Coffee Shops in San Francisco

Looking for meaningful, unconventional and unique experiences

Final Report Capstone 1

I. Problem statement:

Every day, customers of coffee shops are asking for more quality, variety, and care in details. Comfortable, calm, noisy, dynamic, minimalist, colorful or over-decorated spaces contribute to building diverse experiences. Some customers want to be surprised, and others wish to find exactly they expect. It's possible to identify different types of coffee drinkers and with all of the information out there, we can find the best recommendation for each one.

How the matching between people and their coffee shops choices could be improved? The following analysis is focused on the people experiences: the purpose is to discover what criteria people use to decide if an experience is pleasurable or disappointing.

In this way, we can improve the approach of recommendations when you are looking for a specific experience, in this particular case, coffee shops. Additionally, we can find a business opportunity in experiences with less coverage for interested clients and this methodology could be extrapolated to other spectrums of cuisine experiences.

II. Description of the dataset:

a) Yelp Reviews Extraction

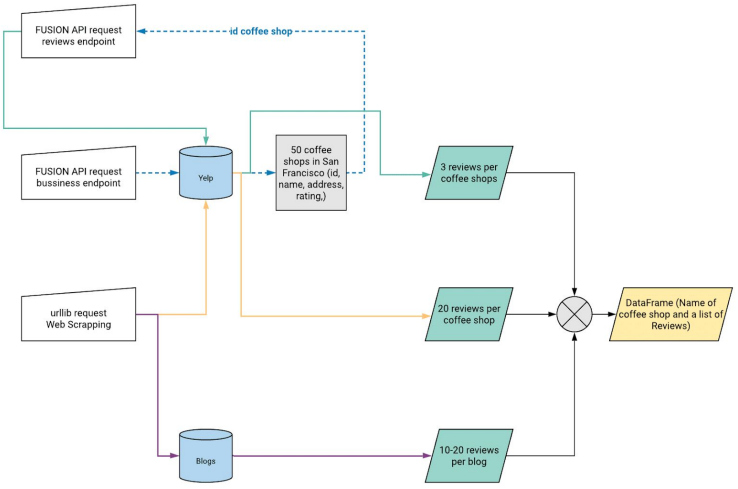

Fig. 1: Data flow from extraction to building of DataFrame

The first step is to send a request to Fusion (Yelp API) using the endpoint /businesses/search to get names/id of coffee businesses in San Francisco. For this

purpose, request parameters are San Francisco (as location), coffee (as the search term) and the maximum limit of data. The output is information about 50 businesses including id, name, address, opening hours, rating, number of reviews. Id's are the inputs to the second request to the endpoint /businesses/{id}/reviews to obtain 3 reviews per each coffee shop. This narrow number of review, encourage us to extract more reviews from Yelp using scraping.

The second step is web scraping with Beautiful soup library, searching Coffee Shops in San Francisco through the Yelp platform. Each page of results has 30 coffee shops. The dataset scraps 180 businesses, including some names gotten with the API request to add more reviews and other news results). The focus in this part is about inspecting the Yelp page and understand the basic syntax of HTML to find the HTML document, titles headings, paragraphs and hyperlinks to open the section that each coffee shops has in Yelp with it owns reviews. At this way, we get 20 reviews of every shop. In this section, all HTML tags are removed to extract names and reviews.

b) Blogs Reviews Extraction

Process of data consists of web scraping in blogs about coffee shops in San Francisco written since 2017. Additionally, for selecting blogs, locations were checked (some blogs include coffee shops from Oakland or Berkeley). In general, these kind of reviews are brief, but you can find a lot of useful words in a couple of phases. In selected blogs, one is related to better coffee shops (authors write about drinks and space features) and two are about better places for working or studying and unique decoration. Blogs chosen are listed below (Table 1).

id	Title of coffee blog	Authors	URL
1	*"28 of San Francisco's Essential Coffee Shops: Hot spots for your morning cup"*	Ellen Fort and Caleb Pershan	https://sf.eater.com/maps/best-coffee-shops-san-francisco-oakland-berkeley
2	*"17 of San Francisco's Most Unique Coffee Shops*	Katie Bush	https://www.venuereport.com/roundups/17-of-san-franciscos-most-unique-coffee-shops
3	*"The Best SF Coffee Shops For Getting Work Done"*	Taylor Abrams, Frida Garza, and Will Kamensky	https://www.theinfatuation.com/san-francisco/guides/the-best-sf-coffee-shops-for-getting-work-done

Table 1: Coffee Blogs chosen to the text-processing

c) Pre-processing

Following diagram shows the Natural Language Processing pipeline used as pre-processing to extract keywords from text data. Remotion of HTML tags it was applied

previously with BeautifulSoup library. In this section, we describe briefly the rest of wrangling steps.

