Machine Learning Report

I. Similarity and Polarity Feature Vector

The dictionary was built using the three blogs cited previously, and the 5% of customers reviews group by rating score. We noted, for example, that we lose valuable information using only the keywords generated by blogs, because people are putting attention in other details of coffee shops that we can't extract from blogs (decoration, cleaning, and dirt in floors, kitchen, bathroom, how dangerous is the neighborhood, etc) and the type of details depends on the score assigned. Thus, we choose randomly the 5% of reviews with score 1, 5% with score 2 and so on, for trying to rescue a text representative sample of what people are talking about in every categorical score.

The pipeline of preprocessing consists of normalization and expansion of the contractions, followed by filtering of special characters, tokenization to split the document into words for deleting the stop words and applying lemmatization. Stop words includes the name of coffee shops (except the words "coffee", "roasters" or others that must be part of the dictionary).

Now, we use these normalized paragraphs to extract keywords by bi-gram TD-IDF. If we want to extract information about the ambient of the coffee, available of sit, what to do there, how crowded is the coffee shop, name of drinks composed with 2 words, the feelings around the coffee drink, food or place, necessary we need bi or even three grams. Bi-gram was selected to find key-phrases because two words are enough to know the meaning of the expressions.

The dictionary contains 5378 items, but after to label them in the corresponding categories and sections, we retrieve 1267 key phrases. After to inspect the dictionary, two new sections were included: ***price*** (on category place), ***snacks*** (on food category) and ***do*** (on category place, including study, work, talking with friends, hanging and elements associated as laptop, wifi, books). The definitive categories used are displayed in Figure 1 and replace the name ***decoration*** for ***ambient*** because this topic includes decoration, size of the business, music; ***go*** includes elements associated with the possibility or not of find a set, grab the coffee to sit or to go, how crowded is the place, availability of tables and ***out*** is related to the view, neighborhood, parking.

Joining these key-words, we build topic documents of every section and we split the customer reviews (paragraphs) into sentences to looking for the similarity between the topic documents with every sentence using cosine similarity. In this way, we have reviews split into sentences and every sentence is represented as a vector with 16 features with the similarity score between the sentence and every topic document, called Similarity feature vector. Additionally, the polarity pattern of every sentence is computed and rescaled to have values between 0 and 1 instead -1 and 1. We did that because, as we noted in the previous analysis, neutral sentences are frequent and they have a polarity pattern around 0.

Multiplying the original polarity scores by the similarity scores cancels a lot of values in every feature and we potentially could lose a lot of information.

Finally, the vectors of sentences are group by review and aggregated using the mean of all the components, to build one vector for review with 16 features, corresponding to a ponderation between the similarity and the polarity of the topics presents on the review.

