (mainly because Fenaroli's handbook includes flavor compounds of more general ingredients into more specific ones like oils and their origins). The VCF was then manually matched because there were times when they needed to look for more generic categories, such as when medieval cuisine included different fish and they associated them with the same flavor profile. Based on this information, they conducted a statistical analysis, observing initially different values due to the VCF containing significantly more flavor compounds. Depending on the flavor compound database used, the research led to contradictory results. On the one hand, they obtained a very strong positive tendency for food pairing in medieval European cuisine using the Fenaroli handbook, on the other hand, they obtained a negative food pairing of this same cuisine using the VCF database, which can be translated as the ingredients used in recipes do not share many flavor compounds. To investigate the cause of this contradictory result, they examined the individual contribution of each ingredient and discovered that, according to each database of flavor compounds, the first contributors are different, for example, VCF is dominated by fish while Fenaroli does not contain many fish and has many gaps from this type, and this instability could be the cause of the conflict. Despite the contradictions, the authors concluded that in general, medieval European cuisine had positive pairing, and that it could have been stronger at the time due to a lack of available ingredients. This study raised new questions for the food pairing community, such as what other parameters play a role in determining the tendency of food pairing?

Another study [22] examined Indian cuisine, presenting a model that quantifies the food pairing pattern and identifies statistical features of this food. They collected 3330 authentic Indian recipes from books and online repositories, then preprocessed them by removing duplicates and ingredients that did not have flavor compounds and replacing some redundant components (canned pineapple = pineapple), additionally they deleted ingredients in the "snack" and "additive" categories, and finally they excluded the recipes that remained with only one

ingredient. After these steps they divided the 2543 remaining recipes into eight regional cuisines: Bengali, Gujarati, Jain, Maharashtrian, Mughlai, Punjabi, Rajasthani, and South Indian. They ended up with 192 ingredients in total, which belonged to 15 food categories (such as spice, vegetable, fruit, fish, animal product, and so on)

Because there are fewer works on Arab cuisine than on Western cuisine, a study was published in 2017 aiming to investigate whether food pairing in Saudi cuisine will prove positive or negative using genetic algorithms [1]. This study presents a data analysis and modeling approach to explore food through their chemical components. The first stage of their work was data collection, which consisted of compiling approximately 100 recipes from Saudi traditional cooking books, such as Rabha Hafzi's "the cooking principles of Saudi and Middle Eastern cuisine." A list of ingredients is gathered from each recipe, and for each ingredient, a list of flavor compounds is collected from the database Fenaroli's handbook of flavor compounds to define its flavor profile. Figure 1.3 illustrates how the data gathered for this study is organized.

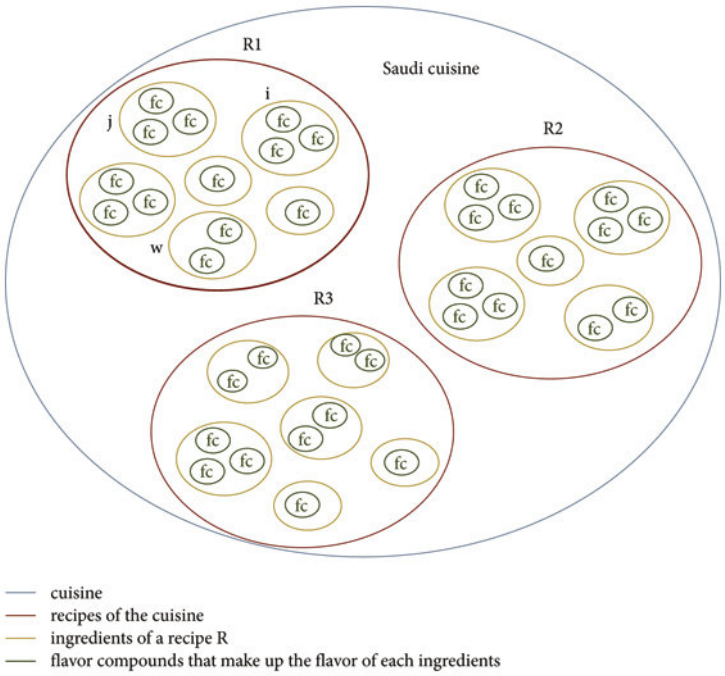


Figure 1.3: Relationship between cuisine, recipes, ingredients, and flavor compounds: each recipe R is made up of a number of ingredients, each ingredient is composed of many flavor compounds [1]

They began with data cleaning and preprocessing after collecting the recipes and removing duplicates, so they unified ingredients with different spellings (sliced tomato = diced tomato =

tomato), and then they built a list of all ingredients occurring in all recipes, and after obtaining the clean list, they translated it from Arabic to English. Once the English version of the ingredients was complete, they compared them to the components in the Fenaroli database, excluded ingredients that did not contain flavor compounds, and then removed recipes with only one ingredient. Finally, they highlighted the ingredients that are unique to Saudi cuisine and contribute to its identity. The principal focus of the study is data modeling, it includes mostly generating a random set of cuisine data using the "copy mutate" genetic algorithm.

Once the random recipes were obtained, they had to neutralize the effect of the copy mutate algorithm's unpredictability by generating 40 sets of 100 random copy mutate recipes, after which they could resume work by calculating the average flavor sharing.

Figure 1.4: Frequency of ingredients: the larger a circle is the higher is the frequency of the ingredient [1]

The visualization of the ingredients allowed to deduce the most used ingredients in Saudi cuisine: black pepper, onion, tomato, garlic, sunflower oil, and cumin, as can be observed in the figure 1.4. The authors expected to see results similar to those of the Indian study, which stated that the pairing of Indian cuisine is negative because both Saudi and Indian cuisine use spices, but the study's conclusion is that the food pairing in Saudi cuisine is positive, which means that they are more likely to use ingredients that share flavor compounds. According to them, this difference may be due to the fact that Saudis use a limited number of spices, whereas Indians have an exceedingly diverse spice culture.

In order to discover statistical patterns in ingredient use and categories in world cuisine,

as well as to investigate culinary evolution by testing algorithms a data-driven study [23] was conducted by . To complete these tasks, they compiled a list of 158544 recipes from around the world and annotated their origin (region, country, or continent). They had recipes from 25 geo-cultural regions, all of which were relatively well represented, with the largest collection containing 23179 recipes (Italy) and the smallest 470 recipes (Central America), they also manually classified each ingredient into 21 categories, such as vegetable, dairy, and meat, and linked them all to FlavorDB components to give them a flavor profile. They calculated a metric to assess ingredient overrepresentation, which quantifies the importance and presence of an ingredient in a region's cuisine in relation to its place in the world cuisine; in this way, they were able to deduce that fish is significantly used in the regions of East Asia and Thailand, in contrast to the rest of the regions; the same finding was made regarding basil in Italy. This metric allowed for the identification of country-specific ingredients as well as differences between regions. Concerning the study of culinary evolution, they imitated it by duplicating and incorporating changes to the ingredients of the initial recipes, and this by using the algorithm proposed by Kimachi et al.[24], which does not place any restrictions on the choice of ingredient replacement so it is a Copy-Mutate Random (CM-R), and then they experimented with variations of this algorithm with different configurations:

- **Copy-Mutate Category only (CM-C)**: selects the replacement ingredient from the same category as the ingredient being replaced.
- **Copy-mutate Mixture (CM-M)**: in which, when replacing, they all do so in the same way as the original algorithm, with the exception of one in the middle of the procedure where one ingredient is replaced by another of the same category.
- **Null Model**:in which no mutation is used.

The CM-C, CM-M, and CM-R algorithms reproduce recipes in accordance with the empirical rank-frequency with 20 ingredients in the initial pool and 4 mutations for CM-R and 6 mutations for CM-M and CM-C, whereas the null model did not reproduce the empirical results. As we can see from the state of the art, studies on Asian and European cuisines currently dominate this field; to address this gap in the literature, our work focuses on examining the hypothesis of food

pairing on traditional Algerian cuisine; for this purpose, we will use a traditional cookbook "La Cuisine Algérienne" written by Bouayed Fatima-Zohra, to gather recipes commonly used in our culture and extract their unique ingredients using natural language processing. In terms of flavor compounds, we will use the FlavorDB database to connect the ingredients with their molecular compounds in order to create a flavor profile. We will present a genetic approach to optimally explore the research of generated components and find the best solutions.

Works	Dataset of recipes	Regions	Results
Article by Ahn et al. [25]	56498	North American (41525)	Positive Food Pairing
		Southern European (4180)	Negative Food Pairing
		Latin American (2917)	Positive Food Pairing
		Western European (2659)	¨Positive Food Pairing
		East Asian (2512)	Negative Food Pairing
Article by Kush et al. [14]	4092	Medieval Europe	Positive Food Pairing
Article by Jain et al. [22]	2543	India	Negative Food Pairing
Article by Al-Razgan et al. [1]	100	Saudi Arabia	Positive Food Pairing

Table 1.1: Summary of studies and their results

1.4 Conclusion

In this chapter, we discussed the origins of food pairing and some important concepts in the subject, as well as the hypothesis that many researchers are interested in, along with recent and influential studies in the field, before briefly explaining our approach to tackling the hypothesis in question.

Chapter 2

Evolutionnary algorithms

2.1 Introduction

In this chapter, the objective is to introduce various theoretical concepts and techniques in relation to our work. We will begin by going through a brief introduction to Evolutionary algorithms with a specific emphasis on Genetic algorithms while exploring the different techniques used, and their applications.

2.2 History and Overview

EAs, are efficient heuristic search methods based on Darwinian evolution that capture global solutions to complex optimization problems with characteristics of robustness and flexibility [26], they are considered to be a subset of Evolutionary Computation (EC) which is a subset of Artificial Intelligence.

EAs have been around since the 1950s and started branching into different subfields a decade later, in fact, one of the earliest use cases of these algorithms was a computational simulation of Evolution by the Norwegian-Italian mathematician Nils Aall Barricelli in 1953 [27]. However, these strategies would only become popular in the 1960-1970s as a result of using them to solve complex engineering [28]. It is speculated that they will be increasingly used and further developed due to the improvements in computational power, more robust and better suited

open-source software libraries as well as the growth in demand for AI-solutions for both the industry and Academia.

2.3 Basic Structure

The goal is to evolve a fixed set of different individuals (solutions), as well as make sure that the fittest and most suitable candidates survive, this is done through a criterion called the fitness value , assigned by the fitness function, which represents the quality of the solution.

The mechanism of an EA is quite similar to that of natural selection, overall, EAs contain four major steps: Initialization, Selection, Genetic operators, and Termination. All EAs start with a pre-defined number of individuals called a population, and then these individuals would be evaluated by the fitness function, and then they would go through a selection process, followed by a combination of genetic operators such as mutation and crossover, this Algorithm will keep iterating until a stopping condition is met.

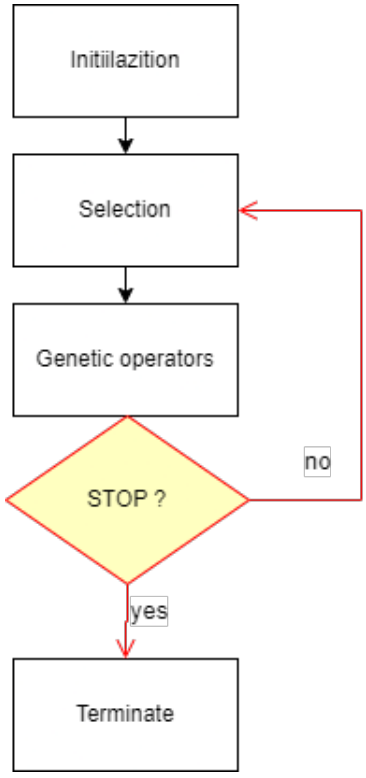


Figure 2.1: Basic structure of an evolutionary algorithm

2.4 Techniques

Despite the different techniques and variations in EAs, they all follow the same principles from Darwinian evolution and only differ in implementation details. They can be grouped into 4 major categories: Genetic Programming (GP), GA, Evolutionary Programming (EP), and Evolutionary Strategy (ES) as shown in figure 2.2.

- **Evolutionary Strategies**: ES are considered to be one of the oldest forms of EAs. They usually involve only mutation and selection, individuals are selected using truncation selection, which systematically removes individuals under a chosen threshold after sorting them based on their fitness [29].

- **Genetic Programming (GP)**: they differ from genetic algorithms in that GP manipulates programs, not strings, in other words, the solutions are in the form of computer programs, and their fitness is determined by their ability to solve a computational problem.

- **Evolutionary programming** : the unique feature that distinguishes EP from the rest of the techniques is the use of relative fitness instead of the raw fitness [30], as illustrated in Table 2.1.

Individual	Raw Fitness	Relative Fitness
A	7.15	1
B	12	4
C	8.9	3
D	4.5	0
E	11	3

Table 2.1: Illustration of relative fitness in EP

- **Genetic Algorithms**: Genetic Algorithms are the most widely used and popular among the EAs, they apply evolution to fixed-length strings. The length of the string represents the dimensionality of the problem [29]. GAs are used to solve complex problems in which the solution space is too large to be explored by traditional methods.

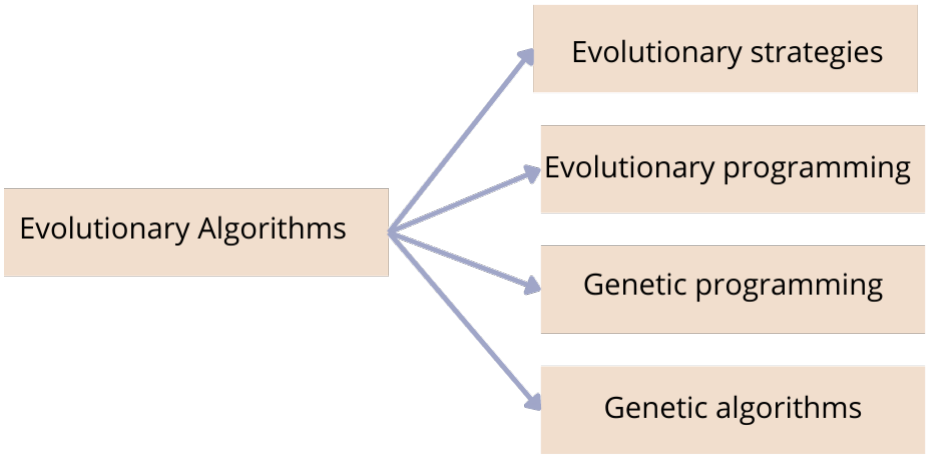


Figure 2.2: Different categories of EAs

2.5 Genetic Algorithms Overview and Structure

The main data structure used in GAs is called a chromosome. Each chromosome represents a potential solution, and within each chromosome, there are different genes that represent bits of information.

We can summarize the essential elements of GAs in five different sections [31]:

- **Encoding the solution**: proper encoding is vital for GAs to work as intended.
- **Population initialization**: a suitable initial population can lead to a rapid convergence of the algorithm.
- **The fitness function**: which assigns a fitness value to each chromosome.
- **Genetic operators**: they allow us to explore the wide space of solutions by diversifying the population.
- **GAs hyper-parameters**: this includes the initial population size, the mutation and crossover rate, and the number of generations or the stopping condition.

The diagram 2.3 illustrates the basic structure of a GA

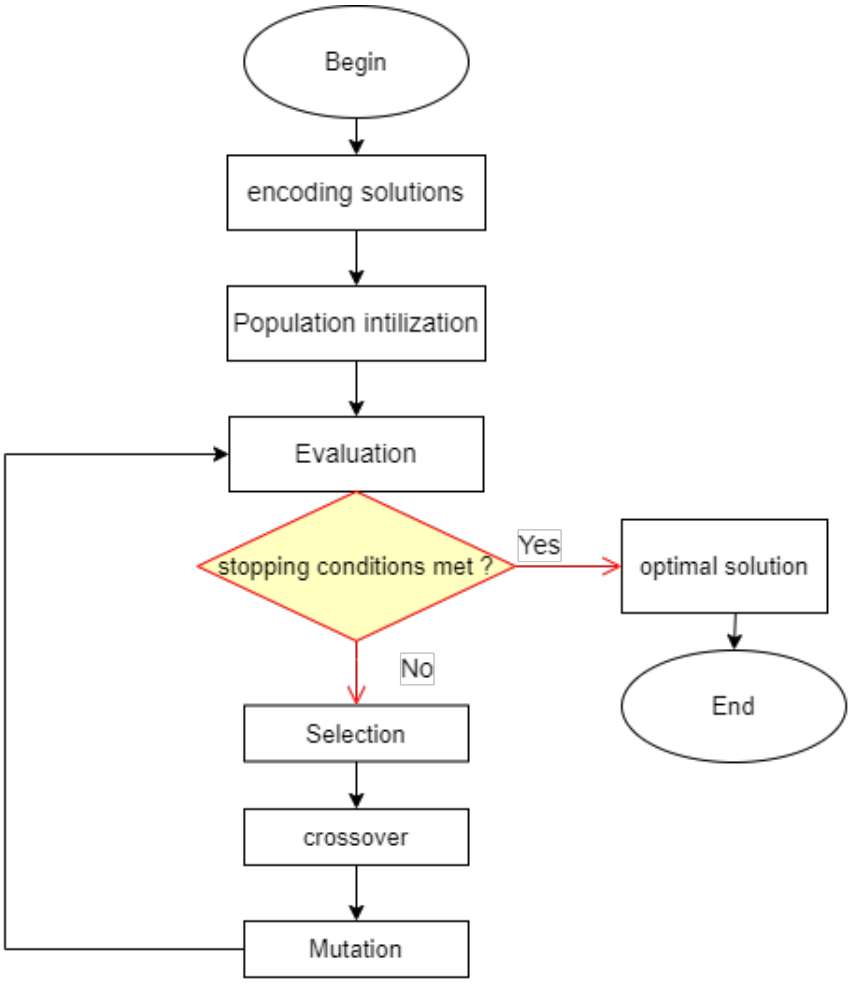


Figure 2.3: Basic structure of a Genetic algorithm

2.5.1 GA workflow

In the following we present the GA workflow:

2.5.1.1 Encoding

Encoding is the first step in the process since we need to represent the variables that we are trying to optimize in a manner that the algorithm can use. There are two types of encoding: *binary encoding* and *real encoding*.

- **Binary encoding**: this representation is the most commonly used, genes are encoded as n-length binary strings 0,1 and are represented in a pre-defined interval, eventually they would have to be converted from binary to a decimal representation in order to calculate the fitness value of each chromosome as shown in figure 2.4.

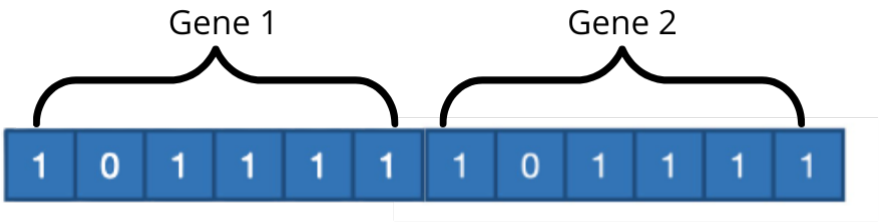


Figure 2.4: Illustration of binary encoding

- **Real encoding**: is simpler than binary representation, where each chromosome is a vector of real values. The evaluation of the fitness is faster in this format since we won't need to decode the binary value. An example of real enconding is given in figure 2.5.

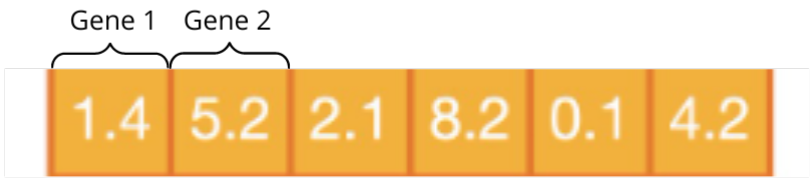


Figure 2.5: Illustration of real encoding

2.5.1.2 Initialization

This can be done in two different ways, either randomly if there is no known information about the global optimum or through an heuristic, if there is a priori knowledge about the problem.

The size of the initial population needs to be optimized to achieve a well-rounded compromise of the cost and the quality of the solution. A population that's too big would require a considerable amount of time and memory, but a population that's too small is most likely to be stuck in a

local optimum. The crossover probability is often chosen between [0.7, 0.99] and the mutation probability is chosen between [0.001,0.01] [32].

2.5.1.3 Fitness evaluation

Genetic algorithms are formulated in terms of maximization, given a real function f with one or multiple variables, and a search space A.

$$max_{x \in A} f(x) \tag{2.1}$$

Where each element of A is a candidate solution, but it can evidently be used in minimization problems through the following transformation: $F(x) = 1/(1 + f(x))$, and we would have this formula as a result.

$$min_{x \in A} g(x) \tag{2.2}$$

Now given these two formulas above, the fitness function would evaluate a solution x and outputs a real value, which would serve as a fitness value.

2.5.1.4 Selection

This step follows initialization, and it is used at the start of each new iteration to pick individuals from the existing population that will serve as parents for the individuals of the next generation. The selection is probability-based, and the probability of an individual being picked is linked to its fitness value, in a way that gives an advantage to individuals with higher fitness values [33]. There are multiple ways of selecting an individual, including but not limited to:

- **Roulette wheel selection**: the probability of selecting an individual is proportionate to its fitness value. This is comparable to using a roulette wheel in a casino. If N represents the cardinality of the population, then the probability of selecting an individual x_i, denoted $p(x_i)$ is equal to:

$$P(x_i) = \frac{F(x_i)}{\sum_{k=1}^{N} F(x_k)} \tag{2.3}$$

- **Rank-based selection**: Rank Selection is similar to Roulette wheel selection[34], the

difference between them is that Rank selection sorts the population first according to their fitness value and ranks them. Then, every chromosome is allocated selection probability with respect to its rank.

- **Tournament selection**: Tournament selection is a method of selecting an individual from a population of individuals. it involves running several tournaments among a few individuals chosen at random from the population., the larger the tournament size, the higher the chance that the best individual will participate in the tournament.

