

CSCI6612 – Visual Analytics

Project proposal

Cooking Partner

by

|  |  |
| --- | --- |
| Team Pandas | |
| Lan Chen | [lan.chen@dal.ca](mailto:lan.chen@dal.ca) |
| Song Zhao | [song.zhao@dal.ca](mailto:song.zhao@dal.ca) |
| Sunit Shah | [sunit.shah@dal.ca](mailto:sunit.shah@dal.ca) |
| Eugene Shishlannikov | [eugene.shishlannikov@dal.ca](mailto:eugene.shishlannikov@dal.ca) |

Table of Contents

[**Introduction, motivation and benefits** 3](#_Toc11605138)

[**Functionality** 3](#_Toc11605139)

[Minimum Functionality 3](#_Toc11605140)

[Expected Functionality 3](#_Toc11605141)

[Bonus Functionality 3](#_Toc11605142)

[**High Level Organization and Design** 4](#_Toc11605143)

[Minimum functionality 4](#_Toc11605144)

[Expected functionality 4](#_Toc11605145)

[Bonus functionality 5](#_Toc11605146)

[**Milestones and timeline** 6](#_Toc11605147)

[References 6](#_Toc11605148)

# **Introduction, motivation and benefits**

Consuming foods is vital to any living being. For us, humans, food consumption have become an inseparable part of a culture long before we begun to develop many of its other aspects. From the very beginning, human’s ration was mostly determined by the local agriculture, and this simple factor caused the high diversity in the recipes around the globe.

Many thousands of recipes exist in the world and it is interesting to visualize the information about them and the relations existing between different cuisines. We will create a web-based application that visualizes some of the information that can be extracted from the dataset [1] we have. Namely, the information that we plan to visualize is the frequency of the mentioning of ingredients in recipes using word cloud, the similarity of the recipes by plotting them on the two-dimensional graph so that similar ones will be located close to each other, and the degree of the similarity between different cuisines using the Sankey diagram.

Using this application user will be able to learn information that is unavailable without the usage of data science and visualization techniques.

# **Functionality**

For all the functionalities, we have decided to use various data mining techniques and Machine learning algorithm to process and visualize the data.

## Minimum Functionality

Our minimum functionality would be to build the word cloud after cleaning the data to find the most used ‘Ingredients’ in the whole datasets of recipes. The goal is to determine the most common and most used ingredients from all the recipes available in the dataset. To visualize this information, we chose to use the ‘Word cloud’ as this is one of the best ways to represent the frequency of ingredients with variation in the size of the name.

## Expected Functionality

After achieving the minimum functionality, our next target is to visualize the clustering of recipes. To achieve this goal, we need to reduce the high dimensionality of recipes into two dimensions in order to make a scatterplot. The goal is to show the similarities between various cuisines all over the world and to detect which countries’ cuisines are similar in terms of ingredients being used.

## Bonus Functionality

Our aim for the bonus functionality is to show the similarities between different cuisines in a more intuitive way. We plan to calculate the similarity weights between all the cuisines and apply the result in a Sankey diagram. Visualization will show the strong and weak connections between cuisines.

# **High Level Organization and Design**

On the left side of the application, there will be three options for users to explore with an explanation for the result below them. On the right side, it shows the visualization of the result. Based on three functionalities, we designed three different visualizations.

## Minimum functionality

This application uses Word Cloud [2] to show the popular ingredients in the world that appear in our dataset. The more popular, the font size is bigger. The wireframe of it shows in Figure1.