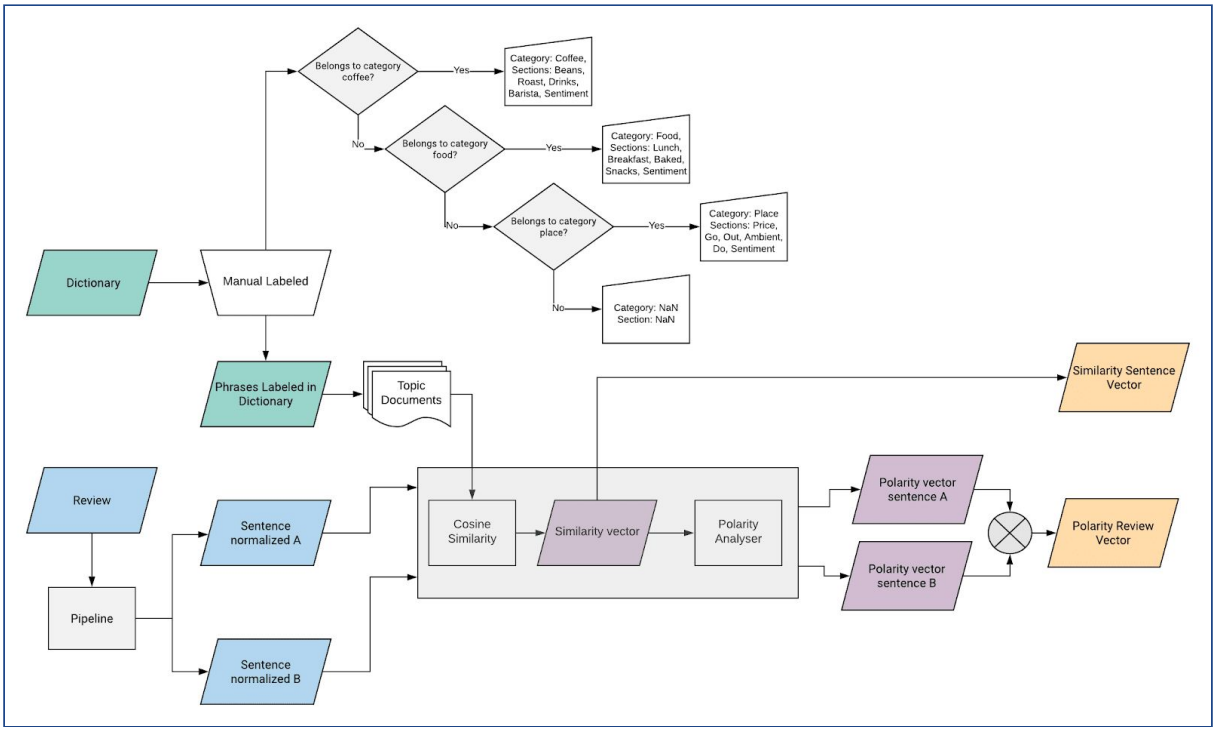

Figure 1: Pipeline building features vectors

II. Supervised Learning

1. Features selection

A statistic chi-squared test provided by SelectKBest was applied to extract the 12 features more strongly related to the label or output variable of the features vector. Tests measure dependence between the features and output variable and lower scores represent independence and then useless for the classification. Following Table 1. is showing the scores for all the features.

ID	SPECS	SCORE	ID	SPECS	SCORE
15	**food sentiment**	5.413	9	**price**	1.068
10	**place sentiment**	3.445	3	**barista**	1.065
4	**coffee sentiment**	2.620	2	**drinks**	0.850

13	**breakfast**	2.499	5	**go**	0.815
8	**ambient**	2.215	14	snacks	0.663
11	**baked**	1.955	12	lunch	0.562
6	**do**	1.552	0	beans	0.165
7	**out**	1.171	1	roast	0.086

Table 1: Selection of features

2. Output variable

If we check the distribution of the polarity pattern of reviews according to the rating score, we can see high overlapping of score 2 in 1 and score 4 in 5, and a trend of score 3 to overlap positive and negative cloud of scores. If you think about that just for one minute, you realize that the concept of a positive and negative connotation is more intuitive than the difference between 1 or 2 scores and together they enclose a more fundamental category. The same case happens with 4 and 5 stars on reviews. After to inspect reviews with 3 stars, it was decided to ignore them in this section, since the criteria to put 3 stars to a slightly positive or slightly negative review is apparently random and depends absolutely on the customers perspective and it is independent and different each other.

Next analysis solves a binary classification problem of positive and negatives reviews.

3. XGBoost

XGBoost is one of the most popular techniques used in Kaggle competition. This variation of boosting is an implementation of Gradient Boosting, an ensemble method that put attention in residuals of previous models and tries to minimize the loss in the following iteration, add a penalty in the objective function.

We start tunning the model using 1000 estimators, using a learning rate of 0.01 (normally is between 0.1 and 0.01), a subsample of 0.8 (best practice is to choose a value between 0.8 and 1), max depth of the trees equals to 3 (default, then we increase the value in 1 and check if the performance offers a significative improvement) and alpha regularization of 1. Additionally, we use the **binary logistic** as the objective function to classify the features vectors.

Splitting data in 80% for training and fitting a model using the values mentioned above, we observe the AUC and error measured aggregating an additional estimator, from 0 to 1000. The results indicate that after to 550 estimators, AUC/error increases/reduces slowly, therefore, we select 550 estimators.