2.5.1.5 Genetic operators

These operators always follow the selection process and they include both *Crossover* and *Mutations*:

- **Crossover**: also known as recombination, is a genetic operator used to maintain genetic diversity it corresponds to the biological crossover that occurs during sexual reproduction and is used to combine the genetic information of two individuals who serve as parents to produce (typically two) offspring [33]. Figure 2.6 demonstrates crossover operation.

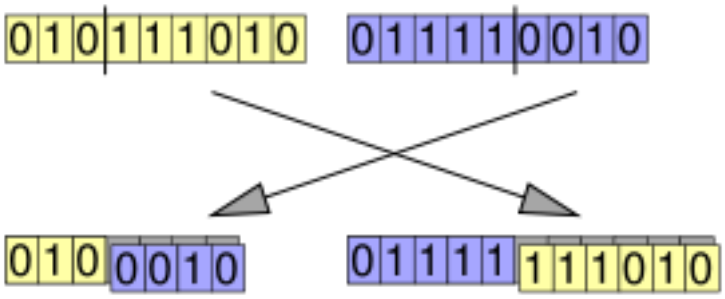


Figure 2.6: Crossover operation [3]

- **Mutation:** which is analogous to biological mutation, is the last operation to be applied to the offspring after the selection and crossover. It is used as an attempt to avoid local minima by preventing the population from becoming too similar to each other [35]. The most common mutation methods are :

  - **Flip bit mutation**: where one gene is randomly selected and its value is flipped as detailed in figure 2.7.

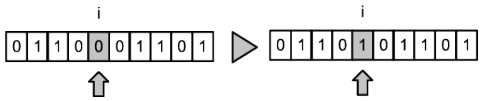


Figure 2.7: Bit-flip mutation [4]

– **Swap mutation**: when applying the swap mutation to binary chromosomes, two genes are randomly selected and their values are swapped. Figure 2.8 displays bit-flip mutation.

Figure 2.8: Swap mutation [4]

– **Inversion mutation**: a random sequence of genes is selected and the order of that sequence is reversed. Figure 2.9

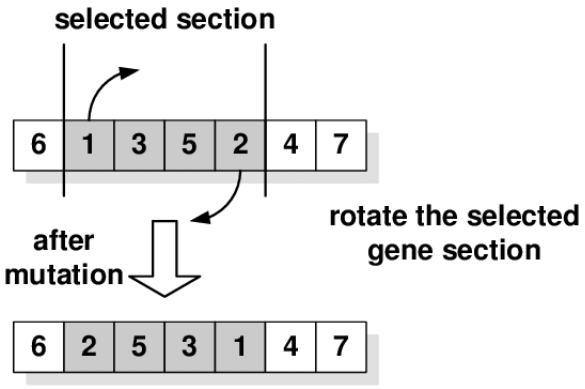


Figure 2.9: Inversion mutation [5]

2.5.1.6 Stopping criteria

This process above will keep iterating until a stopping condition is met. It can be a certain number of generations, a fitness value, or a convergence towards a satisfying solution. It's a delicate step to find the optimal trade-off between time and quality.

2.5.2 GAs applications

Genetic algorithms have been used in a variety of applications including:

1. **Travelling Salesman Problem (TSP)**: which is an NP-hard problem that has numerous implications for tasks like vehicle routing problems, logistics, planning and scheduling.

2. **Neural network hyper-parameter optimization:** the space of hyper-parameters in a Neural network is often large enough that it would be impossible to test all the combinations and that makes GAs a great candidate.

3. **Protein folding simulation:** it's considered an important task in biology and has multiple applications in the medical field such as drug design and discovery[36].

4. **Computational creativity:** which includes different artistic endeavors such as music generation [37].

5. **Timetabling problems**.

2.5.3 GAs advantages

Genetic algorithms have a multitude of advantages compared to other methods, to mention a few:

1. **Parallelism**: since calculating the fitness function is done for each individual independently [38].

2. **Simplicity of the concept** (which makes them often easier to implement in code).

3. **Considerably faster** than traditional brute-force methods (since they are stochastic).

4. **Applicable to multi-objective optimization**.

5. **Exploration of a wide solution space**.

2.5.4 GAs limitations

Although GAs are considered to be powerful and versatile, they still suffer from some drawbacks, including but not limited to:

1. **Computationally expensive**: because GAs often explore a huge space of solutions .

2. **Unguided mutations**: which can hinder a near-optimal solution [39].

3. **Premature convergence:** is generally caused by the loss of diversity within the population[40]

4. **Hyperparameter-tuning:** this includes the probabilities and types of both crossover and mutation, the population size, and the number of generations.

2.6 Conclusion

This chapter provided an overview of the field of evolutionary algorithms, with a focus on genetic algorithms, in order to better understand the techniques we used in our solution.

Chapter 3

Proposed Approach

3.1 Introduction

In this chapter, we will look at the food pairing hypothesis through the lens of evolutionary algorithms to determine if Algerian cuisine has positive or negative tendencies, that is, whether Algerian recipes tend to employ ingredients that share the same flavor components or if the opposite is true. First, we will present the global scheme of our solution. Then we will proceed to detail the methods and techniques used to accomplish all these tasks.

3.2 Global Scheme

The figure 3.1 presents an overview of the proposed approach which investigates the food pairing hypothesis stipulating that the more ingredients share chemical compounds, the more likely they are to be associated in a recipe. This study aims to answer this question for Algerian cuisine and determine its tendency (positive or negative). For this purpose, our approach includes mainly four stages: collecting the dataset and preprocessing it, generating random recipes using genetic algorithms, calculate the average flavor sharing of each of the real and generated recipes, and finally, compare the two cuisines and deduce the Algerian food tendency. The overview of the proposed approach is presented in figure 3.1.

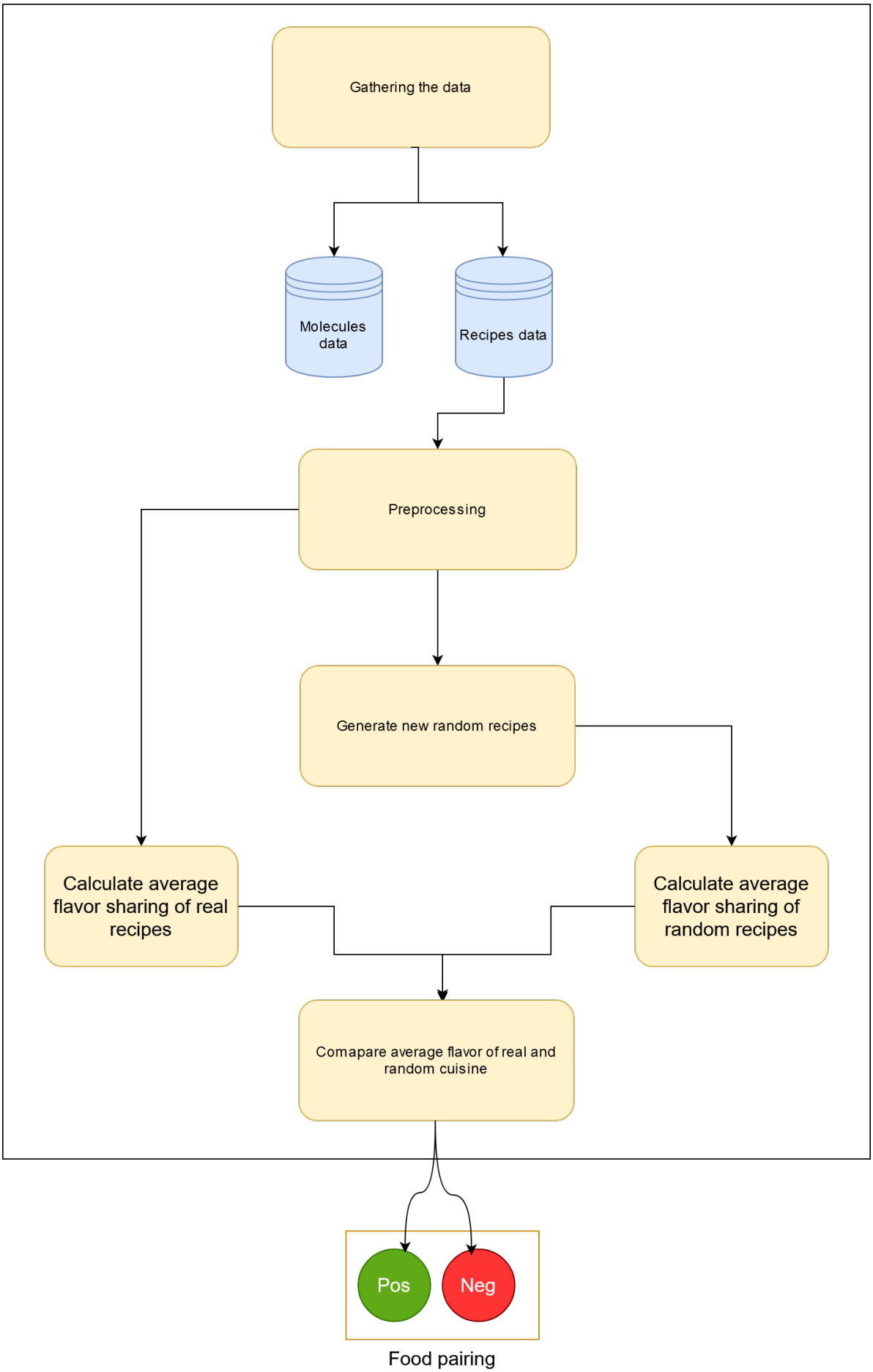


Figure 3.1: Overview of the proposed approach

3.3 Data Collection

Our project necessitates the use of two datasets: one containing traditional algerian recipes and their ingredients and the other the flavor compounds profile of ingredients.

- **Traditional Algerian Recipes Dataset**: To create our own dataset of recipes, we collected recipes from a traditional cookbook authored by Fatima Bouayad [41]. We have chosen this book because it contained authentic traditional food recipes only which dates at least two centries back. Modern Algerian books contains some compilation of Algerian meals that did not meet this requirement.

- **Molecules Dataset**: We used FlavorDB [42], a database that assembles a huge number of ingredients and their flavor molecules profile. However, as FlavorDB does not contain all ingredients needed in Algerian cuisine, we supplemented its information and adapted it to our needs by adding other ingredients with their corresponding molecular profile.

3.4 Pre-processing:

Data pre-processing, or the procedure of turning raw data into a comprehensible and organized format, is a critical stage since the quality of the data utilized directly affects the performance of any algorithm used. In our situation, we deal with a cookbook constituting of recipes, ingredients and their quantity, and preparation instructions, but we only require the unique components of the latter, not to mention the unstructured manner of its writing, which makes it difficult to extract the information needed. The purpose of this operation is to encode each recipe with its accompanying unique ingredients from a raw book format into a clean and structured dataset.

We proceeded through three primary phases to pre-process our data: data cleaning, data integration, and data transformation.

3.4.1 Data cleaning:

It is a technique for removing data that is incorrect, incomplete, or inaccurate from a dataset. For our data, we follow this principles by first correcting any possible spelling errors, especially because the book was scanned, and we did not have the typed version (due to its age), then discarding inaccurate and impertinent recipes for the task, and finally addressing the incomplete information, such as missing ingredients in FlavorDB being replaced with other relevant ingredients, and for other missing ingredients being deleted with their recipes.

3.4.2 Data integration

At this stage, we combine various sources to create a large dataset that contains as much information about the problematic as possible. For example, we utilized this strategy for FlavorDB; most of the data was previously obtained from the website and saved in a file, but when we discovered some missing ingredients, we had to search additional resources in scientific literature in order to add their flavor profiles to our dataset.

3.4.3 Data transformation

In this step, we convert the data into a format that is suitable for the following stages ; in our situation, we normalized our data by assigning a fitness value to the ingredients.
Figure 3.2 illustrates the most important stages of the procedure of data pre-processing.

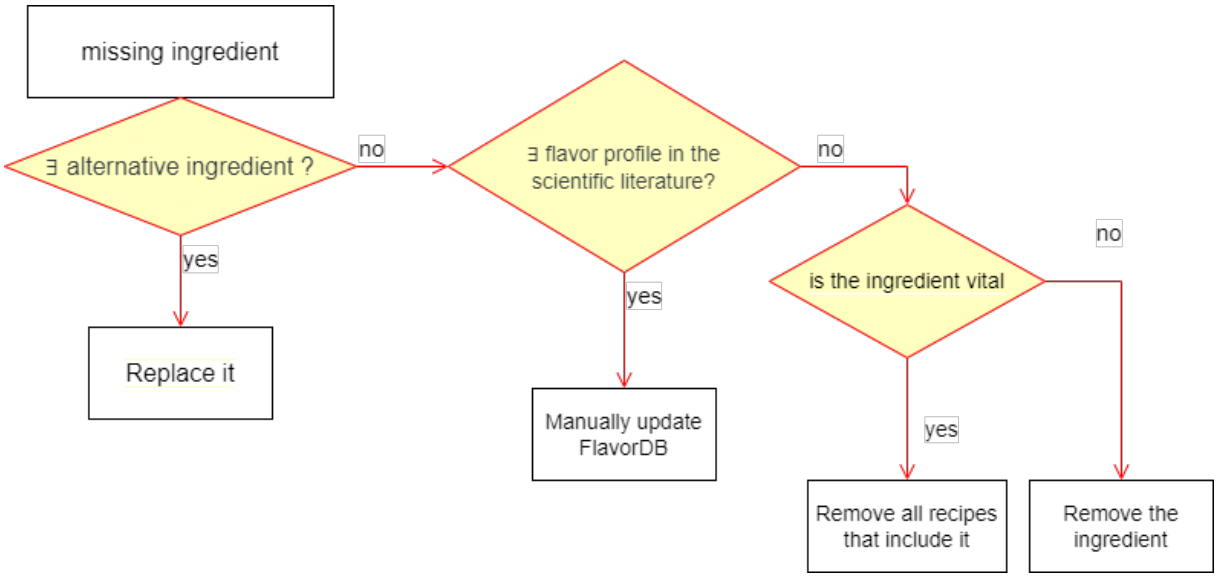


Figure 3.2: Diagram illustrating data pre-processing

3.5 Random recipe generation

To conduct the food pairing analysis, we employed the test presented in one of the field's pioneer studies conducted by Yong-Yeol Ahn at al. [2], which was afterwards applied in all future articles [1] and [14]. If the hypothesis is valid, the set of recipes in the cuisine analyzed should share more flavor components on average than any random set of recipes. The first phase in this test is to generate a collection of random recipes to compare with the real set, and the second step is to calculate the average flavor sharing of the cuisines.

3.5.1 Fitness Value

Attributing a fitness value to each ingredient is a necessary step for generating the random set of recipes, this one being a measure with the role of expressing the value of an ingredient based on implicit information, its frequency, or other factors. This value is in normalized form, therefore we resorted to applying two approaches, the first is used by Al-Razgan et al. [1] namely $FitnessValue(FV)$ the value of the frequency of the ingredients in the recipe normalized between 0 and 1 , and the second is an experimental fitness value we named IFW. The formula

we propose to calculate the fitness value of an ingredient i is:

$$IFW(i) = \frac{Nr(i)}{\sum_{j=1}^{Nr(i)}(N(R_j(i))} \tag{3.1}$$

Where:

Symbol	Description
$IFW(i)/FV(i)$	Fitness value of the ingredient i
$R(i)$	A recipe that contains the ingredient i
$N(R)$	The cardinality of the recipe R
$Nr(i)$	Number of recipes containing the ingredient i
R	represents a recipe

Table 3.1: Annotations in this subsection and their description

3.5.2 Generation process

The generation of the recipes is a crucial part of investigating the hypothesis. For this purpose, we rely on the method used in the recent article [1] which employs a copy mutate algorithm, the stages are summarized in the following algorithm:

Algorithm 1 Genetic algorithm used to generate random recipes [1]

1. Initial pool containing N numbers of ingredients.
2. Generate K number of recipes containing S ingredients randomly constructed from the initial pool of ingredients.
3. From the pool of K recipes select one recipe randomly.
4. Copy the recipe.
5. Mutate the copied recipe M times.
6. Randomly select an ingredient from the copied recipe.
7. Randomly select an ingredient from ingredients pool.
8. Compare fitness value of the two ingredients.
9. Replace the ingredient in the recipe with the ingredient having the higher fitness value.
10. Add the mutated recipe to the pool of recipes.
11. Introduce a new ingredient to the pool of ingredients.
12. Repeat step 3 to 7 until the corpus of random recipes match the number o recipes in the real cuisine.

We begin by initializing a set of ingredients I_0, which we use to generate a fixed-size set of template recipes R_0, and then we randomly select and copy a recipe from R_0 to apply the mutation to it, which is done by selecting a random ingredient i from the latter, randomly selecting another ingredient j from the set I_0, and comparing the fitness value of i and j, if $f(i) > f(j)$, then i is replaced by j in the recipe and removed from the set I_0, the mutation operation is repeated M times, and at its conclusion the mutated recipe is added to the set R_0. The procedure of copying and mutating a recipe, according to this algorithm, must be repeated until the cardinalities of both the real cuisine and the randomly generated one are equal.

For the sake of experimentation, we run another algorithm based on the work proposed in

[23] which employs a similar algorithm but differs from the previous one in that the copy and mutation steps are repeated until the ratio of the number of ingredients in the pool and the number of recipes in the pool corresponds to the ratio of the total number of ingredients and total number of recipes in the pool.

Algorithm 2 Algorithm for copy mutate model [23]

Require: List of ingredients in a cuisine (I),
average recipe size of a cuisine ($\overline{s}$),
size of initial recipe pool (n),
size of initial ingredient pool (m),
total number of recipes in cuisine(N),
number of mutations (M),
ratio of total number of recipes in the cuisine ($\varnothing$)

Ensure: N mutated recipes

for all ingredients i in l **do**
    sample a fitness value for ingredient i
    assign it to i
end for
I0 $\leftarrow$ randomly sample (without replacement) m ingredients from I
$I \leftarrow I - I0$
R0 $\leftarrow$ randomly sample $\overline{s}$ ingredients n times from I0
for $I = 1toN - n$ **do**
    $\delta \leftarrow m/n$
    if $\delta \geq \varnothing$ **then**
        r $\leftarrow$ randomly choose a recipe from R0
        for g=1 to M **do** sample an ingredient i from r sample an ingredient j from I0
            if fitness of i $\geq$ fitness of j **then**
                replace i with in r
            end if
        end for
        $R0 \leftarrow R0 + r$
        $n \leftarrow n + 1$
    else
        choose an ingredient p randomly from I
        $I0 \leftarrow I0 + p$
        $m \leftarrow m + 1$
        $I \leftarrow I - p$
    end if
end for

3.6 Average flavor sharing calculation

The average flavor sharing of a particular cuisine estimates the average number of chemical flavor components shared by all the items used in the recipes. We accomplish this by first calculating the same metric for each recipe, and then computing the mean of the overall cuisine. This procedure applies to both the real and random cuisines. To formalize this procedure, we will use the annotations mentioned below:

The following equation will calculate the number of flavor compounds shared by two ingredients (i, j), which is actually the cardinality of the intersection of the sets $F(i)$ and $F(j)$.

$$Ns(i,j) = |F(i) \cap F(j)| \tag{3.2}$$

This formula is applied to all combinations of the ingredients in the recipe R and plug it into the equation $Ns(R)$ to get the average flavor sharing.

$$Ns(R) = \frac{2}{nR(nR-1)} \sum_{i,j \in R; i \neq j} |F(i) \cap F(j)| \tag{3.3}$$

Ns(R) is the average flavor sharing for one recipe, thus to get the mean of the entire cuisine, we must iterate the process for all the recipes in it, as given in the formula below.

$$Ns = \sum_{R} \frac{Ns(R)}{Nc} \tag{3.4}$$

Where:

Symbol	Description
R	represents a recipe
$Ns(i,j)$	number of shared compounds between the ingredients i and j
$F(i)$	Set of flavor compounds found in ingredient i
$Ns(R)$	average flavor sharing of the recipe R
nR	number of ingredients in the recipe R
Ns	Average flavor of the cuisine
Nc	Number of recipes in the cuisine

Table 3.2: annotations of the formulas above and their description

3.7 Testing the food pairing hypothesis

At this point in the study, we have all the data we need to confirm or refute the above-mentioned hypothesis, namely whether the ingredients associated with Algerian cuisine generally share flavor components or, on the contrary, whether it tends to use elements with the fewest aroma compounds in common, and in order to carry out this task, we must calculate the difference between the average flavor sharing of the real Algerian cuisine "$Ns(real)$" and the random one generated using the genetic algorithm "$Ns(rand)$" according to this equation:

$$\Delta N = Ns(real) - Ns(rand) \tag{3.5}$$

Based on the results of this formula, we can examine the tendencies of Algerian food; if $\Delta N > 0$, the food pairing is positive; otherwise it's negative. Because the copy mutate algorithm we used to generate the random cuisine is stochastic, we repeated the process a hundred times for each different parameter value, and we calculated their average flavor sharing to compare and see if

they all produce roughly the same results.

3.8 Conclusion

In this chapter, we examined the process and the overall steps we followed during the implementation of our approach, beginning with a broad scheme of our solution, and then detailing the strategies involved in each stage.

Chapter 4

Test and Validation

4.1 Introduction

In order to conduct our research, we went through multiple stages of data preparation, starting with collecting traditional lgerian recipes, followed by the necessary text modifications and the extraction of the particular ingredients. And then finally, associating these substances to their molecular components in order to create a flavor profile. This chapter will cover the implementation of our approach and all the steps mentioned above.Finally, we will present the conducted experimentation and obtained results.

4.2 Tools and working environment

In this section, we will be introducing our work environment and all the tools and libraries that we have used:

- Programming language: Python 3
- Libraries used: re, pattern, NLTK, pandas, enchant, deep translator, wordcloud, plotly

4.2.1 Hardware specifications

Three different environments have been used in the implementation of this work:

- **Desktop PC:**
  - GPU: GTX 1060 3GB
  - CPU: I5-3340 CPU @ 3.10GHz
  - RAM: 8,00 GB
  - OS: Microsoft Windows 10 Pro
- **Laptop:**
  - GPU: AMD Radeon R5 M230
  - CPU: i7-4510U CPU @ 2.00GHz
  - RAM: 6,00 GB
  - OS: Microsoft Windows 10 Pro
- **Google Colaboratory:** Google Colab is an online Jupyter Notebooks environment from Google, it allows its users to code without having to set python locally, and it comes equipped with powerful hardware and bandwidth.