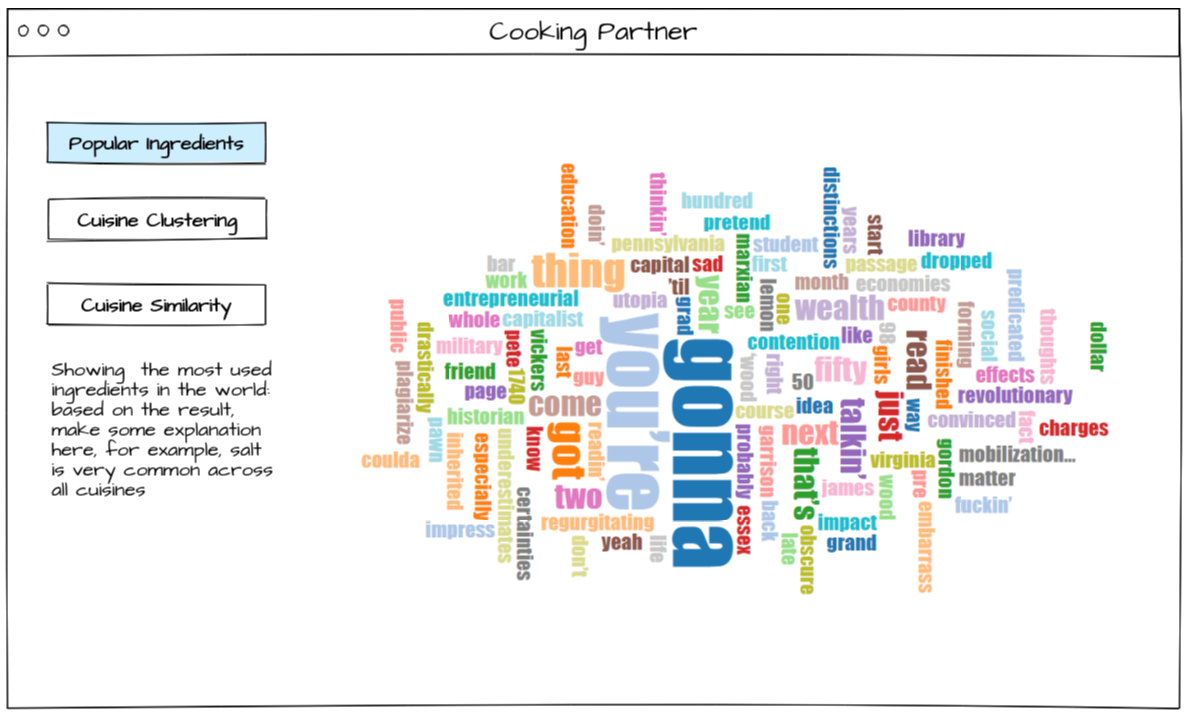


Figure 1 Minimum functionality visualization

## Expected functionality

Each recipe is represented by a vector with several dimensions, where each dimension means one ingredient. We plan to reduce the dimensionality to two in order to make a Scatterplot [3] of the dataset to explore the similarity of the cuisines. If two cuisines overlap, that means they are more similar. If two cuisines are far away from each other, that means they share little to no similarity. The wireframe of the expected functionality user interface is shown in Figure 2 below.

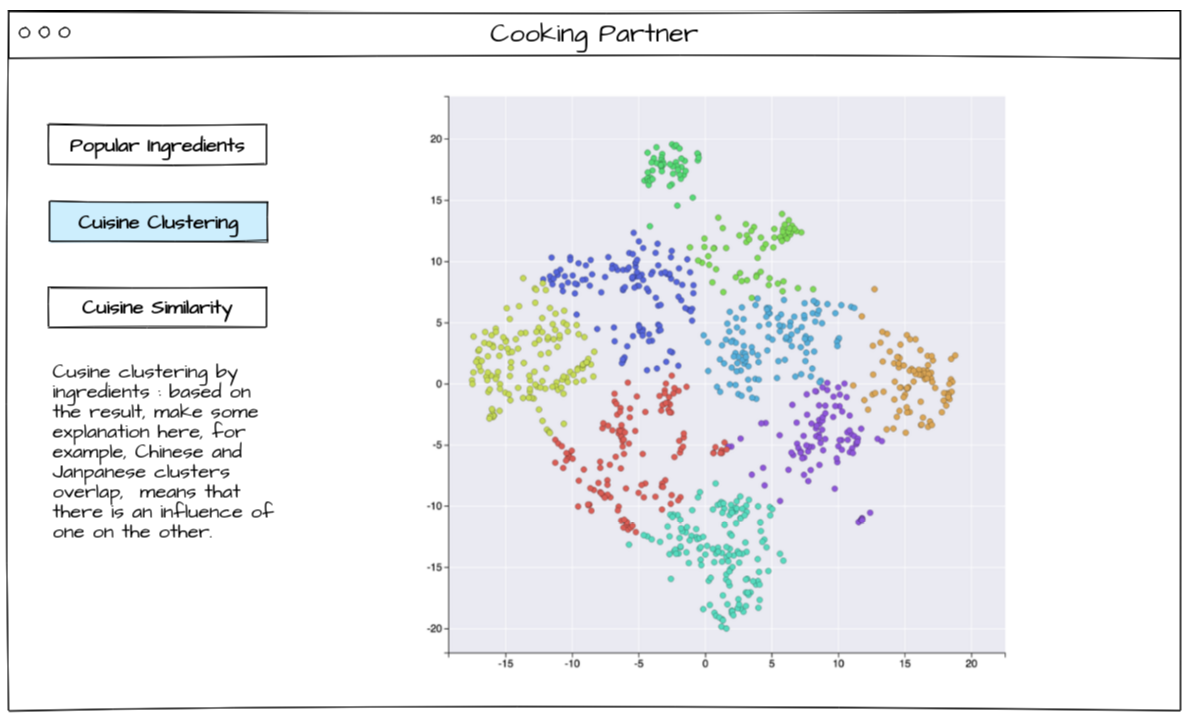


Figure 2 Expected functionality visualization

## Bonus functionality

For a more intuitive visualization of similarity between each cuisine, we plan to use a Sankey diagram [4]. On the left side of the visualization is the source of a cuisine. On the right side, it’s the target of a cuisine. The flows show how similar for the source cuisine to the target cuisine. The thicker the width of the flow, the more similarity between the two cuisines. The wireframe of the user interface it shows in Figure 3 below.