After that, we compare the performance of different ensembled tree models. For instance, the difference between Random Forest and Random Forest Logistic Regression is that the second is the result of a pipeline (fitting the ensemble model using the training data, encoding the result using OneHotEncoder and finally, fitting the logistic model). The training data is split in 50% for fitting the first model and the rest to the second model. In this way, we build a Random Forest, Logistic Random Forest, Gradient Boosting, Logistic Gradient Boosting. all of them using 5 estimators. And we compare the performance of these models with the XGBoost models. The ROC curve of them is presented in Figure 2. As we know, XGBoost is one of the fastest implementations of Gradient Boosting that regularize the trees and it avoids overfitting with randomization (we decide a subsample of 80% training data to train each base model of this gradient boosting).

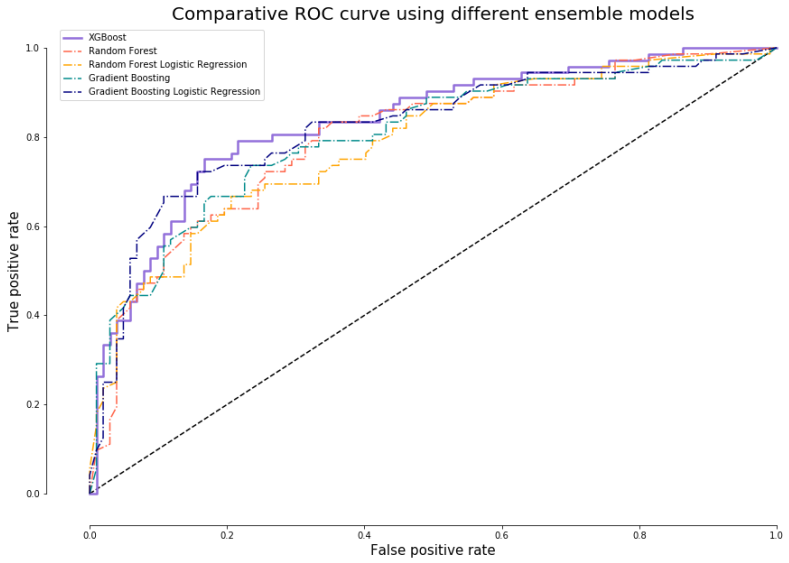

Figure 2: Comparative ROC of XGBoost, Random Forest and Gradient Boosting combined with Logistic Regression.

As we can observe, the performance of Random Forest is better without the Logistic Regression step, but the performance of the general Gradient Boosting increases with the logistic method (GBLR). XGBoost has a ROC curve comparable with the GBLR and checking the AUC and Accuracy Table, we can bear out this. AUC is slightly more for XGBoost than GBLR, but the Accuracy is significantly better.

Model	AUC	Accuracy
XGBoost	**83.46%**	**77.01%**
Random Forest	79.71%	71.84%
Logistic Regression Random Forest	78.50%	71.26%
Gradient Boosting	80.22%	75.29%
Logistic Regression Gradient Boosting	**82.16%**	**73.56%**

Table 2: Performance of the different models

The Confusion Matrix is:

	TP Reviews	TN Reviews
PP Review	77	25
NP Review	15	57

Table 3: Confusion Matrix for XGBoost model fitting

III. Unsupervised Learning

Finally, can we distinguish clusters to separate the coffee shops? What kind of criteria we can use to split them?

The following analysis is centered on topics and how we can use those for clustering the business, using the similarity features vectors computes previously to know the topics on reviews, but in this case we group reviews by coffee shops. We don't want to classify the information depending on the connotation that the customer put on his review, instead how often people is talking about specific topics and how we can utilize that information to generate clusters of coffee. To do that, we group the features by coffee shops with the aggregate function sum.