4.2.2 Python

Python is a high-level, interpreted, general-purpose programming language, designed for simplicity and readability [43],it was developed in 1991 by Guido van Rossum and had grown to become one of the most used programming languages in the world according to stack overflow [44], and it is especially dominant in fields like machine learning and artificial intelligence development in general.

4.2.3 Regular Expressions library

Regular Expression (re) are the backbone of most text processing tasks. We have resorted to using re, which is the default Regular expression module in python. It supports both Unicode strings (str) and 8-bit strings (bytes) [45], and the default supported functionalities proved to be sufficient in our case.

4.2.4 Pandas Library

Pandas is a fast and versatile open-source data manipulation tool [46]. The main data structure used in pandas is called a dataframe which is a tabular representation of data. Dataframe is highly optimized for performance with critical code written in Cython or C.

4.2.5 NLTK Library

Also known as the Natural Language Toolkit, is also a standard open-source python library; developed by Steven Bird and Edward Loper in the Department of Computer and Information Science at the University of Pennsylvania [47]. It provides a wide variety of NLP functionalities: such as stemming, lemmatization, tagging, parsing....

4.2.6 Deep_translator

Deep_translator is an open-source third-party library that supports different ten translating services such as Google Translate, Yandex, and Microsoft translator. It supports a lot of features out of the box such as automatic language detection and batch translation.

4.2.7 Pattern Library

Pattern is a third-party NLP web mining library, that has extensive features in NLP. It shares some of NLTK features (such as n-grams, part of speech tagging, etc...) [48]. But it also includes other useful features such as sentiment analysis, spellchecking, singularizing and pluralizing.

4.2.8 Pyenchant Library

Pyenchant is a package that contains a set of python bindings for Enchant [49], which is a spellchecking library that wraps a number of different spelling libraries and supports multiple languages.

4.2.9 Wordcloud

Wordcloud, is a visualization tool that enables word cloud generation, with full customization such as support of stop words, coloring, size, etc...

4.2.10 Plotly Library

Plotly is an interactive, open-source, and browser-based graphing library for Python [50]. It comes prepackaged with 3D graphs, animations as well as extensive controls over the plots.

4.3 Dataset

4.3.1 Traditional Algerian recipes data

To collect the data, we chose Ms. Bouayed Fatima Zohra's well-known algerian book "*la cuisine algérienne*" [41]. It contains 400 authentic Algerian recipes from diffrent algerian regions, and various categories of foods, e.g breads, vegetables, soups, deserts, pastry... etc. However, it presents a number of difficulties, primarily due to its age and formatting issues in the text. So, before we could proceed with the actual pre-processing of the text, we had to manually review it first.

4.3.2 Molecules dataset

FlavoDB [42] is the database we used to associate each ingredient with their flavor profile. It is a repository of molecular features. This database is the most suitable for our work because it focuses on the chemical aspect of flavors and provides natural sources of the ingredients [42]. Unlike other databases, we found FoodDB, for example, compiles ingredients molecules but not on the flavor point of view. Another example of a database that did not match our requirements is Flavornet which is a database of flavor molecules but lacks information about their natural sources [51].

FlavorDB was created to collect the most comprehensive resources about flavor molecules into a single database. This was accomplished by collecting all of the necessary information from various resources such as FoodDB, Flavornet, SuperSweet, BitterDB, and many others. It covers 25595 flavor molecules, 2254 of which are associated with 936 ingredients [42].

4.4 Preprocessing

The first step was to separate and save the recipe's composition ingredients list from the preparation of the food. For this purpose, we used a regular expression regex that delimits the recipes with the starting key word "composition" and ending with any number (d+).

```
import re
import pandas as pd
def GetTheSentences(textfile,start,end):
    """
    returns chunk of text between two custom delimieters
    Arguments:
        textfile: a String representing the path to the file
        start: a string representing the first delimiter
        end: a string representing the second delimiter
    returns:
        a list of Recipes
    """
    ls=[]
    with open(textfile,encoding='UTF-8') as fp:
        for result in re.findall(f'{start}(.*?){end}',fp.read(),re.S):
            ls.append(result)
    return ls

ingridient_list=GetTheSentences("dataset Cuisine Alg\\recettes_rasha.txt",'
    COMPOSITION.','\d+\.')

recipe_names=GetTheSentences("dataset Cuisine Alg\\recettes_rasha.txt","\d
    +\.",'COMPOSITION')
```

```
22 data = {'recipe_name':recipe_names, 'Ingredients':ingridient_list}
23 df = pd.DataFrame(data)
```

Listing 4.1: Python snippet for seperating the recipes

4.4.1 Cleaning recipes dataset

4.4.1.1 Discarded irrelevant categories

We excluded some categories of the book from the start because they were irrelevant to our context:

- the category of "drinks," which contains 5 recipes that do not represent Algerian cuisine (such as coffee...)
- the category of "preserved vegetables and fruits," in total 25 recipes, that detail of how preserve vegetables and fruits. However, it concern a specific ingredient with the same manner of preparation with few ingredients like salt and water.

4.4.1.2 Supplementary words

After obtaining the list of ingredients for each recipe, we noticed that in addition to the ingredients, the list contains supplementary words such as measurements (e.g. 1 Kg, a handful of, grilled, . . . etc.) and descriptive adjectives, which we eventually eliminated while being cautious about accidentally removing ingredients composed of several words like: « pomme de terre, huile d'olive . . . ».

4.4.1.3 Multiple ingredients in the same line

In some cases, multiple ingredients are mentioned in the same line, and there were two variations of this issue:

- The first was when the recipe provided alternative ingredients for example: "1 cuillerrée de beurre ou de smen"; and since we strived to keep the recipe's unique ingredients, we decided to keep only the first suggestion.

- The second was when important ingredients were mentioned in the same line but separated by the connector "et"(and) for example: “1 verre d’huile et de beurre”. And considering the fact that the elements in question are not optional, we replaced the "et" (and) with "-" to treat the second ingredient as a separate and independent one.

```
df['recipe_name']=df['recipe_name'].apply(lambda row:(row.split('\n')))
# makes sure the ingredients are in lower-case
df['Ingredients']=df['Ingredients'].str.lower()
# a small sample of stopwords we removed
df['Ingredients']=df.Ingredients.str.replace(r"\bk?g (de)?\b|louche|gousses
    ?|pour friture|[0-9] ou [0-9]", ' ')
# splitting the recipes into lists of ingredients
df['Ingredients_alpha']=df['Ingredients'].apply(lambda row:row.split('-'))
# removing trailing white space
df['Ingredients_alpha']=df['Ingredients_alpha'].apply(lambda row:[w.strip()
     for w in row])
# removing empty words
df['Ingredients_alpha']=df['Ingredients_alpha'].apply(lambda row:[w for w
    in row if w !=''])
# removing the second part of an "or" statement
df['Ingredients_alpha']=df['Ingredients_alpha'].apply(lambda row:[re.sub(r"
    \bou\b.*",'',w) for w in row])
# removing duplicate ingredients
df['Ingredients_alpha']=df['Ingredients_alpha'].apply(lambda row : list(set
    (row)))
```

Listing 4.2: Python snippet containing a subset of preprocessing

4.4.1.4 Compounded ingredients

The compounded elements peculiar to Algerian cuisine, such as " Dersa, Hror, Ras el hanout, pate de dates . . . ” were one of the most troubling challenges we came across. These elements, often cited as ingredients in recipes, but are actually made up of multiple basic components that are given in an other separate recipe. And since our study relies on the atomic ingredients, we replaced the composed ingredients with their constituents and then deleted the original recipe

from the dataset if it existed. In table 4.1, we present the set of composed ingredients and the list of their basic elements. In addition, we enumerate in table 4.2 the list of composed spices and their composition that were replaced with. Finally, table 4.3 presents the list of composed recipes and their composition that were replaced with.

Name of recipe of composed ingredients deleted	***List of the elements in the recipe***
Pate de dattes	Eau de fleur oranger, miel, clous de girofle, cannelle, dattes, cubèbe, huile
Sucre inverti	Eau, alun, sucre
Viande séchée/qaddid/ khli	Bœuf, huile, ail, carvi, poivre rouge, poivre noir, coriandre, cumin, sel
d'youl	Semouline, farine, sel, eau
Plombs	semoule, semouline, sel
Cheveux d'ange/ vermicelle pour gateaux	Farine, sel, eau, moelle épinière de bœuf, huile
Levain	Farine, vinaigre, pain, sel, eau
Tlitli	Semouline, sel, eau
Crêpes	Semouline, farine, sel, eau
Confiture d'abricot	Abricots, sucre

Table 4.1: List of the recipes replaced as composed ingredients and deleted from the database

Name of composed spices deleted	*List of the basic ingredient's constituents*
chermoula	Oignon, ail, persil, poivre rouge, cumin, poivre noir, citron, farine, sel, huile
Dersa	Ail, cannelle, cumin, piment, sel, laurier
Ras lhanout	Cumin, gingembre, sel, poivre, cannelle, coriandre, piment de la Jamaïque, Clou de girofle, piment

Table 4.2: List of composed spices replaced with their composition

Crêpes	Semouline, farine, sel, eau
Confiture d'abricot	Abricots, sucre

Table 4.3: List of composed recipes replaced with their composition

4.4.1.5 Synonymous ingredients

We found some elements that were referred to differently in various passages so we unified those synonyms by choosing one representative word, for example "berkoukes" and "petits plombs" were represented by "petits plombs".

4.4.1.6 Final version of the recipes dataset

With the exception of this final obstacle, we had a fairly clean list of ingredients at this point, with just the unique constituents remaining. The wordclouds in figures 4.1 and 4.2 show the contrast between the dataset before the cleaning process and after, respectively.

Figure 4.1: Wordcloud demonstrating the dataset before preprocessing. The size of the words is an indicator of the frequency of the terms, in red are represented the ingredients and in green: are shown the words not related to the unique ingredient

Figure 4.2: Wordcloud showing the database after the cleaning process. the size of the text indicates the frequency of the ingredient in the database

4.4.2 Linking recipes data to molecules data

After cleaning the recipes dataset, the next step was to connect them to FlavoDB. We utilized FlavoDB for this task because it was readily available. But we had to first translate the unique ingredients obtained to English, as the cookbook we utilized for our dataset was originally written in French, while the data in FlavoDB is in English.

4.4.2.1 Mistranslation issues

We faced few mistranslation issues during the automatic translation process. First and foremost with a number of distinct ingredients that gave the same translation, which is problematic since it would link a certain element to the incorrect flavor profile of another ingredient. For example, "piment rouge" and "poivre rouge" both translate to "red pepper", but when we check the database, we find that "red pepper" isn't even listed, so we looked for other names for the similar ingredients, and "piment rouge" was substituted with "capsicum".

The second issue was that the translation output sometimes didn't match the names of the ingredients in FlavorDB. Generally speaking, the database of flavor molecules records ingredients in the singular form, but the translation output for some components was in the plural form. To solve this problem, we used the library Enchant to check the spelling first and then correct any potential misspellings, as well as the function singularize() from the web mining module pattern.en to return the singular form of plural nouns.

4.4.2.2 Missing ingredients in FlavorDB

Despite the great diversity of FlavorDB's 936 ingredients, we discovered that some ingredients were missing in the latter during the step of associating the recipe database with FlavorDB, such as ,olive oil, baking powder ... To remedy this, we took several steps by considering each case separately:

- **Adding the flavor profile to FlavorDB:** We used this method first wherever possible to avoid altering the recipe database as much as possible, and we did so by looking for a list of flavor molecules that characterize the item in other sources. This approach has worked

well with olive oil, which we were able to include thanks to the article of Silava and al. [52]. Baking powder is made up of an acid, a base, and an inert component that helps keep the reactive elements apart and prevents a premature reaction. In most current baking powders, the acid is baking soda (called sodium carbonate in the molecular database) and the base is tetrasodium pyrophosphate, with maize starch serving as the inert component [53].

- **Replacing an ingredient with a suitable substitute:** For certain ingredients that don't exist in FlavorDB, we chose to replace them with other ingredients found there, such as "smen" with "butter" or "levure boulangère" with "levain", which is used as a substitute in all bread items.

- **Remove the ingredient from the recipe:** For the unimportant ingredients in a recipe, we decided to remove them, and keep the recipe. We applied this technique when it comes to "fat" for example, where the ingredient was not always obligatory.

Mhadjeb Carrés de pâte farcis
Chalvoq

Table 4.4: List of recipes where fat was optional

- **Deleting the recipe from the database:** This decision is the last resort we had in hands, after failing too find the volatile compounds of the ingredient or not finding a substitute for the latter, we chose to remove the recipe. The table 4.5 demonstrates the complete list of recipes deleted and the ingredients missing.

Missing ingredient	*Recipe removed*
Cédrat	Trandj msakker / Cédrats confis
Gras	l-Maghlouga / Crêpes farcies
Mauve	Khoubbiz- Bqoul- Moudjdjir / Mauve aux légumes
Cervelle	Mokh b'tomatich / Cervelle à la tomate
Tripes	Dowwara b'l-khodra / Tripes aux légumes
Moelle épiniere de boeuf	Qtâyef / Vermicelles pour gâteaux
	Qtâyef khobza / Gâteaux aux cheveux d'ange
Langue de veau	Mtewwma b'l-lsân / Langue de veau à l'ail
	Lsân mekhft / Tranches de langue de veau frites
Feuille de vigne	Dalya m'aamra b roz / Feuilles de vigne farcies au riz
	Dolma dalya / Feuilles de vigne farcies
Boyau	Osbâne / Panse farcie confite
	Dolma masrân / Boyau farci en sauce
	Dolma masrân maa el-khodra / Boyau farci et légumes
	Kesksou b'l- osbâne -Taâm b'l-bekbouka / Couscous à la panse farcie
Foie	Kebda mchermla / Foie en sauce
	Kebda b'l-bsel / Foie aux oignons
	Kebda b'tomatich / Foie en sauce tomate
	Melfouf- Boulfâf / Brochettes de foie
Tete d'agneau ou de mouton	Bouzellouf b'l-mermez / Tête de mouton à l'orge concassée
	Bouzellouf mfawwer / Tête de mouton à la vapeur
	Bouzellouf marqa / Tête de mouton en sauce
	Bouzellouf loubya / Tête de mouton aux haricots secs
	Bouzellouf chtetha-Hergma / Tête d'agneau ou de mouton en sauce

Table 4.5: List of the missing ingredients in FlavorDB and the recipes deleted because these ingredients were important.

4.5 Visualisation

To summarise, we started with 400 recipes from the reference cookbook, discarded 78 for the previously reasons outlined in this chapter and summarised in the table 4.6, and ended up with 322 recipes in total for our database, with 116 unique ingredients. Dominent ingredients in algerian recipes are presented in figure 4.4. Similarly, figure 4.3 illustrates the most common ingredients in Algerian cuisine and their occurrences. The size of the bubbles indicates the frequency of the ingredients.

Reason for removal	**Number of recipes removed**
Recipes from "drinks" category	5
Recipes from "preserves" category	25
Recipes of composed ingredients	8
Missing ingredients in FlavorDB	24

Table 4.6: Summary for recipes deleted from the database

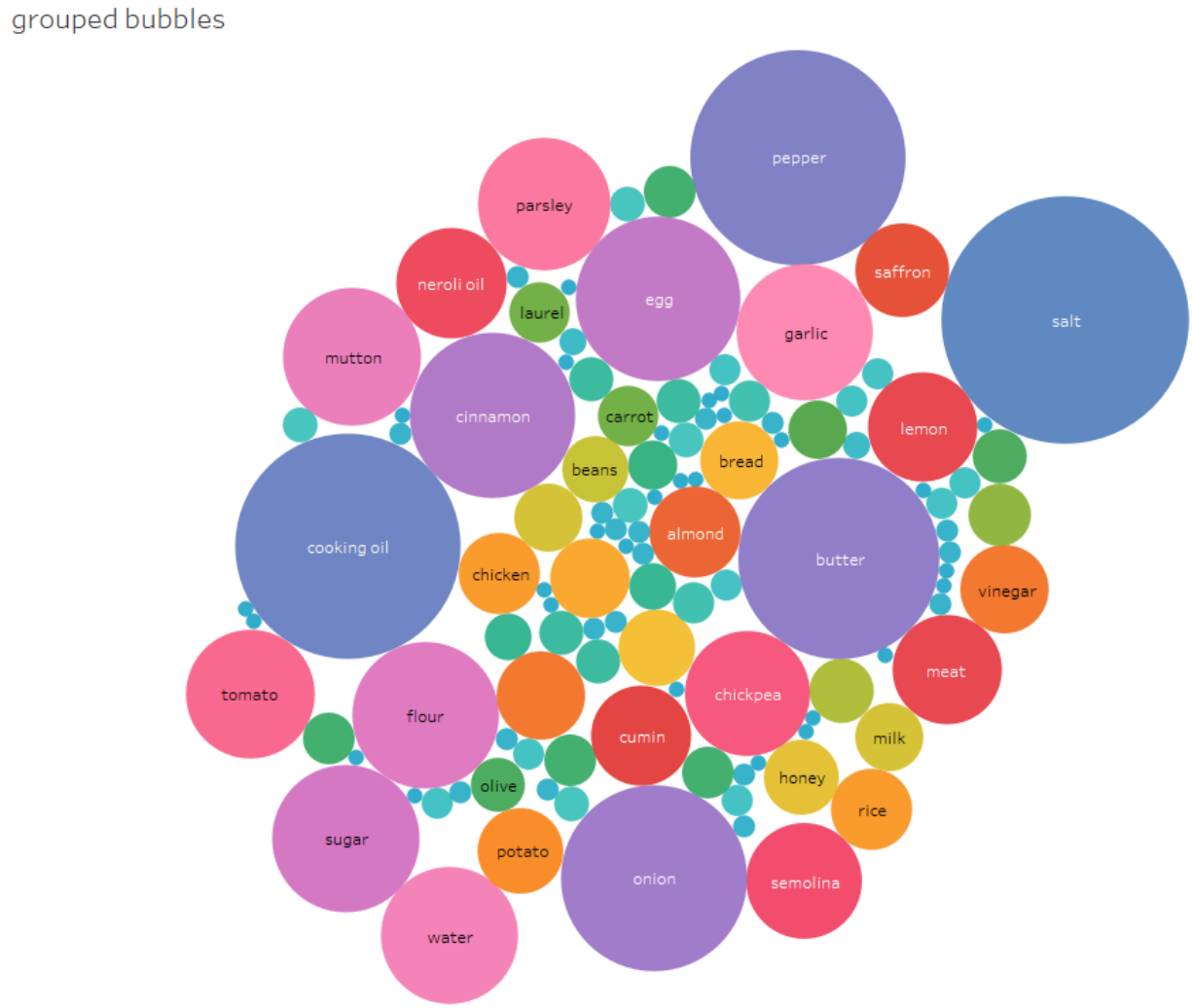


Figure 4.3: Illustration showing the most common elements in our database and by extension in the Algerian cuisine and their occurrences. The size of the bubbles indicates the frequency of the ingredients

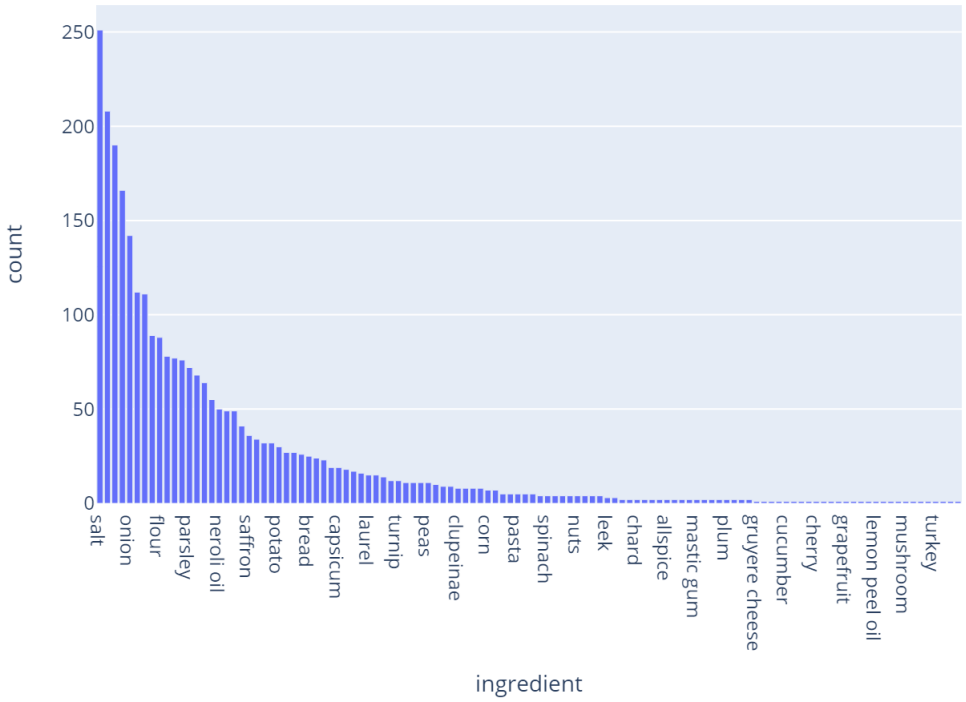


Figure 4.4: Barchart showcasing the difference in the occurrence of certain dominant ingredients and other less recurrent

We decided to separate the savory recipes and the pastry recipes into two distinct databases for the purposes of this study. The savory food database contains 232 recipes, with each recipe containing 9.87 ingredients on average (minimum of 2 and maximum of 23 ingredients), whereas the pastry recipes dataset contains 90 recipes in total, with an average of 6.3 ingredients used in a recipe (minimum of 2 and maximum of 12 ingredients). The histograms in the figures 4.5 and 4.6 show the distribution of the number of ingredients in recipes in both datasets.