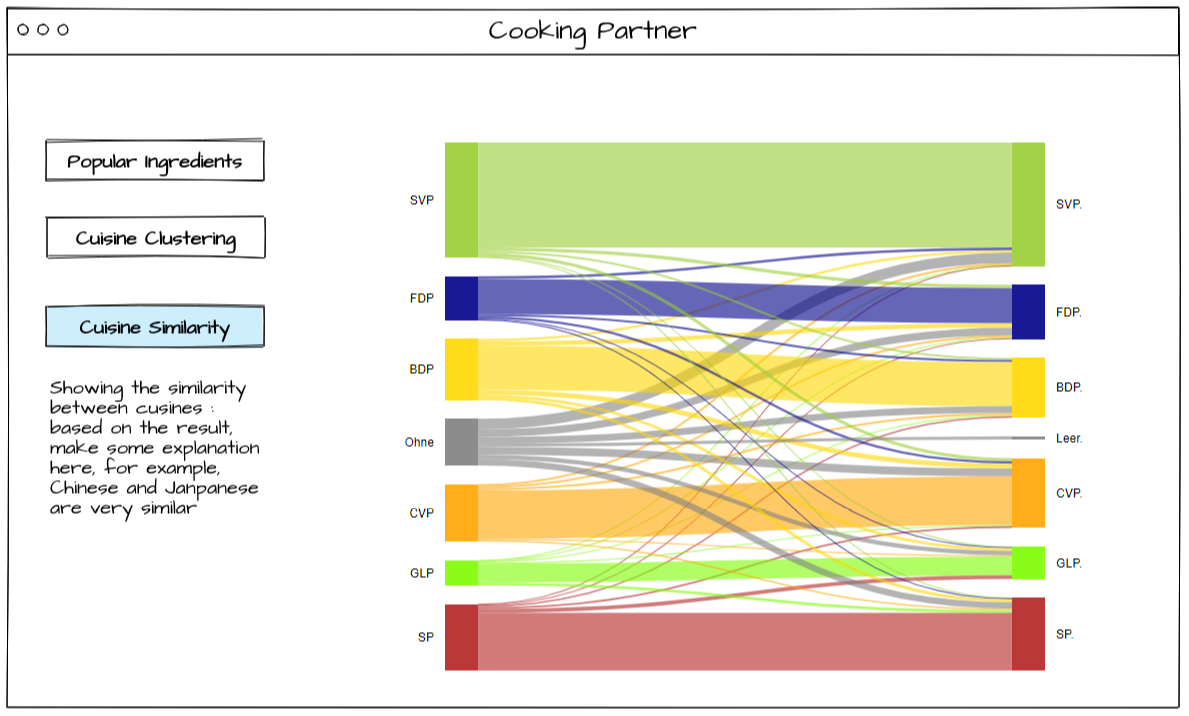


Figure 3 Bonus functionality visualization

# **Milestones and timeline**

|  |  |  |  |
| --- | --- | --- | --- |
| **No.** | **Due Date** | **Milestone** | **Responsibility** |
| 1 | 2019-06-23 | Clean missing data and bad data, convert and normalize it. | Eugene, Sunit |
| 2 | Build collaborative environment for programming using Github, build the frame of front-end | Lan, Song |
| 3 | 2019-06-30 | Learn visualization and machine learning techniques | Eugene, Sunit, Lan, Song |
| 3 | 2019-07-07 | Develop minimum functionality | Lan, Song |
| 4 | Develop expected functionality | Eugene, Sunit |
| 5 | 2019-07-21 | Develop the back-end of the bonus functionality | Eugene, Sunit |
| 6 | Develop the front-end of the bonus functionality | Lan, Song |
| 7 | 2019-07-26 | Polish and wrap the project | Lan, Song |
| 8 | 2019-07-26 | Finish project report and project presentation material | Eugene, Sunit |

# References

*D3 Wordcloud Example*. (n.d.). Retrieved from https://bl.ocks.org/jyucsiro/767539a876836e920e38bc80d2031ba7

*How Coursera uses Data Visualization and Clustering to Categorize Content*. (n.d.). Retrieved from https://www.analyticsvidhya.com/blog/2019/01/coursera-data-driven-content-categorization-algorithm/

*Party Flow Sankey*. (n.d.). Retrieved from https://bl.ocks.org/cdermont/846051eb548e846eac60

*Used dataset source*. (n.d.). Retrieved from https://github.com/conwayyao/Recipe-Analysis/blob/master/README.md