Firstly we tried to find groups of coffee shops using all features. In this section, we apply K-Means followed to Agglomerative Clustering to find unsupervised groups of coffee shops. We implement a two-phase solution using k-Means with 12 clusters and then, as a second stage, a hierarchical clustering of 3 clusters. The results are displayed below:

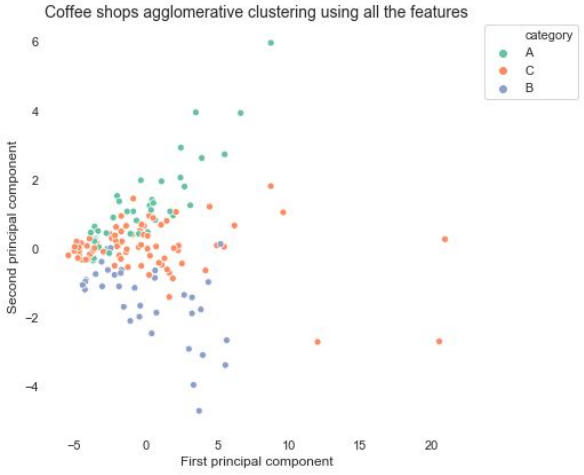

Figure 3: Agglomerative Clustering of coffee shops using the 16 features

Which features are determining this categorization? What if we inspect some boxplot of the values to different features? As we can see on the boxplots of Figure 4, Cluster B include coffee shops with more mentions associated with the coffee cup itself (beans, drinks, sentiments about the drinks). Otherwise, Cluster A has the lowest coffee mentions, but instead, people are talking about the food (sentiments, baked food).

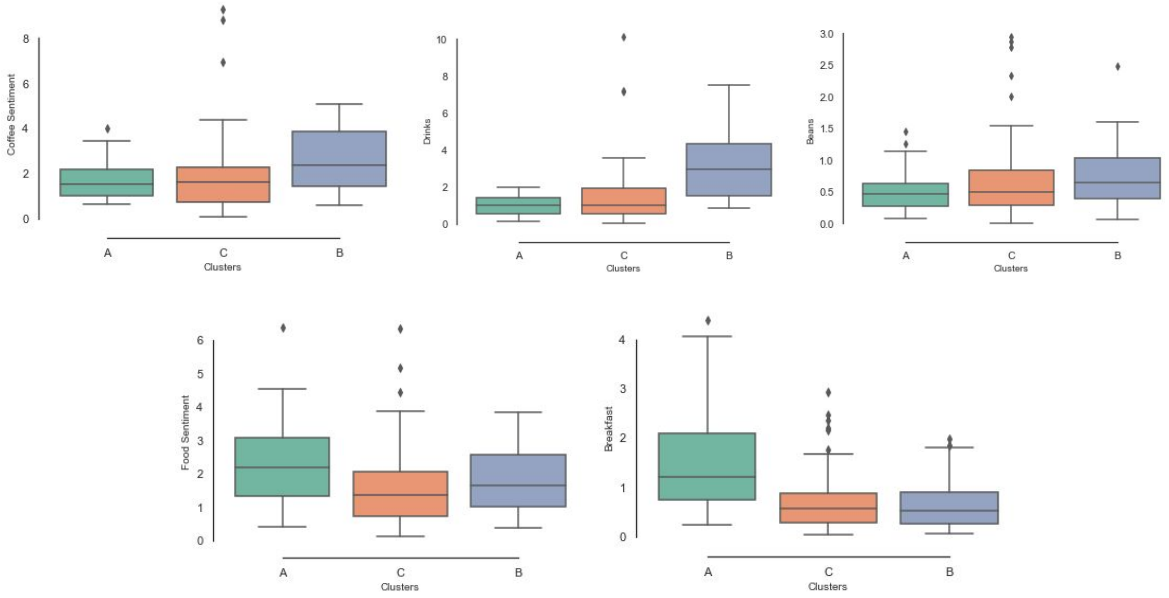

Figure 4: Representation of features values in boxplots split into the resultant clusters of Hierarchical Clustering

Secondly, we analyzed the features split into the original categories: coffee (beans, roast, baristas, drinks, sentiments), food (baked, breakfast, lunch, snacks, sentiments) and place (go, do, out, ambient, price, sentiments). To do that, hierarchical models were trained for building dendrograms and determine the number of clusters that allows splitting the coffee shops into reasonable groups and what criteria we can use to understand and interpret the results.