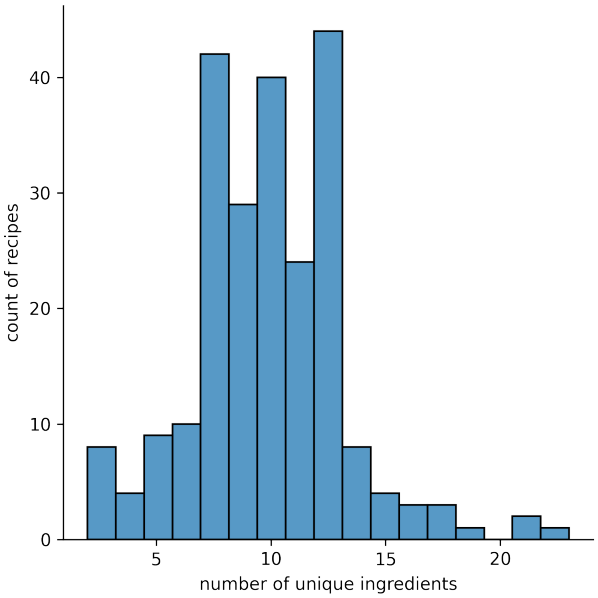


Figure 4.5: Histogram of the number of ingredients in recipes of the food database

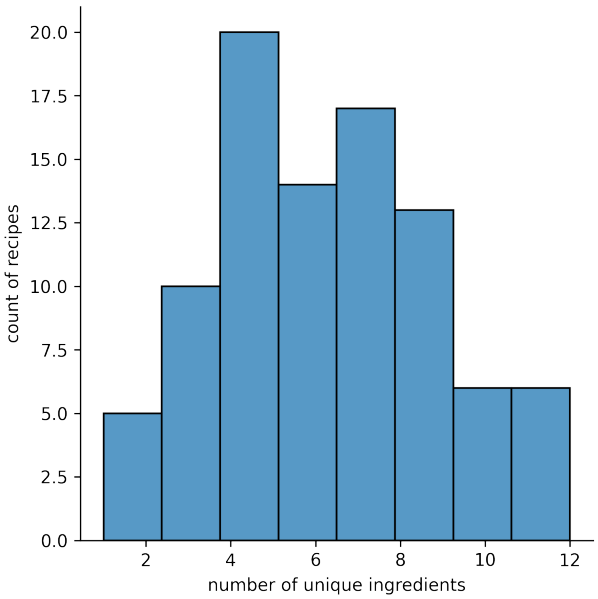


Figure 4.6: Histogram of the number of ingredients in recipes of the pastry database

4.5.1 Characteristic ingredients of Algerian cuisine

According to these statistics, the most common ingredients in Algerian cuisine are: “pepper, cooking oil, sugar, eggs, cinnamon”, but to get a clearer picture of the representative elements used in Algerian cuisine, we removed additive ingredients like salt, cooking oil, water, and sugar, and we discovered that the majority of recipes use the following ingredients: “pepper, butter,

onion, cinnamon, and eggs". As for the most co-occurring ingredients in the recipes, they are listed in the tables 4.7 and 4.8. We extracted them, for the pastry dataset and the food dataset separately.

Ingredients	***Count***
(Butter, Pepper)	89
(Onion, Pepper)	73
(Butter, Onion)	65
(Butter, Cinnamon)	63
(Mutton, Onion)	62
(Cinnamon, Pepper)	61

Table 4.7: List enumerating the most common co-occuring ingredients in the food database

Ingredients	***Count***
(Butter, Flour)	27
(Flour, Egg)	26
(Butter, Neroli Oil)	24
(Butter, Egg)	19
(Almond, Butter)	18

Table 4.8: List enumerating the most co-occurring ingredients in the pastry database

4.6 Generation Process

In this section, we will use genetic algorithms to generate random recipes, and then apply the formula delta $\Delta Ns = Ns(real) - Ns(rand)$ to confirm the hypothesis. We tested different configurations for different variables, and after enumerating them, we will compare them and deduce a conclusion for food pairing in Algerian cuisine.

4.6.1 Fitness value

As mentioned in Chapter 3, each ingredient must first be assigned a fitness value, and we experimented with two different fitness functions to do so.

- FV : The first method employed by Al-Razgan et al. [1] which calculates the normalized frequency of the ingredient i in the corpus of recipes and it ranges from 0 to 1.
- IFW: The second formula weights the fitness value by evaluating the importance of the ingredient in a recipe in relation to the corpus of total recipes detailed in equation 3.1.

Both fitness functions require numerical data, but our data (the ingredients) is nominal, so it must be converted to a numerical form, and for this we used the one hot encoding, so each recipe is represented by a vector with 116 values corresponding to all the ingredients present in the cuisine, with each value being either 1 or 0, depending on whether the recipe contains the ingredient or not.

The table 4.9 gathers both fitness values of the most recurrent ingredients in both datasets (savory food and pastry).

Dataset	Ingredient	FV Normalized frequency fitness value	IFW fitness value
Savory food dataset	meat	0.21120689655172414	0.09158878504672897
	butter	0.521551724137931	0.0965682362330407
	mutton	0.33620689655172414	0.0912280701754386
	salt	0.9267241379310345	0.1053921568627451
	cheese	0.034482758620689655	0.08080808080808081
	cinnamon	0.4051724137931034	0.0888468809073724
	onion	0.6120689655172413	0.09543010752688172
	pepper	0.40086206896551724	0.049102428722280884
	cooking oil	0.7844827586206896	0.09816612729234088
	bread	0.04310344827586207	0.0847457627118644
	egg	0.3232758620689655	0.0949367088607595
	chickpea	0.27586206896551724	0.0862533692722372
	parsley	0.3103448275862069	0.09836065573770492
	potato	0.12931034482758622	0.08403361344537816
	saffron	0.15086206896551724	0.08883248730964467
Pastry datase	flour	0.1810344827586207	0.09012875536480687
	banana	0.004310344827586207	0.1
	cinnamon	0.07758620689655173	0.12162162162162163
	cherry	0.004310344827586207	0.1
	lemon	0.07758620689655173	0.14634146341463414
	neroli oil	0.15086206896551724	0.12962962962962962
	melon	0.004310344827586207	0.1
	orange	0.021551724137931036	0.15151515151515152
	apple	0.004310344827586207	0.1

Table 4.9: Comparison between FV and IFW fitness values

4.6.2 Recipe generation

Then, in order to generate a set of random recipes, we tested two variations of the copy mutate genetic algorithm that differed in the halting condition:

- The first algorithm is based on a stopping condition that limits the number of created recipes to the size of our initial dataset.
- The other algorithm limits recipe production according to the ratio of the number of recipes in the corpus to the number of ingredients in the pool.

We repeated the process 100 times for each generation test, testing different parameters. The two algorithms we tested with are based on three parameters to control:

1. Recipe size: we defined the size of the recipe to be generated based on the average size of the recipes in our dataset, therefor for the food dataset with an average of 9.87 ingredients per recipe, we set the size to 10, however the pastry dataset has an average of 6.3 ingredients per recipe, as a consequence we set the size to 6.
2. Number of initial recipes(templates): we evaluated 4 different values for this parameter (3, 5, 7 and 9).
3. Initial ingredients pool size: we evaluated 3 diffrent values or this parameter ($1.5 * recipesize$, $2 * recipesize$ and $3 * recipesize$) therfore the test values for the savory food dataset were 15,20 and 30, and the values of the pastry dataset were 9, 12 and 18.

before applying this step, we calculated the average flavor of the real cuisine which resulted for the savory food dataset in $Ns(real) = 28.845858822764793$ and for the pastry dataset in $Ns(real) = 14.469971139971143$.

In the following sections, we will go over all the tests and results that we ran to generate the recipes

4.6.3 Experiments using Algorithm 1

After estimating the fitness value of each component using FV normalised frequencies and then IFW formula, we used our copy-mutate technique to construct 100 sets of randomised cuisines by altering the initial size of the recipes pool.

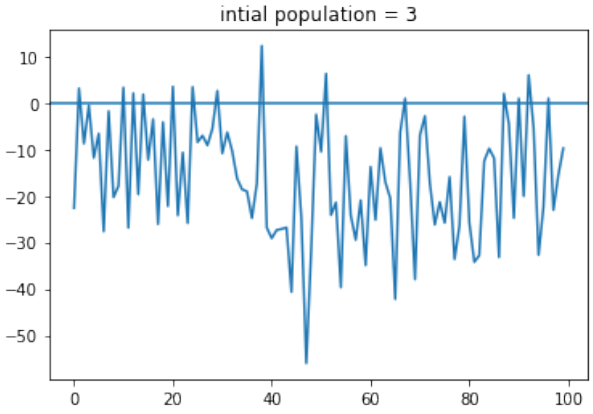
4.6.3.1 Template size parameter

Savory food dataset

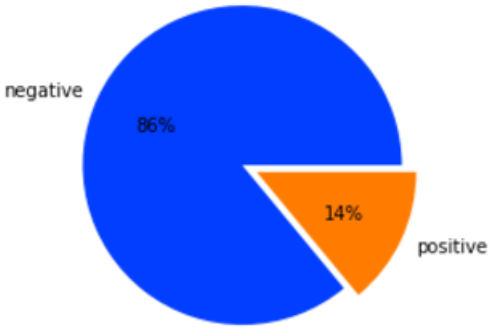
The table 4.10 summarizes the results of the different configurations tested on the savory food dataset, and the figures 4.7, 4.8, 4.9 and 4.10 show the delta variations for the normalized frequency fitness values and the figures 4.11, 4.12, 4.13 and 4.14 show the results for the IFW fitness values at each iteration:

		FV normalized frequency	Observation	IFW	Observation
pop 3	min delta	-55.97	negative food pairing in 86 % of the cases	-51.09	negative food pairing in 88% of the cases
	max delta	12.24		13.34	
	avrg delta	-16.004		-15.86	
pop 5	min delta	-40.58	negative food pairing in 93% of the cases	-43.99	negative food pairing in 85% of the cases
	max delta	12.24		10.19	
	avrg delta	-16.2		-13.28	
pop 7	min delta	-44.78	negative food pairing in 94% of the cases	-39.1	negative food pairing in 95% of the cases
	max delta	8.06		9.02	
	avrg delta	-16.9		-15.21	
pop 9	min delta	-40.93	negative food pairing in 97% of the cases	-39.56	negative food pairing in 86% of the cases
	max delta	4.32		8.42	
	avrg delta	-17.72		-13.99	

Table 4.10: Conclusion and comparison of the results of normalized frequency and IFW on different values for template size in algorithm 1 on savory dataset

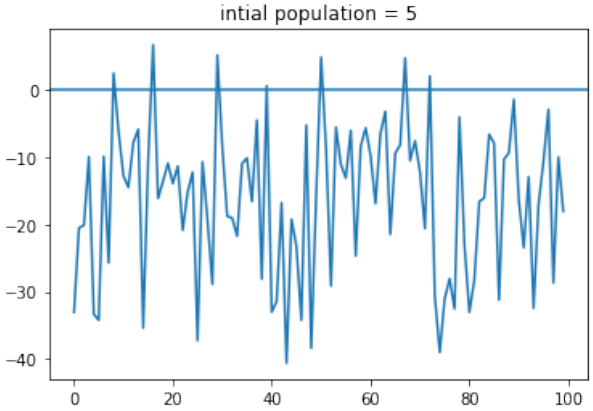


(a) Lineplot of delta variation

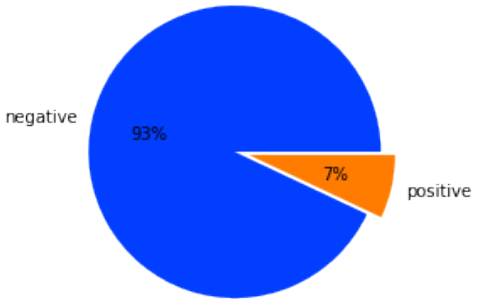


(b) Piechart of food pairing results

Figure 4.7: Results for initial population = 3 with normalized frequency

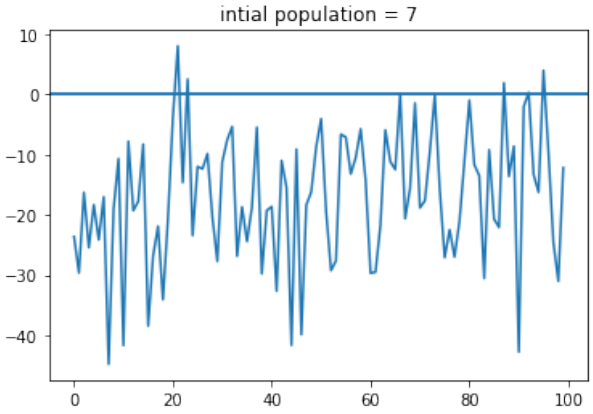


(a) Lineplot of delta variation

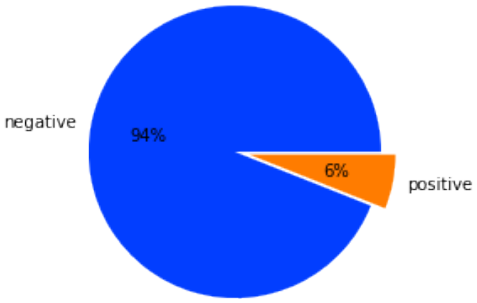


(b) Piechart of food pairing results

Figure 4.8: Results for initial population = 5 with normalized frequency

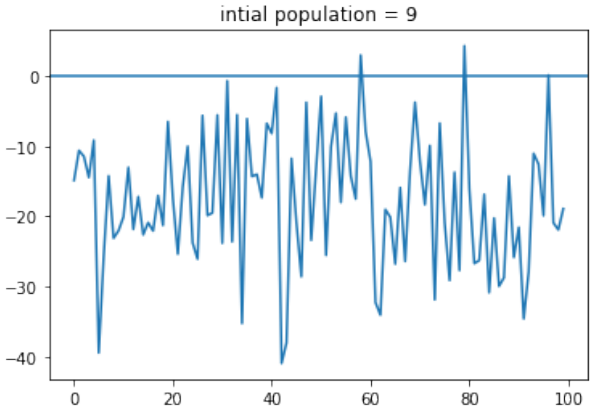


(a) Lineplot of delta variation

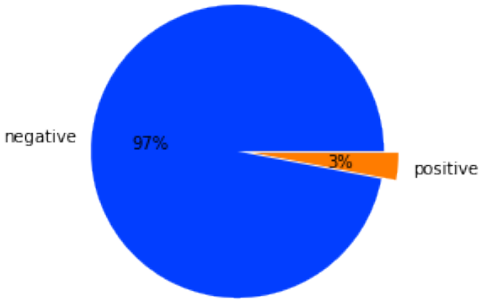


(b) Piechart of food pairing results

Figure 4.9: Results for initial population = 7 with normalized frequency

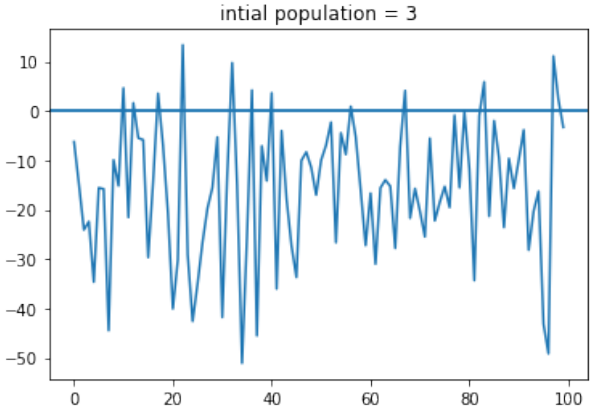


(a) Lineplot of delta variation

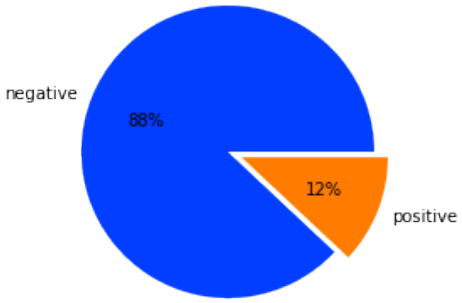


(b) Piechart of food pairing results

Figure 4.10: Results for initial population = 9 with FV normalized frequency

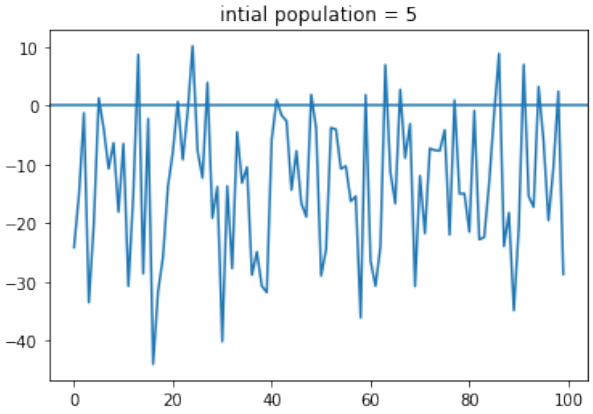


(a) Lineplot of delta variation

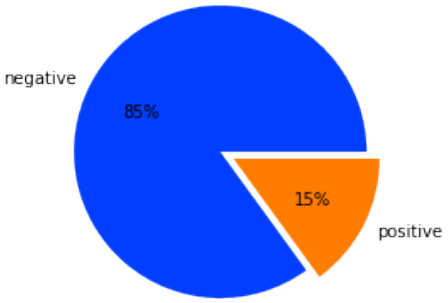


(b) Piechart of food pairing results

Figure 4.11: Results for initial population = 3 with IFW formula

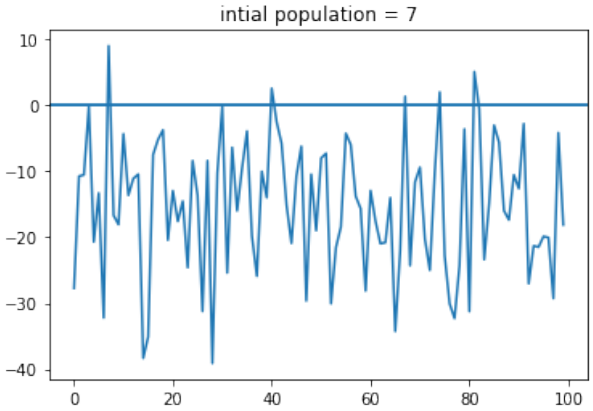


(a) Lineplot of delta variation

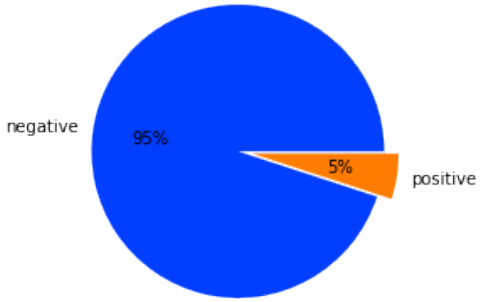


(b) Piechart of food pairing results

Figure 4.12: results for initial population = 5 with IFW formula

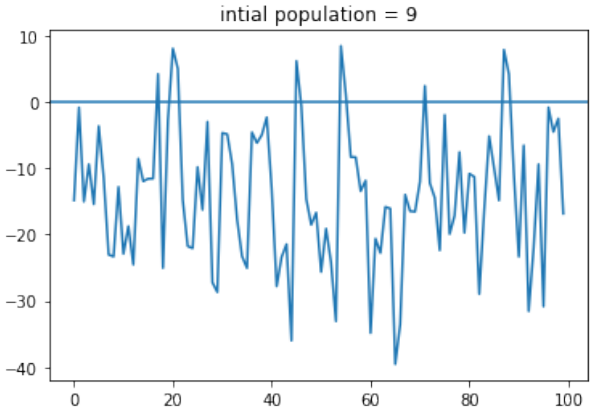


(a) Lineplot of delta variation

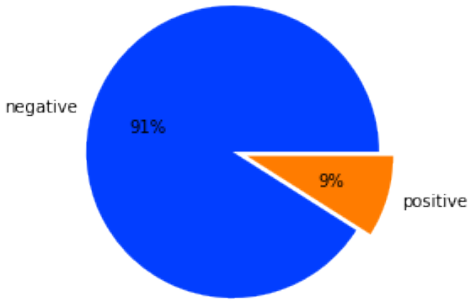


(b) Piechart of food pairing results

Figure 4.13: Results for initial population = 7 with IFW formula



(a) Lineplot of delta variation



(b) Piechart of food pairing results

Figure 4.14: Results for initial population = 9 with IFW formula

After examining the experimental results, we notice that the fitness values represented by the IFW have a slightly higher average difference between $Ns(real)$ and $Ns(rand)$ than the results of the normalised frequencies, but the two techniques reach the same conclusion when it comes to the tendencies of food pairing in the savory Algerian cuisine, presenting a negative leaning with a minimum percentage of 85. And for the few remaining cases that gave a positive difference, we can consider them as outliers, since they represent a very small proportion of the tests we performed.

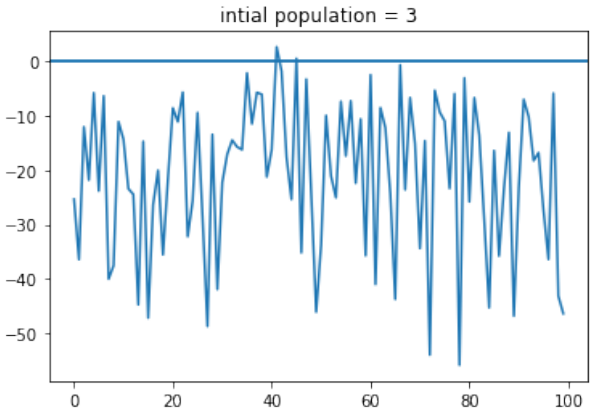
Pastry dataset

The table 4.11 summarizes the results for the pastry dataset in the different configurations tested, and figures 4.15, 4.16, 4.17 and 4.18 show the delta variations for the normalized frequency

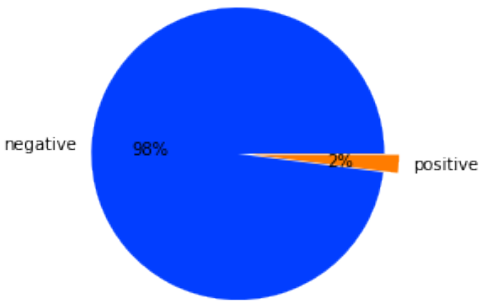
fitness values and figures 4.19, 4.20, 4.21 and 4.22 show the results for the IFW fitness values at each iteration:

		normalized frequency	observation	IFW	observation
pop 3	min delta	-55.8	negative food	-73.37	negative food
	max delta	2.57	pairing in 98%	-3.12	pairing in 100%
	avrg delta	-20.96	of the cases	-29.89	of the cases
pop 5	min delta	-55.37	negative food	-65.4	negative food
	max delta	4.04	pairing in 98%	3.53	pairing in 98%
	avrg delta	-21.03	of the cases	-26.29	of the cases
pop 7	min delta	-52.9	negative food	-57.55	negative food
	max delta	-1.1	pairing in 100%	-3.13	pairing in 100%
	avrg delta	-25.24	of the cases	-25.36	of the cases
pop 9	min delta	-57.8	negative food	-71.56	negative food
	max delta	1.37	pairing in 99%	0.4	pairing in 99%
	avrg delta	-24.06	of the cases	-27.54	of the cases

Table 4.11: Conclusion and comparison of the results of normalized frequency and IFW on different values for template size in algorithm 1 on pastry dataset

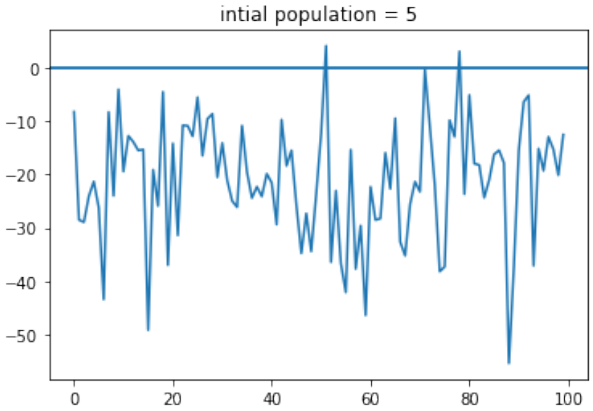


(a) Lineplot of delta variation

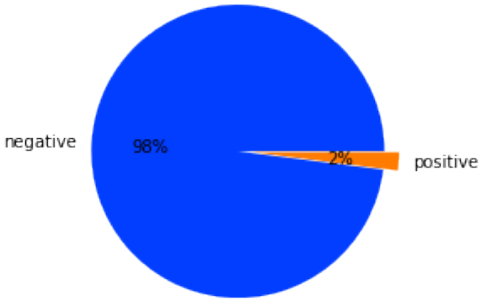


(b) Piechart of food pairing results

Figure 4.15: Results for initial population = 3 with FV normalized frequency

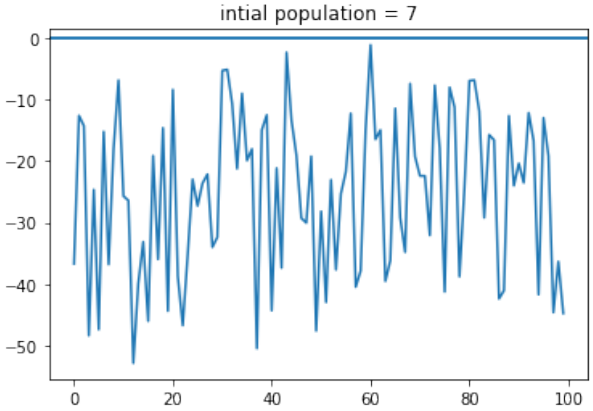


(a) Lineplot of delta variation

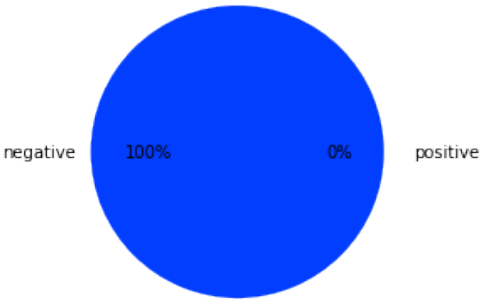


(b) Piechart of food pairing results

Figure 4.16: Results for initial population = 5 with FV normalized frequency

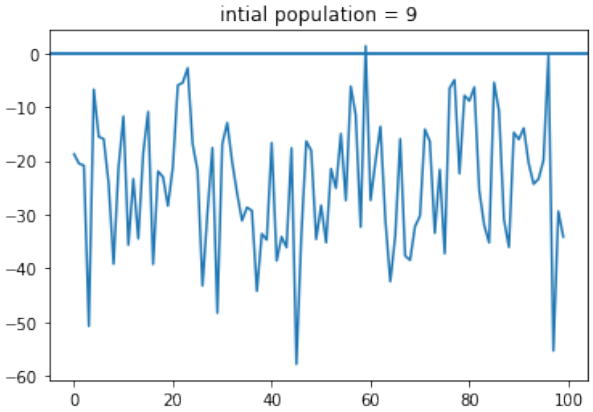


(a) Lineplot of delta variation

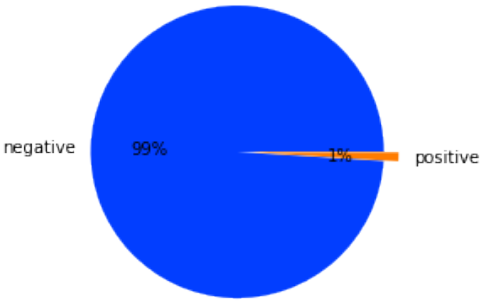


(b) Piechart of food pairing results

Figure 4.17: Results for initial population = 7 with FV normalized frequency

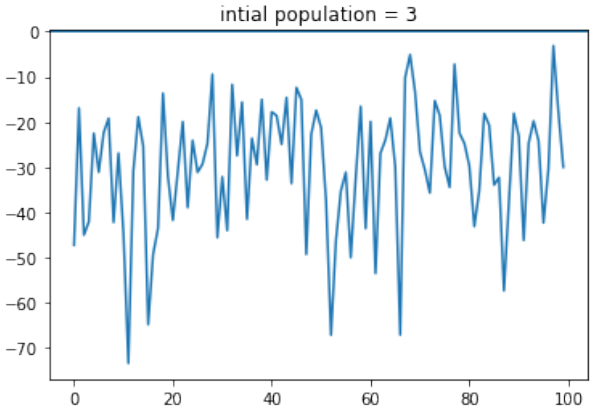


(a) Lineplot of delta variation

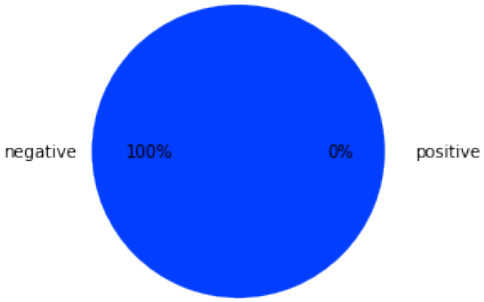


(b) Piechart of food pairing results

Figure 4.18: Results for initial population = 9 with FV normalized frequency

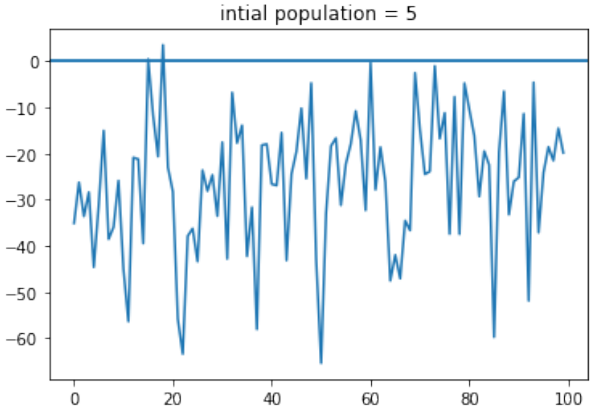


(a) Lineplot of delta variation

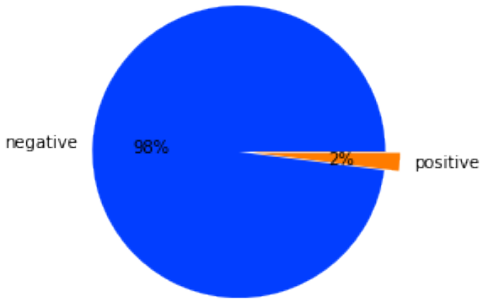


(b) Piechart of food pairing results

Figure 4.19: Results for initial population = 3 with IFW formula

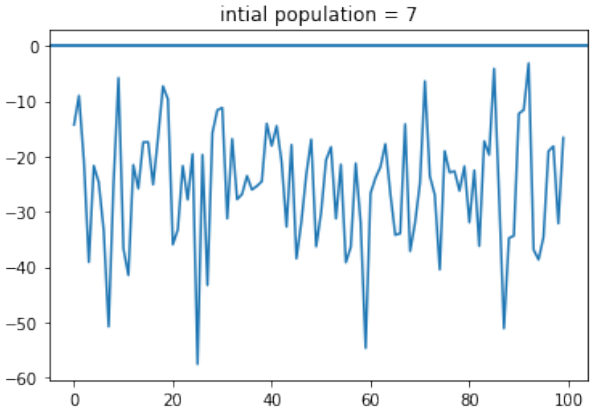


(a) Lineplot of delta variation

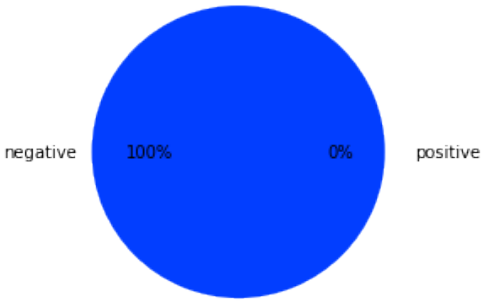


(b) Piechart of food pairing results

Figure 4.20: Results for initial population = 5 with IFW formula

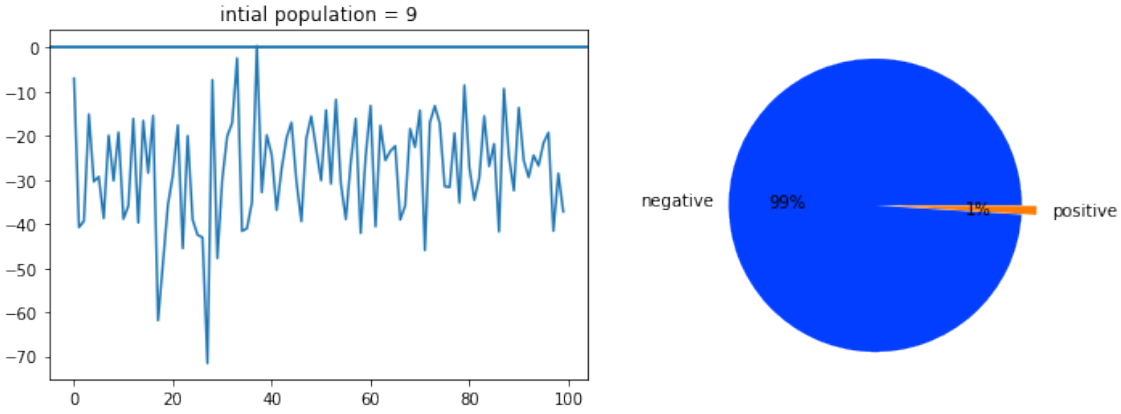


(a) Lineplot of delta variation



(b) Piechart of food pairing results

Figure 4.21: Results for initial population = 7 with IFW formula



(a) Lineplot of delta variation

(b) Piechart of food pairing results

Figure 4.22: Results for initial population = 9 with IFW formula

The experimental results on the pastry dataset show that when calculating the difference between Ns(real) and Ns(rand), the IFW fitness values tend to give slightly lower results than the normalised frequencies, but in all configurations tested, the conclusion is the same for both techniques, the algerian pastry has a negative food pairing.

4.6.3.2 Ingredient initial pool size parameter

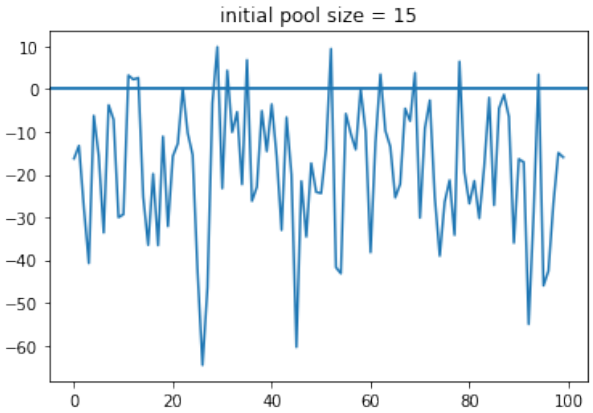
The previous tests show that regardless of the values assigned to the recipe template size, the food pairing is negative, so we performed test sets on the initial ingredient pool variable with two new values, 1.5*recipe size and 2*recipe size, in addition to the one we tested for the previous tests (3*recipe size), and by setting the template size to 5.

Savory food dataset

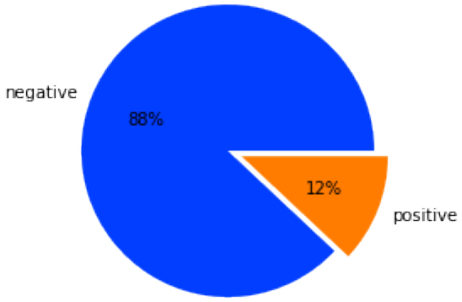
The table 4.12 and figures 4.23, 4.24, 4.25 and 4.26 display the results of the tests conducted on the savory food dataset.

		normalized frequency	observation	IFW	observation
Pool size = 15	min delta	-64.57	negative food	-46.54	negative food
	max delta	9.83	pairing in 88%	12.40	pairing in 93%
	avrg delta	-18.50	of the cases	-16.57	of the cases
Pool size = 20	min delta	-50.77	negative food	-48.78	negative food
	max delta	14.34	pairing in 96%	9.33	pairing in 89%
	avrg delta	-18.05	of the cases	-14.52	of the cases

Table 4.12: Conclusion and comparison of the results of normalized frequency and IFW on different values for initial pool in algorithm 1 on savory dataset

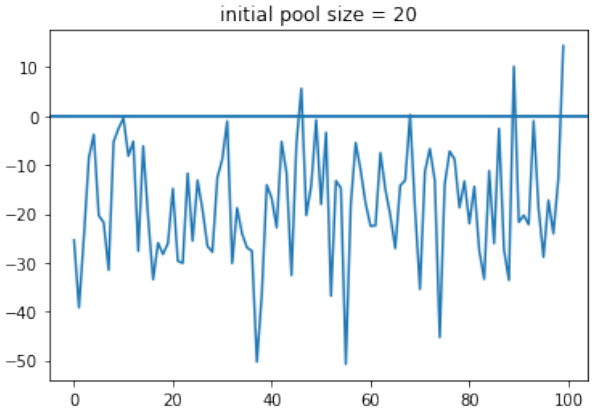


(a) Lineplot of delta variation

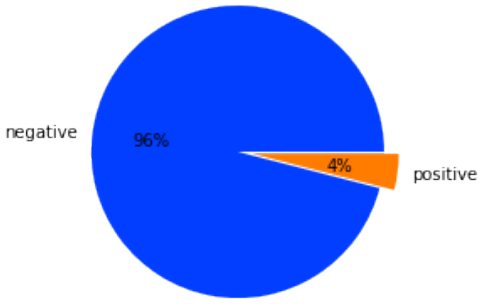


(b) Piechart of food pairing results

Figure 4.23: Results for initial pool size = 15 with FV normalized frquency

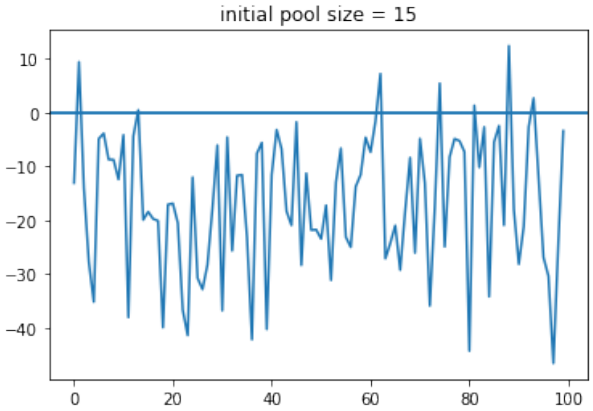


(a) Lineplot of delta variation

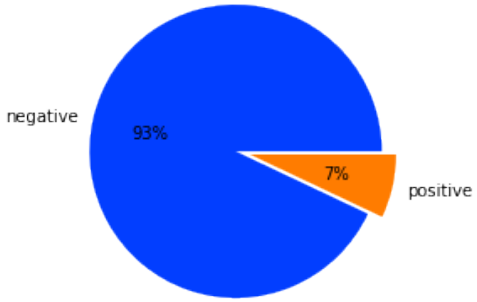


(b) Piechart of food pairing results

Figure 4.24: Results for initial pool size = 20 with FV normalized frquency

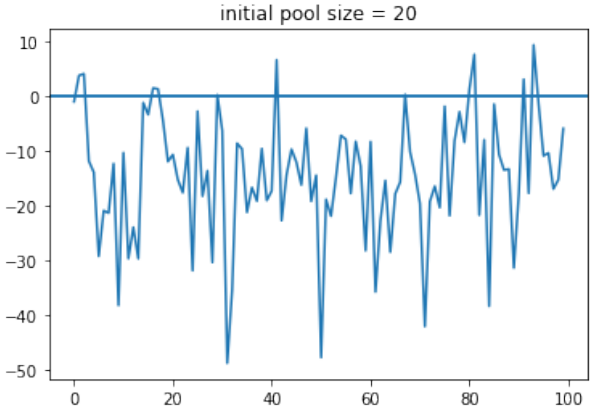


(a) Lineplot of delta variation

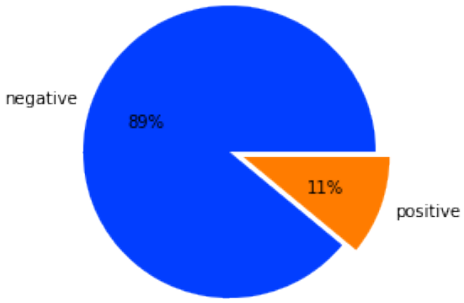


(b) Piechart of food pairing results

Figure 4.25: Results for initial pool size = 15 with IFW formula



(a) Lineplot of delta variation



(b) Piechart of food pairing results

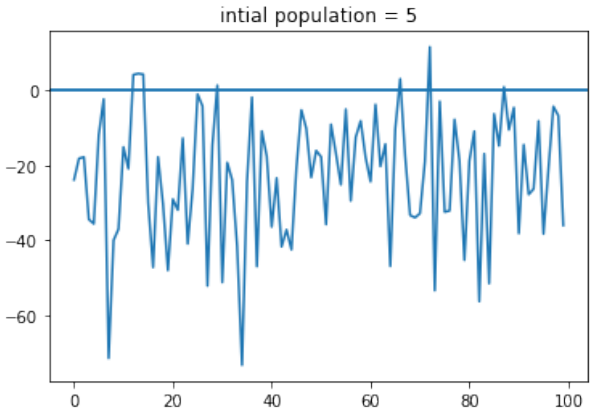
Figure 4.26: Results for initial pool size = 20 with IFW formula

Pastry dataset

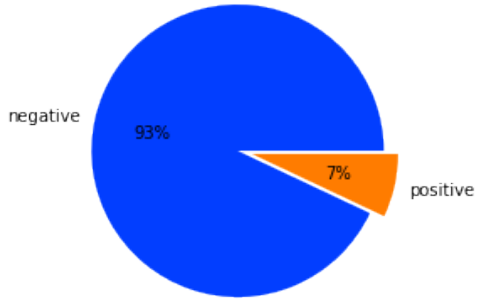
The table 4.13 and figures 4.27, 4.28, 4.29 and 4.30 display the results of the tests conducted on the savory food dataset.

		normalized frequency	observation	IFW	observation
Pool size = 9	min delta	-73.39	negative food	-81.72	negative food
	max delta	11.62	pairing in 93%	4.53	pairing in 98%
	avrg delta	-22.93	of the cases	-29.67	of the cases
Pool size = 12	min delta	-52.35	negative food	-75.57	negative food
	max delta	8.93	pairing in 96%	-0.86	pairing in 100%
	avrg delta	-19.81	of the cases	-26.74	of the cases

Table 4.13: Conclusion and comparison of the results of normalized frequency and IFW on different values for initial pool in algorithm 1 on pastry dataset