1. Coffee category

A dendrogram of the coffee shops considering only sections related to coffee is displayed below. This structure is the result of an **Agglomerative Clustering** that uses the method `complete`, that links clusters using the less similar points (or far away observations) and `cosine` distance as metric. The resultant groups are compact and highly similar.

Cosine reduces the noise into account the shape of variables, more than their values and it is useful when you have many variables and you are not sure about the significance of them into the model. After to test variables using chi-square we know which variables have more influence for training models, but in this case, we separate the categories (thus, we separate the most decisive variables) to force that another variables take more protagonism and determine the label of the features vectors.

We extract 6 groups inspecting the clusters generated at a distance of 0.25. And we can note from dendrogram, there is a group that contains only one variable, the Laundré coffee shop. Therefore, in practice we have five clusters, showed in Figure 6 in a bidimensional representation (using PCA to generate coordinates in axis x and y).

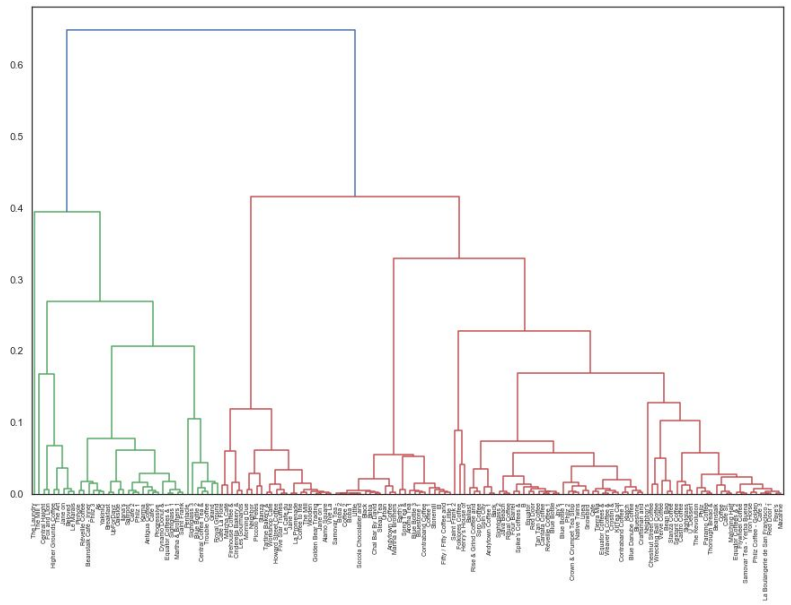

Figure 5. Dendrogram of coffee shops clustering by coffee features

The size of the elements depends on the sum of **roasting expressions** for every coffee business. We can observed, for example, that **Cluster F** has the higher roasting sum, followed by the **Cluster B** and **D**. The clusters with less allusion to the roasting of the seeds are E and A. Thus, information about **roasting** allows to distinguish categories of coffee shops and it is useful for the formation of clusters.

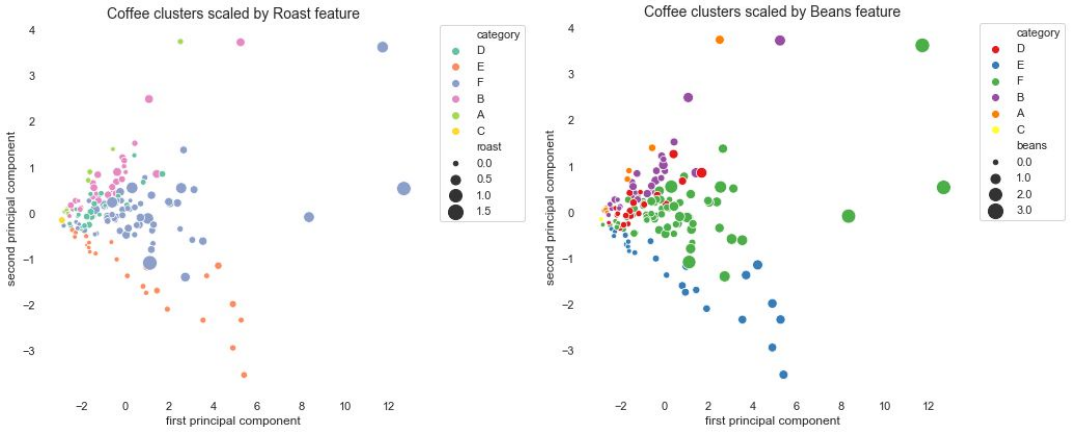

Figure 6. Clusters using the Coffee Category with the size of points adjusted to different features

In other words, the clusters generated by the coffee categories split the coffee shops according to the coffee features and roasting is a influyent variable, describing how much attention their customers that wrote reviews put on the roasting. The same exercise with other features reveals that the mentions about beans are different between the clusters. Indeed, beans and roasting are highly mentioned in the same cluster, that includes some of the most iconic San Francisco coffee shops as Andytown Coffee Shops, Mazarine, Four Barrel, Paramo, Red Door, Ritual Coffee, Saint Frank, Sightglass, Equator, Blue Bottle, Wrecking Ball. Meanwhile, the tiniest cluster, A includes bakeries and stores where coffee

itself is not the most essential part of the experience (Le Marais, Art's Cafe) and The Laundry represent a cluster as unique element and it is a kind of gallery for events, artists, innovators, and other creative types. The **Cluster E** contains other coffee shops, but mostly tea, chai and chocolatier business.

2. Food category

Using the same procedure, we build the dendrogram of the hierarchical clustering for food features (Figure 7) and we find 5 clusters using 0.35 as a threshold of the dendrogram. The clusters as a bidimensional representation are shown in Figure 5: to the left, we can note that the baked is a topic predominant in **Cluster C**, where we can find some coffee shops outstanding for their pastries and baked items.

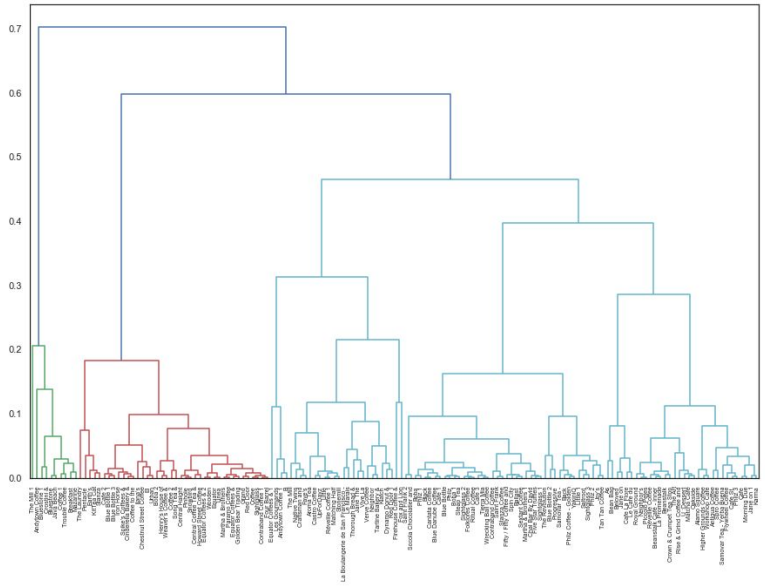

Figure 7. Dendrogram of coffee shops clustering by food features

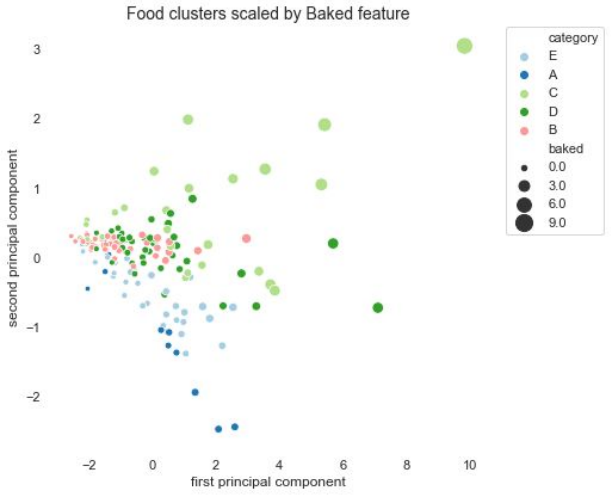

Figure 8. Clusters using the Food Category with the size of points adjusted to Baked feature

3. Place category

The dendrogram of the hierarchical clustering for place features is displayed on Figure 9 and we extract 7 groups using a threshold of 0.15 on the dendrogram.