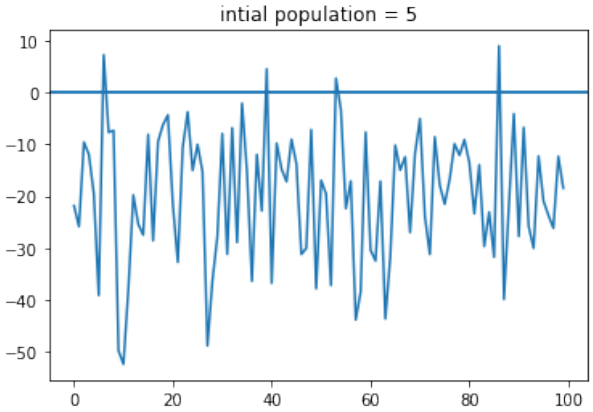


(a) Lineplot of delta variation

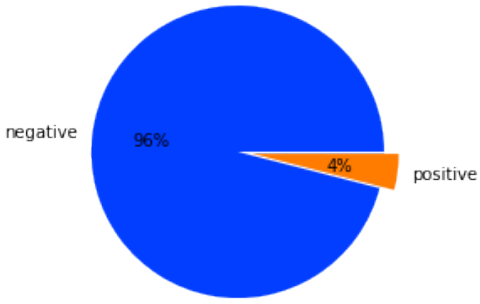


(b) Piechart of food pairing results

Figure 4.27: Results for initial pool size = 9 with FV normalized frequency

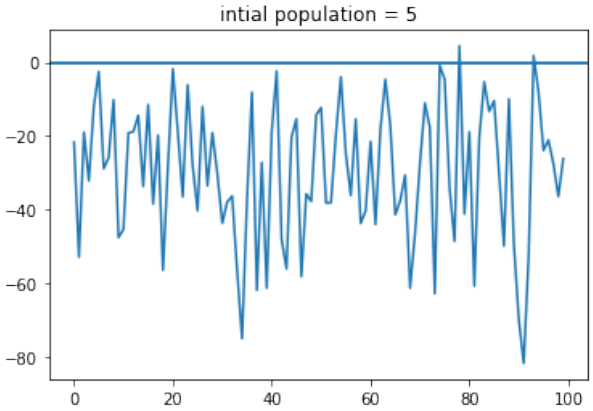


(a) Lineplot of delta variation

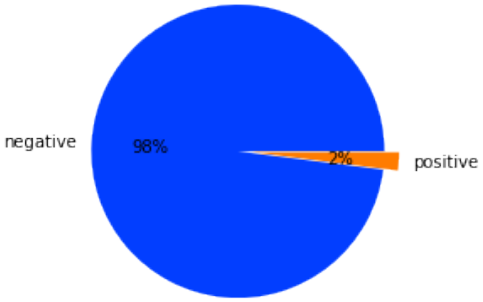


(b) Piechart of food pairing results

Figure 4.28: Results for initial pool size = 12 with FV normalized frequency

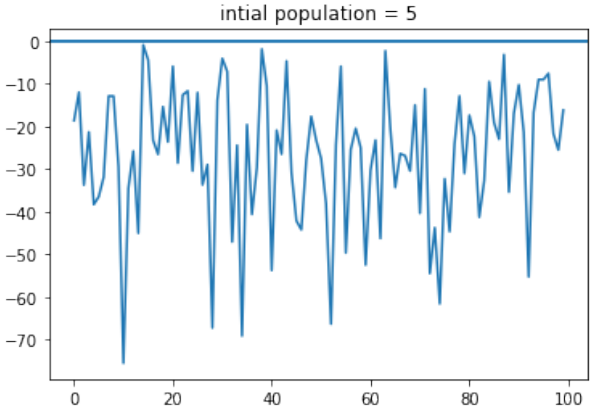


(a) Lineplot of delta variation

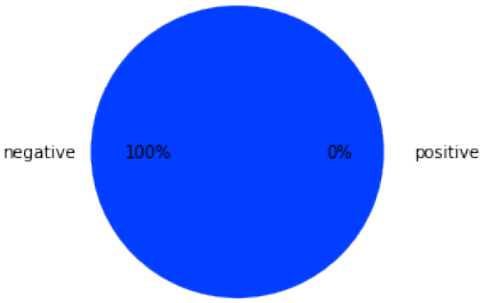


(b) Piechart of food pairing results

Figure 4.29: Results for initial pool size = 9 with IFW formula



(a) Lineplot of delta variation



(b) Piechart of food pairing results

Figure 4.30: Results for initial pool size = 12 with IFW formula

4.6.4 Experiments using Algorithm 2

We went through the same stages with the second algorithm, testing different configurations of the algorithm's variables, namely the recipe's template size and the ingredient initial pool size, on both savory food and pastry datasets.

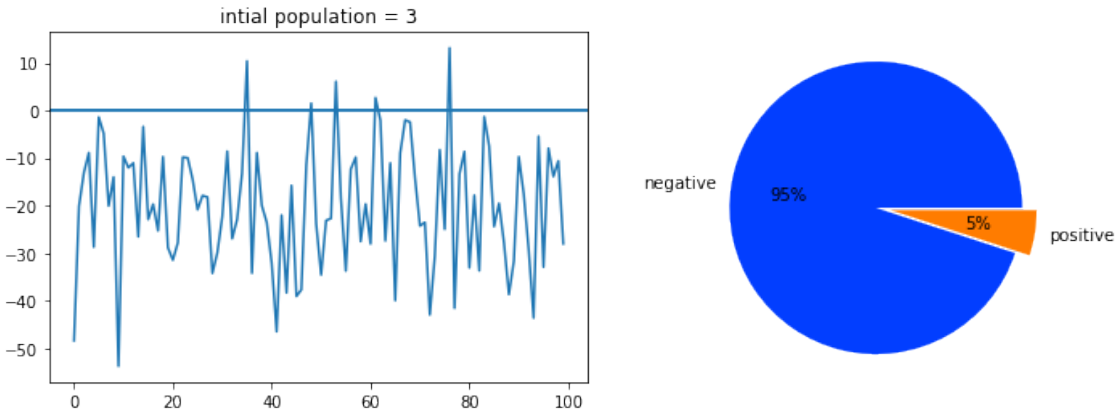
4.6.4.1 Recipes' template size parameter

Savory dataset

The table 4.14 and figures 4.31, 4.32 4.33, ‘4.34, 4.35, 4.36, 4.37 and 4.38 show the results of the tests ran on the savory dataset.

		normalized fequency	Observation	IFW	Observation
pop 3	min delta	-53.78	negative food	-49.54	negative food
	max delta	13.16	pairing in 95 %	13.44	pairing in 89%
	avrg delta	-20.01	of the cases	-14.85	of the cases
pop 5	min delta	-52.51	negative food	-46.49	negative food
	max delta	7.74	pairing in 94%	5.55	pairing in 92%
	avrg delta	-16.39	of the cases	-17.59	of the cases
pop 7	min delta	-45.11	negative food	-41.57	negative food
	max delta	1.85	pairing in 98%	9.69	pairing in 94%
	avrg delta	-19.89	of the cases	-17.28	of the cases
pop 9	min delta	-41.92	negative food	-42.6	negative food
	max delta	2.05	pairing in 98%	7.37	pairing in 92%
	avrg delta	-17.72	of the cases	-16.47	of the cases

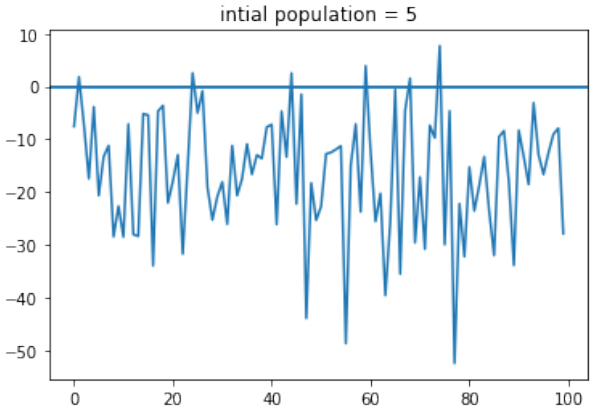
Table 4.14: Conclusion and comparison of the results of normalized frequency and IFW on different values for template size in algorithm 2 on savory dataset



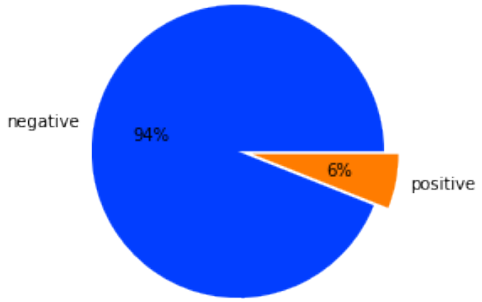
(a) Lineplot of delta variation

(b) Piechart of food pairing results

Figure 4.31: Results for initial population = 3 with FV normalized frquency

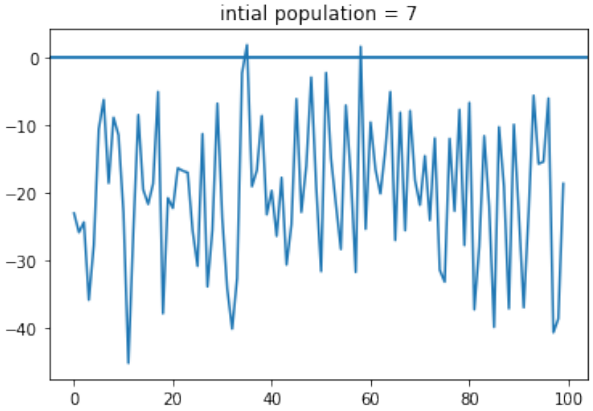


(a) Lineplot of delta variation

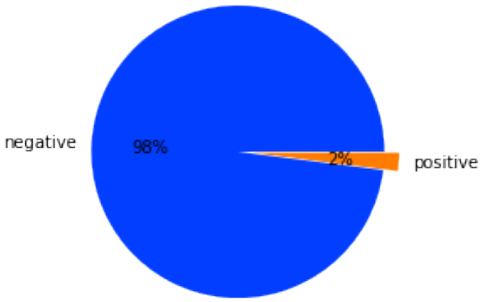


(b) Piechart of food pairing results

Figure 4.32: Results for initial population = 5 with FV normalized frequency

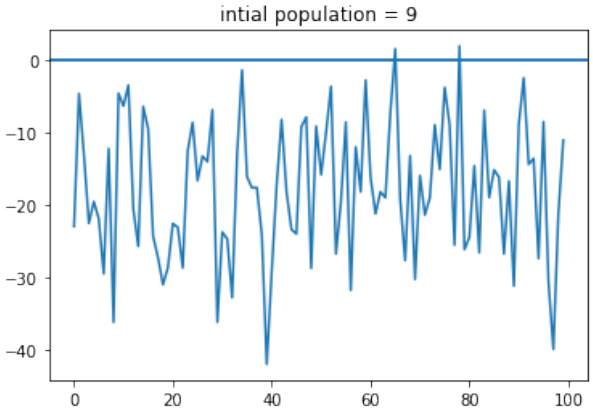


(a) Lineplot of delta variation

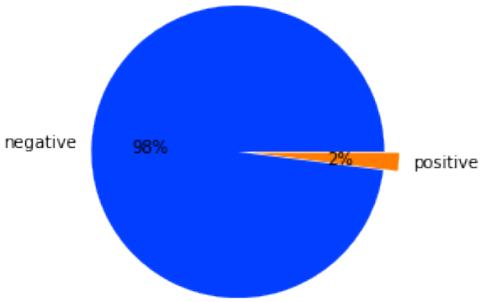


(b) Piechart of food pairing results

Figure 4.33: Results for initial population = 7 with FV normalized frequency

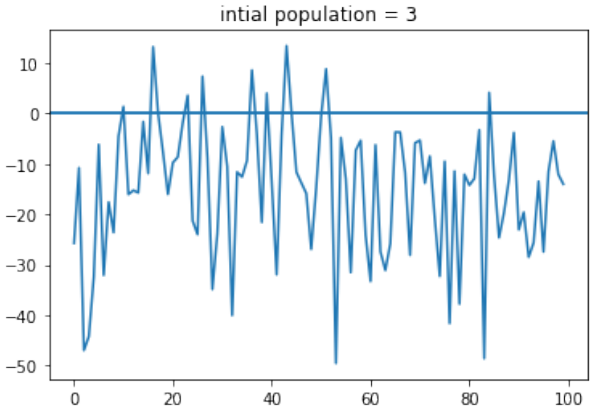


(a) Lineplot of delta variation

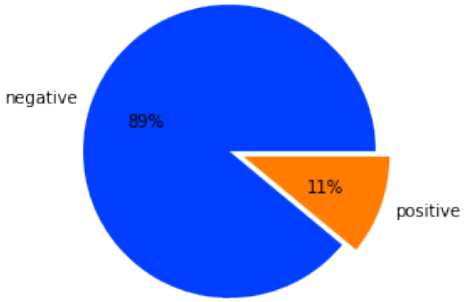


(b) Piechart of food pairing results

Figure 4.34: Results for initial population = 9 with FV normalized frequency

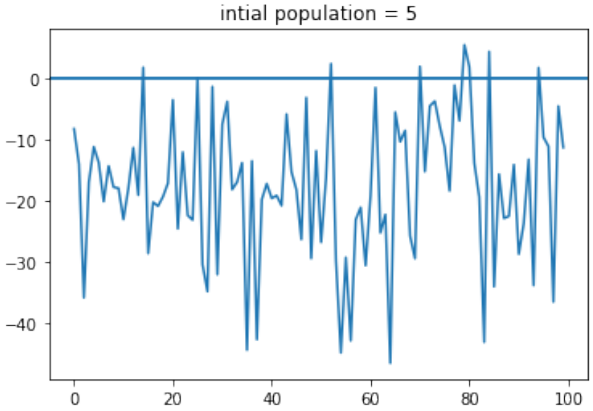


(a) Lineplot of delta variation

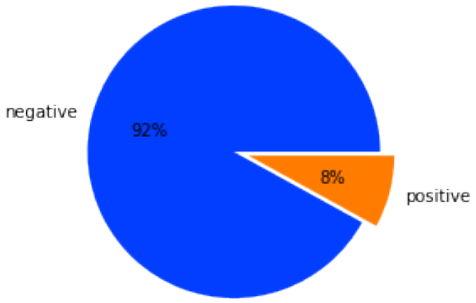


(b) Piechart of food pairing results

Figure 4.35: Results for initial population = 3 with IFW formula

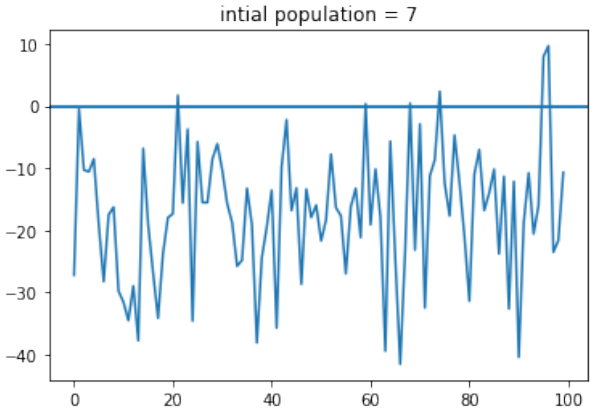


(a) Lineplot of delta variation

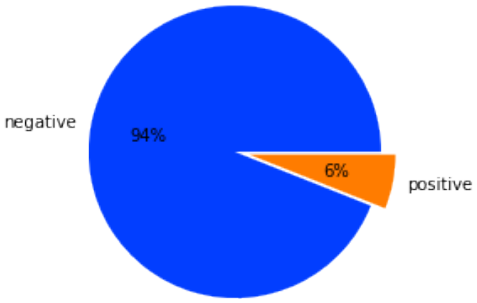


(b) Piechart of food pairing results

Figure 4.36: Results for initial population = 5 with IFW formula

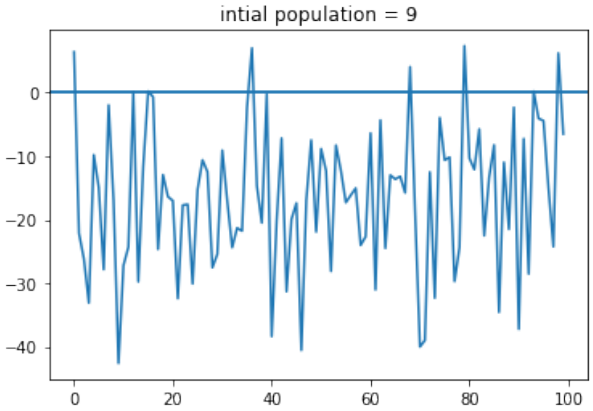


(a) Lineplot of delta variation

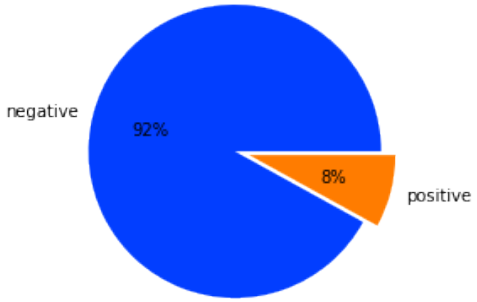


(b) Piechart of food pairing results

Figure 4.37: Results for initial population = 7 with IFW formula



(a) Lineplot of delta variation



(b) Piechart of food pairing results

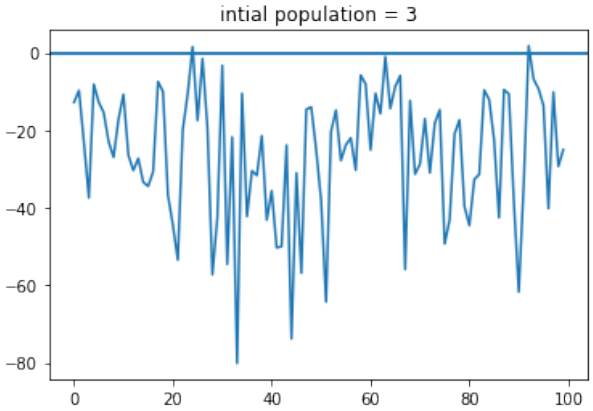
Figure 4.38: Results for initial population = 9 with IFW formula

Pastry dataset

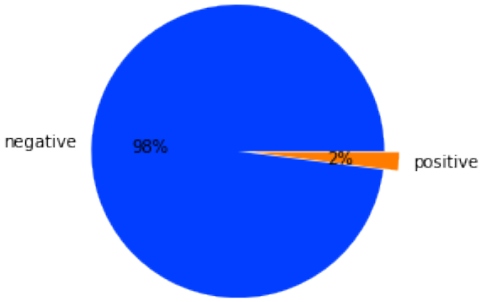
The table 4.15 and figures 4.39, 4.40, 4.41, 4.42, 4.43, 4.44, 4.45 and 4.46 show the results ran on the pastry dataset.

		normalized fequency	Observation	IFW	Observation
pop 3	min delta	-80.11	negative food	-63.31	negative food
	max delta	1.76	pairing in 98 %	2.40	pairing in 98%
	avrg delta	-26.13	of the cases	-24.54	of the cases
pop 5	min delta	-61.87	negative food	-71.46	negative food
	max delta	2.80	pairing in 98%	-1.51	pairing in 100%
	avrg delta	-27.05	of the cases	-29.17	of the cases
pop 7	min delta	-62.33	negative food	-78.39	negative food
	max delta	4.20	pairing in 98%	-4.24	pairing in 100%
	avrg delta	-27.56	of the cases	-28.74	of the cases
pop 9	min delta	-55.88	negative food	-71.85	negative food
	max delta	-2.06	pairing in 100%	-6.46	pairing in 92%
	avrg delta	-27.63	of the cases	-29.64	of the cases

Table 4.15: Conclusion and comparison of the results of FV normalized frequency and IFW on different values for template size in algorithm 2 on pastry dataset

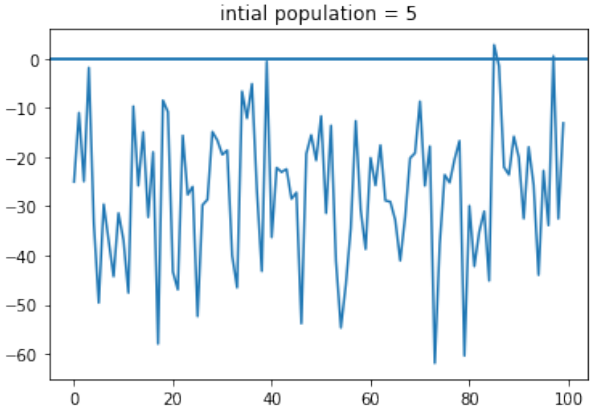


(a) Lineplot of delta variation

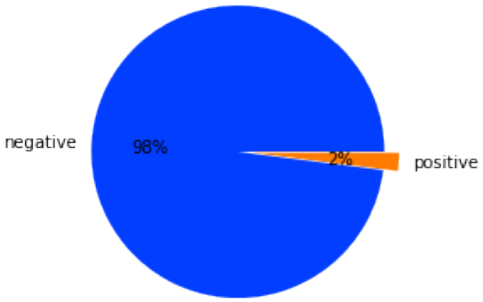


(b) Piechart of food pairing results

Figure 4.39: Results for initial population = 3 with FV normalized frequency

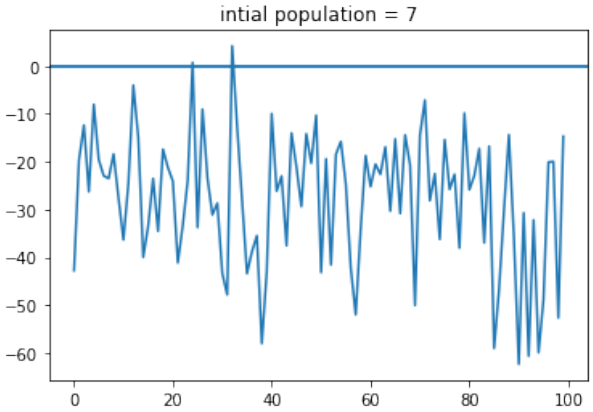


(a) Lineplot of delta variation

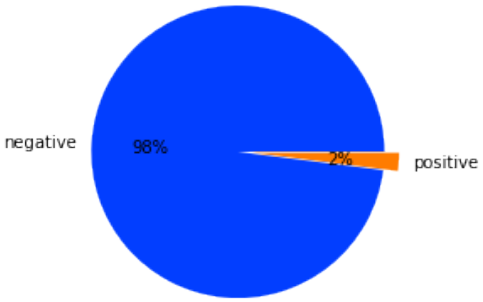


(b) Piechart of food pairing results

Figure 4.40: Results for initial population = 5 with FV normalized frequency

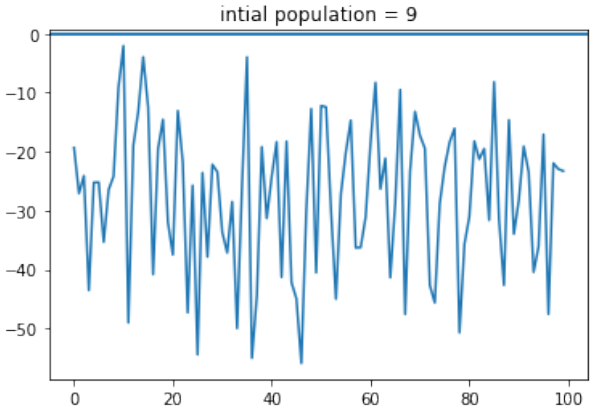


(a) Lineplot of delta variation

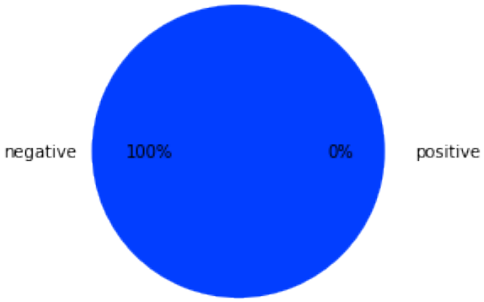


(b) Piechart of food pairing results

Figure 4.41: Results for initial population = 7 with FV normalized frequency

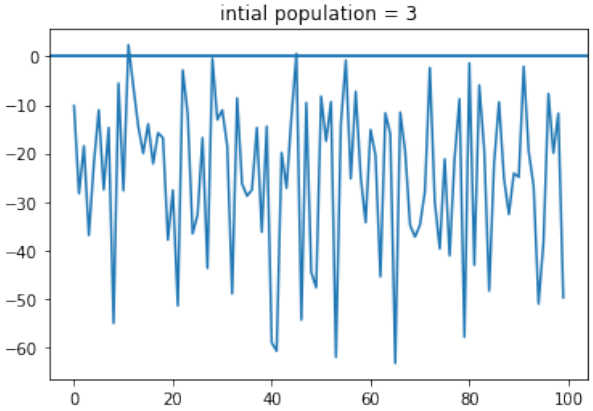


(a) Lineplot of delta variation

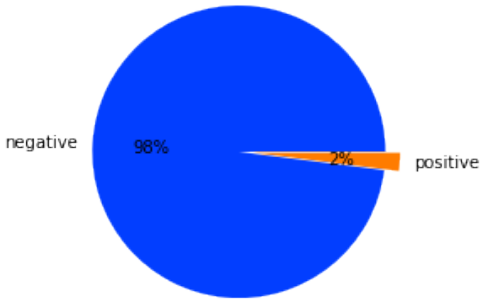


(b) Piechart of food pairing results

Figure 4.42: Results for initial population = 9 with FV normalized frequency

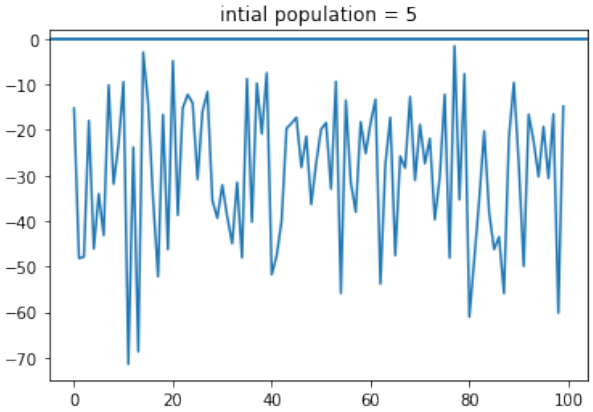


(a) Lineplot of delta variation

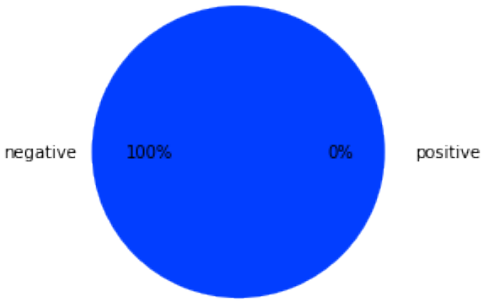


(b) Piechart of food pairing results

Figure 4.43: Results for initial population = 3 with IFW fromula

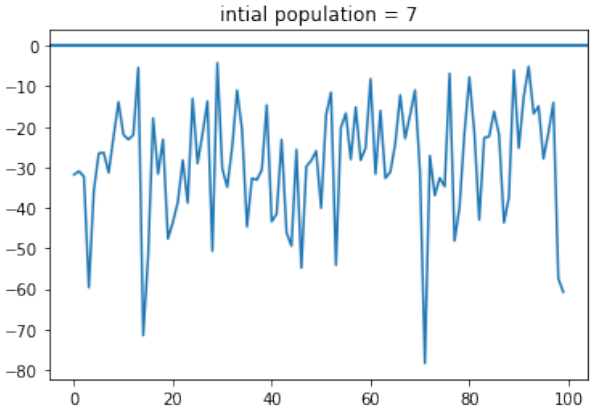


(a) Lineplot of delta variation

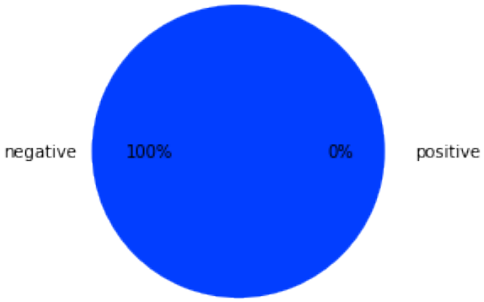


(b) Piechart of food pairing results

Figure 4.44: Results for initial population = 5 with IFW fromula

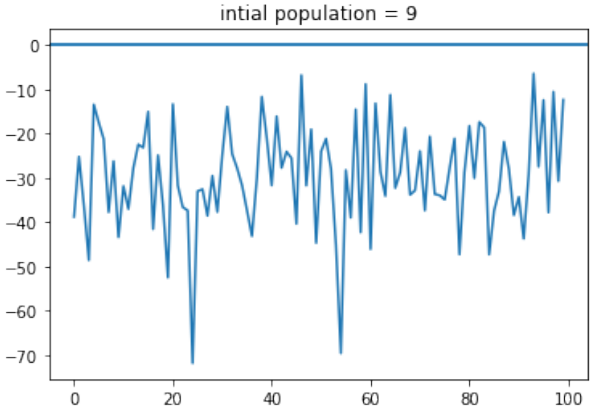


(a) Lineplot of delta variation

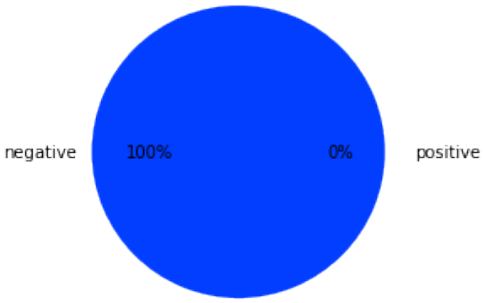


(b) Piechart of food pairing results

Figure 4.45: Results for initial population = 7 with IFW fromula



(a) Lineplot of delta variation



(b) Piechart of food pairing results

Figure 4.46: Results for initial population = 9 with IFW fromula

4.6.4.2 Ingredients initial pool size

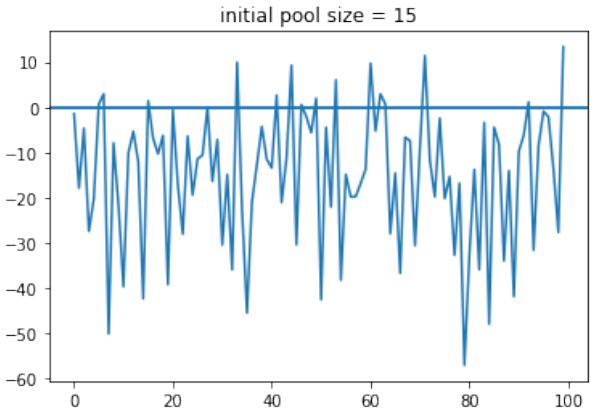
The table 4.17 and figures 4.51, 4.52, 4.53 and 4.54 show the results ran on the pastry dataset using different initial pool size.

Savory food dataset

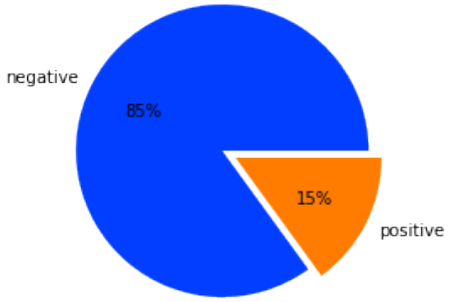
The table 4.16 and figures 4.47, 4.48, 4.49 and 4.50 show the results ran on the savory food dataset.

		normalized fequency	observation	IFW	observation
pool size = 15	min delta	-57.16	negative food	-61.68	negative food
	max delta	13.34	pairing in 85%	11.47	pairing in 89%
	avrg delta	-14.96	of the cases	-18.45	of the cases
pool size = 20	min delta	-49.42	negative food	-46.85	negative food
	max delta	14.36	pairing in 89%	9.05	pairing in 86%
	avrg delta	-16.71	of the cases	-13.67	of the cases

Table 4.16: Conclusion and comparison of the results of FV normalized frequency and IFW on different values for initial pool in algorithm 2 on savory dataset

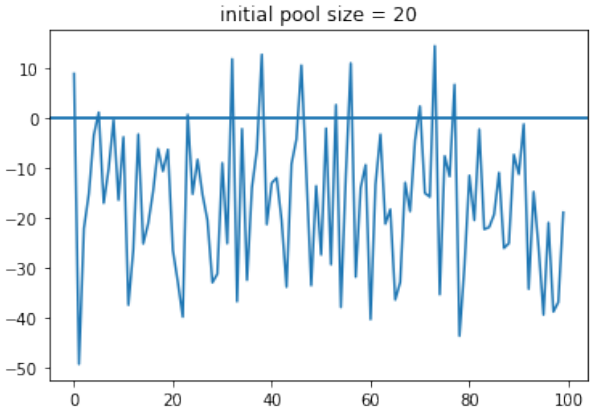


(a) Lineplot of delta variation

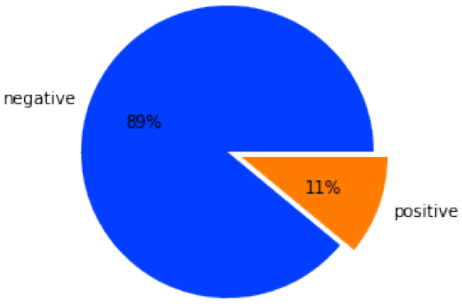


(b) Piechart of food pairing results

Figure 4.47: Results for initial pool size = 15 with normalized frequency

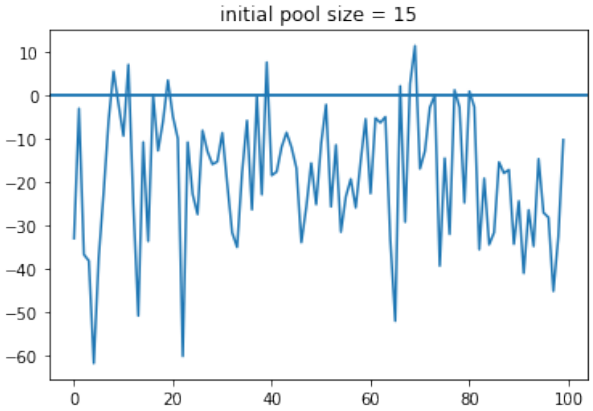


(a) Lineplot of delta variation

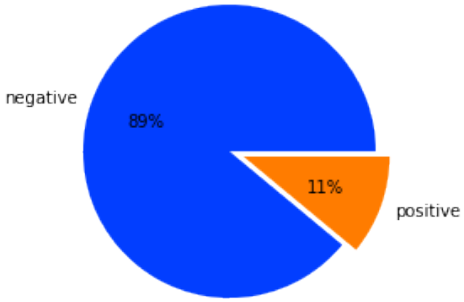


(b) Piechart of food pairing results

Figure 4.48: Results for initial pool size = 20 with normalized frequency

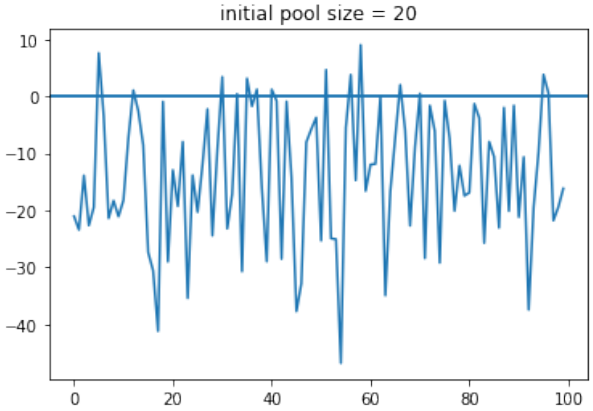


(a) Lineplot of delta variation

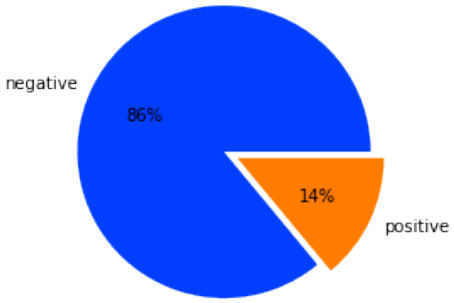


(b) Piechart of food pairing results

Figure 4.49: Results for initial pool size = 15 with IFW fromula



(a) Lineplot of delta variation



(b) Piechart of food pairing results

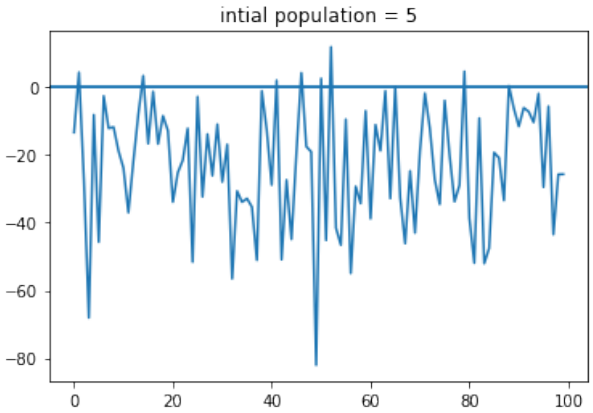
Figure 4.50: Results for initial pool size = 20 with IFW fromula

Pastry dataset

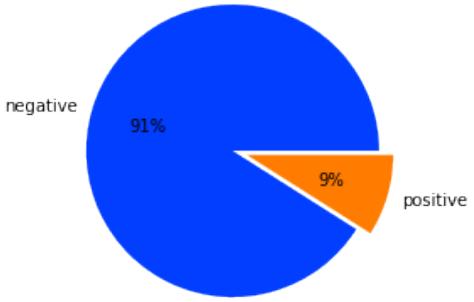
The table 4.17 and figures 4.51, 4.52, 4.53 and 4.54 display the results of the tests conducted on the pastry dataset.

		normalized fequency	observation	IFW	observation
pool size = 9	min delta	-82.04	negative food	-86.00	negative food
	max delta	11.84	pairing in 91%	8.84	pairing in 97%
	avrg delta	-22.88	of the cases	-28.28	of the cases
pool size = 12	min delta	-92.31	negative food	-48.78	negative food
	max delta	3.14	pairing in 98%	9.33	pairing in 95%
	avrg delta	-26.05	of the cases	-14.52	of the cases

Table 4.17: Conclusion and comparison of the results of FV normalized frequency and IFW on different values for initial pool in algorithm 2 on pastry dataset

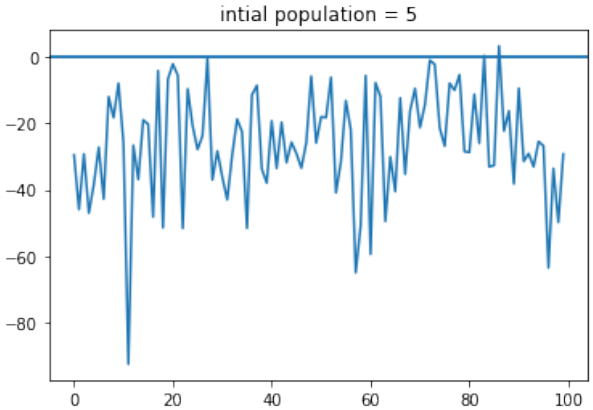


(a) Lineplot of delta variation

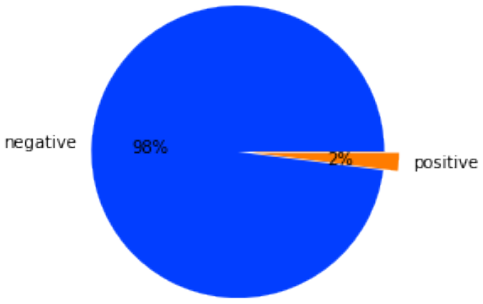


(b) Piechart of food pairing results

Figure 4.51: Results for initial pool size = 9 with FV normalized frequency

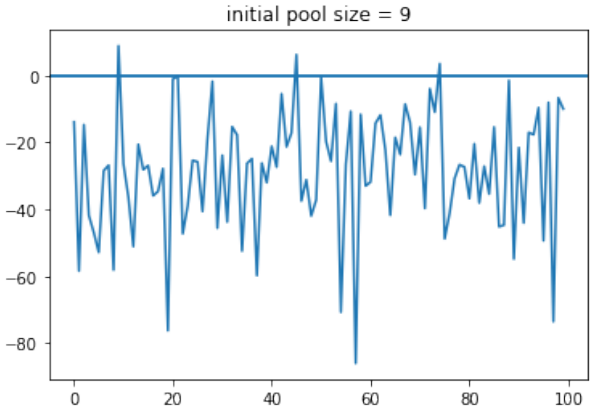


(a) Lineplot of delta variation

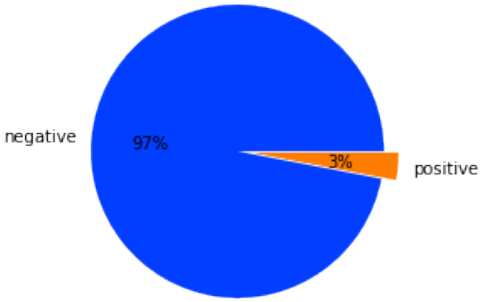


(b) Piechart of food pairing results

Figure 4.52: Results for initial pool size = 12 with FV normalized frequency

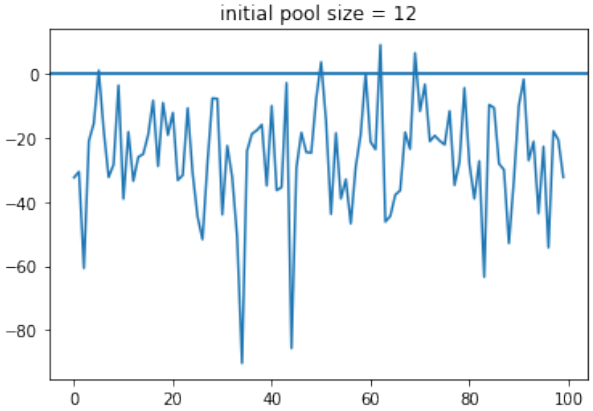


(a) Lineplot of delta variation

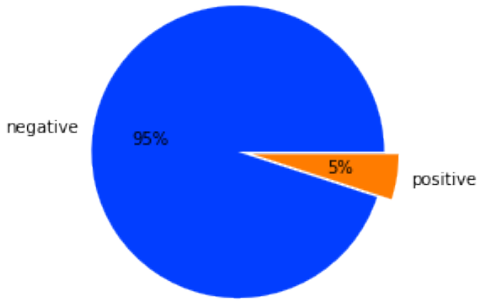


(b) Piechart of food pairing results

Figure 4.53: Results for initial pool size = 9 with FV normalized frequency



(a) Lineplot of delta variation



(b) Piechart of food pairing results

Figure 4.54: Results for initial pool size = 12 with FV normalized frequency

4.6.5 Discussion

The subject of our thesis is the food pairing hypothesis, which proposes that humans tend to prefer in their cooking the association of ingredients that share as many chemical components as possible, and more precisely flavor profiles. Through our research we investigated the question to confirm or refute the hypothesis for Algerian cuisine.

Although food pairing is a relatively new field, some scientists have already investigated the proposition for several world cuisines using various approaches: (North American, Western European, Southern European, Latin American, and East Asian cuisines were studied using bipartite graphs in the article by Ahn et al. [2], Indian cuisine was studied in the article by Jain

et al. [22] and Saudi Arabian cuisine was discussed in the article by Al-Razgan et al. [1]. This discipline, however, has yet to be introduced in North Africa, and to the best of our knowledge, this work is the first to tackle the subject with Algerian cuisine as the subject of study.

We can refute the food pairing hypothesis for Algerian cuisine and conclude that it is negative, implying that the ingredients most commonly combined in this region do not share many chemical components.

Our research is based on authentic recipes from a well-known book, and we used two algorithms to perform several tests on their variables in order to obtain reliable results. Nonetheless, we ran into some issues due to the lack of flavor profiles for certain Algerian ingredients, such as "la tete de mouton," which forced us to delete the recipes that used these ingredients. We also note that our work does not cover all of Algeria in a diverse manner, and that there is a significant lack of Sub-Saharan cuisine, owing to the limited documentation available on the latter.

At first glance, one might presume that the century-long French occupation of Algeria has influenced our cuisine, and thus the results would be positive, similar to the trends in Western Europe, of which France is a part. However, contrary to this initial belief, one can conclude that the Algerian cuisine and French cuisine have different tendencies. The second important observation that our results reveal is their similarity with the conclusion reached by the article by Ahn et al. [2] concerning South European cuisine, which has a negative food pairing. This rapprochement in the results indicates that geographical proximity represents an important and determining factor in the evolution of cuisines in nearby regions.

4.7 Algerian food dashboard

To accompany our research process and to help demonstrate our work, we have built a Dashboard using a python framework called Streamlit, which would allow users to interact with the dataset and experiment with different values, as well as visualize the results

4.7.1 Features

This tool supports multiple functionalities divided into two main parts:

4.7.1.1 Dataset overview

This section helps the user explore the dataset, it includes:

- **Support for custom file uploads**: it only includes the CSV format for now.
- **Co-occurrence graph**: each node of the graph represents an ingredient, and each edge is weighted according to the value of co-occurrence of these ingredients, i.e. the higher the value of co-occurrence, the thicker the edge would be.
- **Summary statistics**: it includes statistical values about recipes, such as the minimum, maximum and average length ...
- **Bar chart**: which gives additional information about ingredients and their frequency.
- Average flavor sharing calculation.