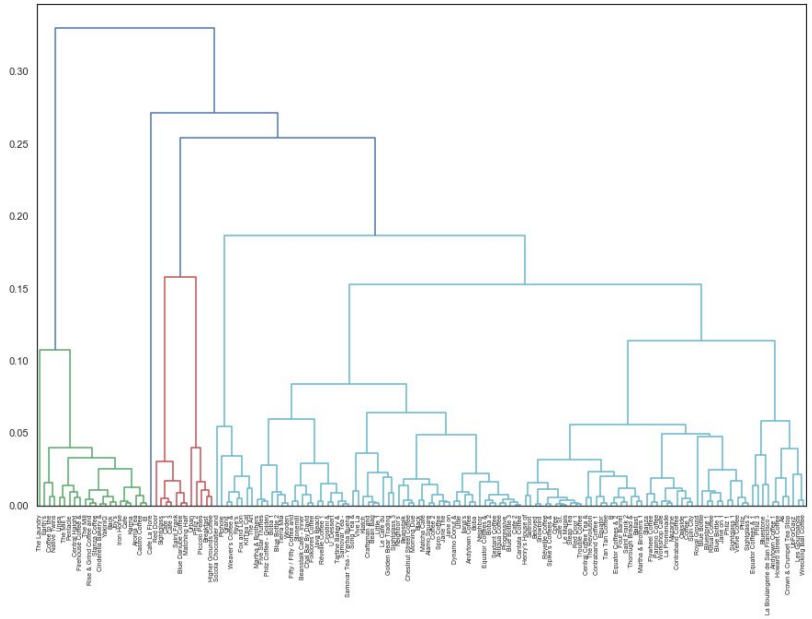

Figure 9. Dendrogram of coffee shops clustering by food features

A bidimensional representation is exposed in Figure 10. Feature **Do** is about what to do in the coffee shop. **Cluster B** contains huge points (it means that reviews in this cluster have higher mentions related to this feature. The coffee shops inside this cluster are Blue Danube Coffee, Matching Half, Red Door, Saint Frank, and Sightglass, that include space inside the store to study or working on your laptop. Piccolo Petes, Urban, and Higher Grounds Coffee are part of **Cluster C** and they represent three coffee shops where people commonly going to share with friends (Higher Ground has a bar style).

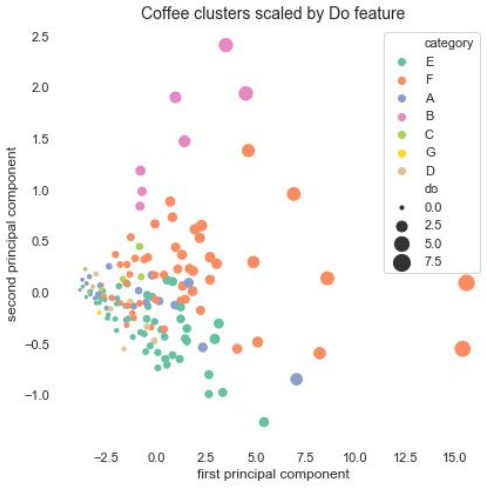

Figure 10. Clusters using the Place Category with the size of points adjusted to Do feature

Cluster D contains some business as KitTea Cat, Socola Chocolatier, Pinhole, and Faye's Coffee and it's interesting to note that all of them put special attention to the experience: KitTea has cats around the shop, Socola is all about chocolate, Pinhole surprises with colorful walls and drawings in coffee cups and Faye's is surrounded by books; **Cluster F** includes coffee shops concentrated in crowded places with a lot of movement during weekdays (financial district, along to Market and Mission streets) and high mentions about what to do there.