4.7.1.2 Dataset Generation

This page is to replicate the generation process in the food pairing hypothesis. It includes:

- Control for different parameters, such as:
  - Number of iterations
  - Number of ingredients in the initial pool
  - Recipe length
  - Number of initial templates
- Line chart with mark points
- Food pairing calculation

4.7.2 User Guide

When opening the tool it will prompt the user to the home screen, and the user has the option to choose between Dataset overview or Data generation. Figure 4.55 shows the home screen of the dashboard.

Figure 4.55: Home screen of dashboard

When clicking on the dataset overview, an option to upload a CSV file will be displayed, which corresponds to the Dataset, and on the sidebar, there are option to choose the number of top co-occurrences to plot, summary statistics, a bar chart for ingredient frequency, and average flavor sharing for the chosen dataset as shown in 4.56.

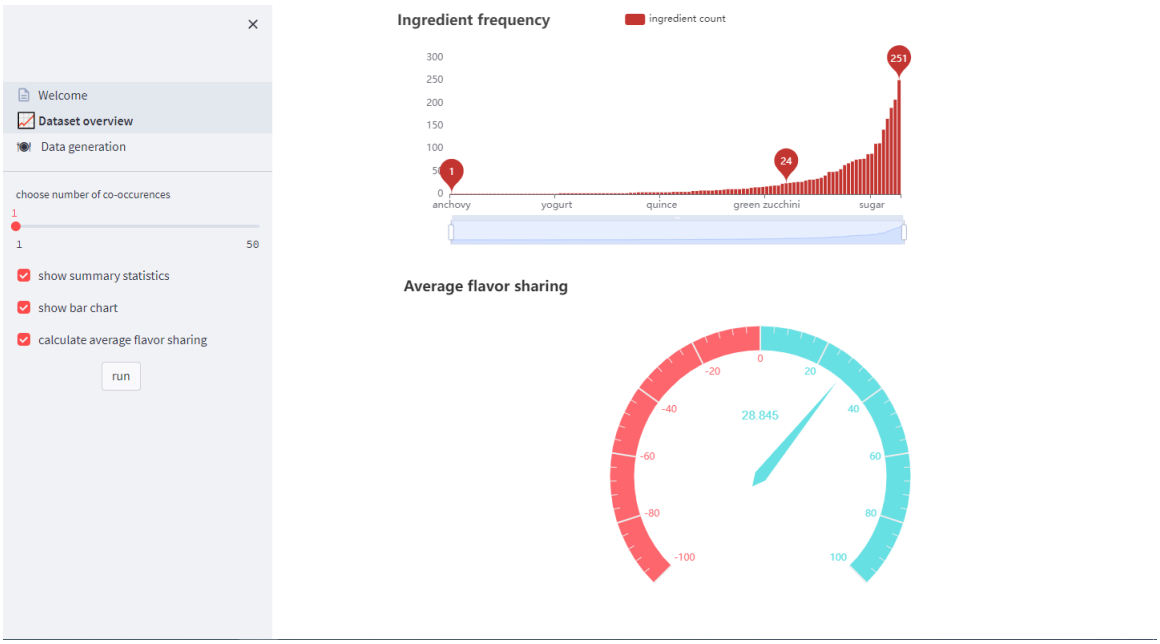


Figure 4.56: Screen capture of dataset overeview

For the data generation aspect, we have multiple variables to control the generation process,

as well as a line chart to display the different values of the average flavor sharing as illustrated in figure 4.57 and figure 4.58.

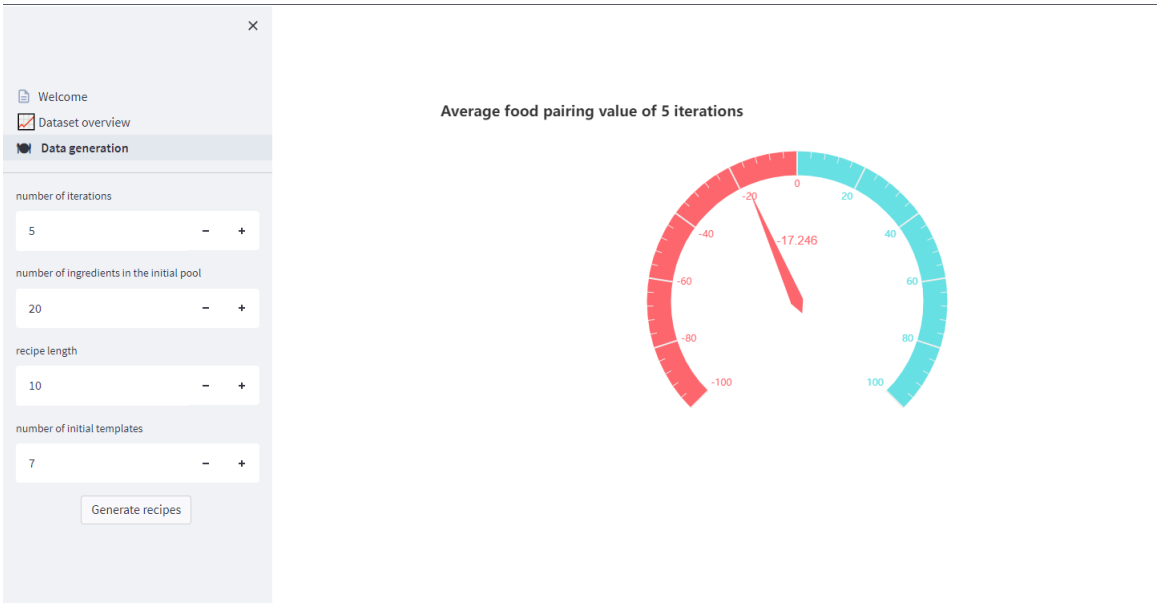


Figure 4.57: Screen capture of the food pairing results

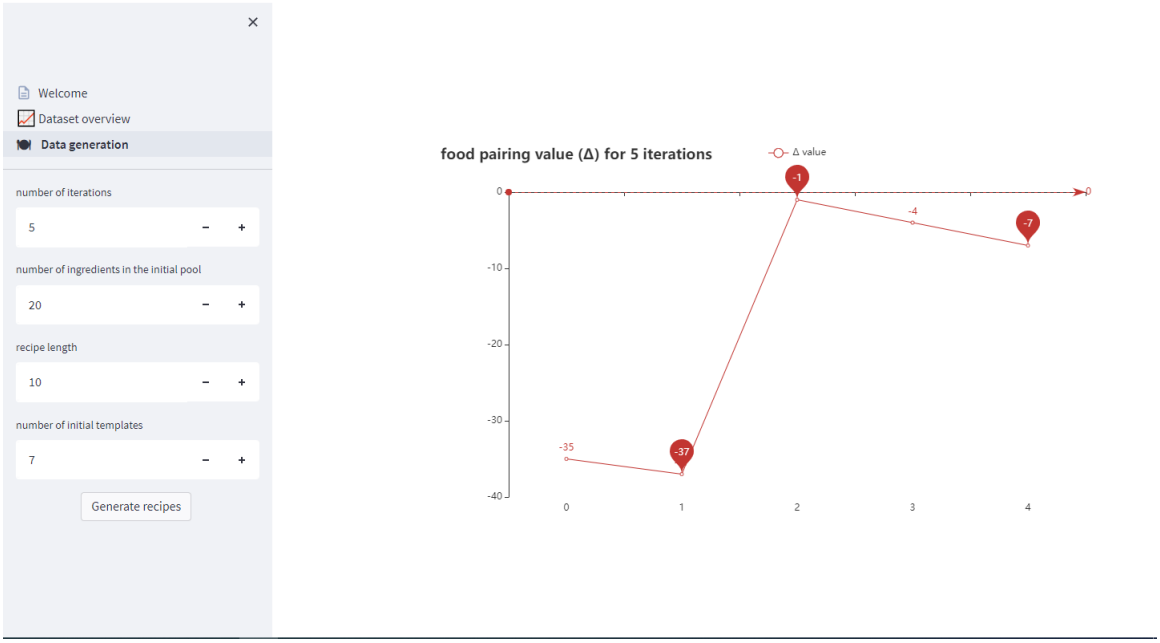


Figure 4.58: Screen capture of the food pairing results with variation of 5 iterations

4.8 Conclusion

In this chapter, we saw all of the steps we took to implement our solution to verify the food pairing hypothesis in Algerian cuisine. Beginning with our working environment and tools, then

the process of collecting the corpus of Algerian recipes and the pre-processing methods we used, then all of the tests we performed to ensure the reliability of our conclusion, and finally we discussed the different configuration of the proposed approach, the results obtained, and their implications.

Genral Conclusion

This thesis aimed to introduce the field of food pairing into Algerian gastronomy by identifying the nature of food pairing that distinguishes it based on the blumenthal hypothesis, and by asking what this reveals about the identity and influences of Algerian cuisine.

Our study came to a conclusion, proving that Algerian cuisine has a negative food pairing trend, i.e. the ingredients frequently associated in recipes do not share many aromatic molecules. We also observed that the trend of southern European cuisine (European Mediterranean countries) is negative, indicating that the culinary culture and preferences of these two neighbouring regions are similar.

To reach this conclusion, we collected Algerian recipes from a reference book, then used natural language processing methods to extract the unique ingredients of each recipe and associate them with a flavour profile, which consisted of the aromatic molecules responsible for their flavours, allowing us to study the chemical components shared by the ingredients. We used a genetic algorithm to optimise the search in order to study the space of the generated components.

Perspectives

Our conclusion answers the basic hypothesis of food pairing, and opens doors to innovative perspectives in Algerian gastronomy. The answer to this question is a key to a new alliance between Algerian chefs and scientists to inaugurate the universe of unlikely associations. To begin with, with a large volume of ingredients and infinite combinations to be explored in search of ideal pairs, in our case ingredients with the fewest chemical components in common, computer scientists can intervene to apply deep learning techniques in order to facilitate the process, so that chefs can then exploit this information to elaborate a dish with the right dosage and balance

on the gustatory and sensory levels, because most importantly , we should not ignore the fact that cooking is an art and a sensitive and human expression, and scientific rationality should not make gastronomy deviate from it.

It should be noted that the work we have done is mainly focused on the computational aspect of food pairing which explains our inability to address the problem of the lack of flavour profile of certain ingredients, however it would be interesting to collaborate with chemists to remedy this lack by extracting the flavor molecules of these ingredients and formulating their flavor profile. Finally, one of the other perspectives that our research provides is an examination of the supposition that geographical proximity is important in the formation of nearby region cuisines, extending the study to the rest of the Mediterranean countries, primarily Moroccan and Tunisian cuisine.

Bibliography

[1] Muna Al-Razgan, Shahad Tallab, and Taha Alfaqih. Exploring the Food Pairing Hypothesis in Saudi Cuisine Using Genetic Algorithm. *Mathematical Problems in Engineering*, 2021:1–16, December 2021.

[2] Yong-Yeol Ahn, Sebastian E. Ahnert, James P. Bagrow, and Albert-László Barabási. Flavor network and the principles of food pairing. Technical Report arXiv:1111.6074, arXiv, November 2011. arXiv:1111.6074 [physics] type: article.

[3] File:Computational.science.Genetic.algorithm.Crossover.Cut.and.Splice.svg - Wikimedia Commons last visit: 25-05-2022. https://commons.wikimedia.org/wiki/File:Computational.science.Genetic.algorithm.Crossover.Cut.and.Splice.svg.

[4] Omar Rifki. *A Study on Robust Structures Discovery and its Economic Applications : Combination of Algorithm and Simulation*. PhD thesis, March 2016.

[5] Ming-Shen Jian, Ta-Yuan Chou, Kun-Sian Sie, and Long-Yeu Chung. Adaptive Life-Cycle and Viability based Paramecium- Imitated Evolutionary Algorithm. *WSEAS Transactions on Computers*, 8:1358–1367, January 2009.

[6] Matthew Sadiku, Sarhan Musa, and Tolulope Joshua Ashaolu. Food Industry: An Introduction. *International Journal of Trend in Scientific Research and Development*, Volume-3:128–130, June 2019.

[7] Heston Blumenthal. *The Fat Duck Cookbook*. Bloomsbury USA, London, 8981st edition edition, October 2009.

[8] Best of the Best |The Fat Duck, last visit: 22-06-2022. https://www.theworlds50best.com/awards/best-of-the-best/the-fat-duck.html.

[9] Ryan Sutton. The World's 50 Best Restaurants 2016: The Full List of Winners, last visit: 22-06-2022. https://www.eater.com/2016/6/13/11923536/worlds-50-best-restaurants-2016, June 2016.

[10] By Barry Neild CNN. World's 50 best restaurants for 2018, last visit: 22-06-2022. https://www.cnn.com/travel/article/worlds-best-restaurants-2018/index.html.

[11] Research about Moleculer Cuisine Application as an Innovation Example in Istanbul Restaurants. *Procedia - Social and Behavioral Sciences*, 195:446–452, July 2015. Publisher: Elsevier.

[12] Dauro M. Zocchi and Michele F. Fontefrancesco. Traditional Products and New Developments in the Restaurant Sector in East Africa. The Case Study of Nakuru County, Kenya. *Frontiers in Sustainable Food Systems*, 4, 2020.

[13] John Kennedy and Ian Cosnett. Food flavours biology and chemistry. *Carbohydrate Polymers - CARBOHYD POLYM*, 46:296–296, November 2001.

[14] Kush R. Varshney, Lav R. Varshney, Jun Wang, and Daniel Myers. Flavor Pairing in Medieval European Cuisine: A Study in Cooking with Dirty Data. Technical Report arXiv:1307.7982, arXiv, July 2013. arXiv:1307.7982 [physics] type: article.

[15] Mansi Goel and Ganesh Bagler. Computational gastronomy: A data science approach to food. *Journal of Biosciences*, 47, March 2022.

[16] Yasmina SELLAM. *Mémoire culinaire de l'Algérie, histoire de recettes*. 2022.

[17] +UNESCO. Les traditions du couscous à l'UNESCO – un exemple de coopération culturelle internationale, last visit: 22-06-2022. https://fr.unesco.org/news/

traditions-du-couscous-lunesco-exemple-cooperation-culturelle-interna December 2020.

[18] How many languages are there in the world?, last visit: 25-05-2022. https://www.ethnologue.com/guides/how-many-languages, May 2016.

[19] NATURAL LANGUAGE PROCESSING | Meaning & Definition for UK English | Lexico.com, last visit: 25-05-2022. https://www.lexico.com/definition/natural_language_processing.

[20] The NLP pipeline, last visit: 25-05-2022. https://pythonwife.com/the-nlp-pipeline/.

[21] Heston Blumenthal. Weird but wonderful. *The Guardian*, May 2002.

[22] Anupam Jain, Rakhi N. K, and Ganesh Bagler. Spices form the basis of food pairing in Indian cuisine. Technical Report arXiv:1502.03815, arXiv, February 2015. arXiv:1502.03815 [physics, q-bio] type: article.

[23] Rudraksh Tuwani, Nutan Sahoo, Navjot Singh, and Ganesh Bagler. Computational Models for the Evolution of World Cuisines. In *2019 IEEE 35th International Conference on Data Engineering Workshops (ICDEW)*, pages 85–90, April 2019. ISSN: 2473-3490.

[24] Osame Kinouchi, Rosa Diez-Garcia, Adriano Holanda, Pedro Zambianchi, and Antonio Carlos Roque. The non-equilibrium nature of culinary evolution. *New Journal of Physics*, 10:073020, July 2008.

[25] Yong-Yeol Ahn, Sebastian E. Ahnert, James P. Bagrow, and Albert-László Barabási. Flavor network and the principles of food pairing. Technical Report arXiv:1111.6074, arXiv, November 2011. arXiv:1111.6074 [physics] type: article.

[26] B. Galvan, Greiner D., Périaux J., M. Sefrioui, and G. Winter. Parallel Evolutionary Computation for Solving Complex CFD Optimization Problems : A Review and Some Nozzle Applications. In K. Matsuno, A. Ecer, N. Satofuka, J. Periaux, and P. Fox, editors,

Parallel Computational Fluid Dynamics 2002, pages 573–604. North-Holland, Amsterdam, January 2003.

[27] Nils Aall Barricelli. Numerical testing of evolution theories. *Acta Biotheor*, 16(1):69–98, March 1962.

[28] W. Vent. Rechenberg, Ingo, Evolutionsstrategie — Optimierung technischer Systeme nach Prinzipien der biologischen Evolution. 170 S. mit 36 Abb. Frommann-Holzboog-Verlag. Stuttgart 1973. Broschiert. *Feddes Repertorium*, 86(5):337–337, 1975. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/fedr.19750860506.

[29] Andrew N. Sloss and Steven Gustafson. 2019 Evolutionary Algorithms Review. Technical Report arXiv:1906.08870, arXiv, June 2019. arXiv:1906.08870 [cs] type: article.

[30] Brandon Morgan. Unit 5) Evolutionary Programming, last visit: 25-05-2022. https://towardsdatascience.com/unit-5-evolutionary-programming-cced3a00166a, July 2021.

[31] Youcef Merabti. Optimisation des réseaux de neurones MLP par l'algorithme hybride AG-RT pour le contrôle d'un système non linéaire. 2015. Accepted: 2019-01-17T06:24:03Z Publisher: Université Larbi Ben M'hidi.

[32] ZERARI Naima. LES ALGORITHMES GENETIQUES EN MAINTENANCE. Master's thesis, Univesity of Batna, 2014.

[33] Eyal Wirsansky. *Hands-On Genetic Algorithms with Python: Applying genetic algorithms to solve real-world deep learning and artificial intelligence problems*. Packt Publishing, 1er édition edition, January 2020.

[34] Rakesh Kumar and Jyotishree. Blending Roulette Wheel Selection & Rank Selection in Genetic Algorithms. *IJMLC*, pages 365–370, 2012.

[35] Mutation (genetic algorithm), last visit: 25-05-2022. https://en.wikipedia.org/w/index.php?title=Mutation_(genetic_algorithm)&oldid=1091156249, June 2022. Page Version ID: 1091156249.

[36] Steffen Schulze-Kremer. Genetic Algorithms and Protein Folding. In David M. Webster, editor, *Protein Structure Prediction: Methods and Protocols*, Methods in Molecular Biology™, pages 175–222. Humana Press, Totowa, NJ, 2000.

[37] Filippo Carnovalini and Antonio Rodà. Computational Creativity and Music Generation Systems: An Introduction to the State of the Art. *Frontiers in Artificial Intelligence*, 3, 2020.

[38] Genetic Algorithms and its use-cases in Machine Learning, last visit: 25-05-2022. https://www.analyticsvidhya.com/blog/2021/06/genetic-algorithms-and-its-use-cases-in-machine-learning/.

[39] Amita Malav, Kalyani Kadam, and Pooja Kamat. PREDICTION OF HEART DISEASE USING K-MEANS and ARTIFICIAL NEURAL NETWORK as HYBRID APPROACH to IMPROVE ACCURACY. *International Journal of Engineering and Technology*, 9:3081–3085, August 2017.

[40] Ms Shikha Malik and Mr Sumit Wadhwa. Preventing Premature Convergence in Genetic Algorithm Using DGCA and Elitist Technique. https://www.semanticscholar.org/paper/Preventing-Premature-Convergence-in-Genetic-Using-Malik-Wadhwa/db7678845db91e5cdabbd2767de11c1e85eae6ae, 2014.

[41] Bouayed Fatima-Zohra. *La cuisine algérienne*. Temps Actuels, January 1970.

[42] Neelansh Garg, Apuroop Sethupathy, Rudraksh Tuwani, Rakhi NK, Shubham Dokania, Arvind Iyer, Ayushi Gupta, Shubhra Agrawal, Navjot Singh, Shubham Shukla, Kriti Kathuria, Rahul Badhwar, Rakesh Kanji, Anupam Jain, Avneet Kaur, Rashmi Nagpal, and Ganesh Bagler. FlavorDB: a database of flavor molecules. *Nucleic Acids Research*, 46(D1):D1210–D1216, January 2018.

[43] Welcome to Python last visit: 25-05-2022. https://www.python.org/.

[44] Stack Overflow Developer Survey 2021, last visit: 25-05-2022. https://insights.stackoverflow.com/survey/2021/?utm_source=social-share&utm_medium=social&utm_campaign=dev-survey-2021.

[45] re — Regular expression operations — Python 3.10.5 documentation, last visit: 25-05-2022. https://docs.python.org/3/library/re.html.

[46] pandas - Python Data Analysis Library, last visit: 25-05-2022. https://pandas.pydata.org/about/index.html.

[47] Natural language toolkit, last visit: 25-05-2022. https://www.nltk.org/.

[48] Pattern, last visit: 25-05-2022. https://github.com/clips/pattern, June 2022. original-date: 2011-05-03T15:29:01Z.

[49] Enchant, last visit: 25-05-2022. https://abiword.github.io/enchant/.

[50] Plotly Python Graphing Library, last visit: 25-05-2022. https://plotly.com/python/.

[51] H. Arn and T. E. Acree. Flavornet: A database of aroma compounds based on odor potency in natural products. In E. T. Contis, C. T. Ho, C. J. Mussinan, T. H. Parliment, F. Shahidi, and A. M. Spanier, editors, *Developments in Food Science*, volume 40 of *Food Flavors: Formation, Analysis and Packaging Influences*, page 27. Elsevier, January 1998.

[52] Marco Silva, A.M. Freitas, Maria Cabrita, and Raquel García. Olive Oil Composition: Volatile Compounds. pages 21–22. February 2012.

[53] John Brodie and John Godber. Bakery Processes, Chemical Leavening Agents. In *Kirk-Othmer Encyclopedia of Chemical Technology*. John Wiley & Sons, Ltd, 2007. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/0471238961.0308051303082114.a01.pub2.