A dictionary to expand English contractions is utilized to help with text standardization process. After that, all special characters/whitespace/ are removed and each document (paragraph) is split into sentences. Then, tokenization of sentences is applied to remove stopwords (like prepositions, articles and all words that appear frequently in the text but they don't have significance). Finally, the sentences are joined as a paragraph again.

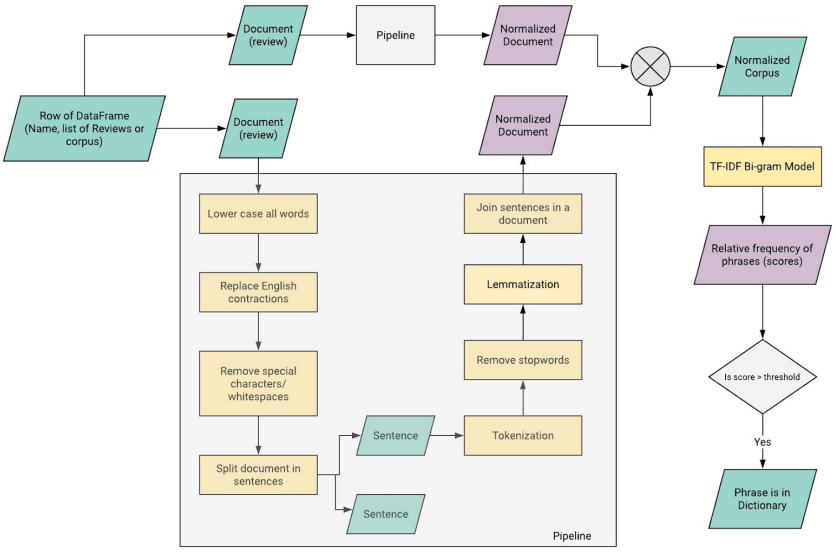

Fig. 2: NLP pipeline of text data

Doing the above pipeline with all reviews per coffee shops, it is possible building a normalized corpus that we use as the input to TF-IDF Model, that assigns a normalized frequency (score) to each word. This score is directly proportional to the frequency of the term in a document but inversely proportional to the frequency of the term in the whole corpus. It prevents assigning more importance to repetitive words that overshadow others.

III. Initial findings from the exploratory analysis

1. What blogs are talking about coffee shops?

What kind of information we can extract from coffee blogs? The following analysis was applied to blog id 1 included in the project. We extracted TF-IDF feature vectors for the text to find meaningful pairs of terms for classifying them in categories and sections relatives to the complete experience in the coffee shop. Key-pairs with a frequency normalized less than 20% were filtered and the dictionary used to label phases includes a round of one thousand of key-pairs. We define a **category** and a **section** for each key-pair. Criteria for the manual labeling was the following: a) only are considered the phrases with concrete and valid meaning and b) if a phrase classifies as meaningful, it belongs to:

- **Coffee**: all phrases relatives to types of drinks, beans, roasters, baristas, special types of sugar, milk, and items involving the experience about the cup of coffee. Sections are Baristas, Roasting, Beans, Drinks, Sentiment, None.

- **Food**: phrases about pastries, donuts, bagels, baked items in general, sandwiches, phrases relatives to breakfast and lunch with Sentiment, Breakfast, Baked (tiny items to eat in the coffee shop), Lunch and Brunch, None as sections.

- **Place**: place features as decoration, description of inside and outside spaces, parklets, sunsets from the seat, music, gardens, streets around; about the service itself (coffee to here, to go, wifi). Sections included are Decoration, To here, To go, Outside, Sentiment, Size of the coffee shop, wifi, None.

- **None**: all the rest.

Now, we want to discover which ones play a fundamental role in reviews.

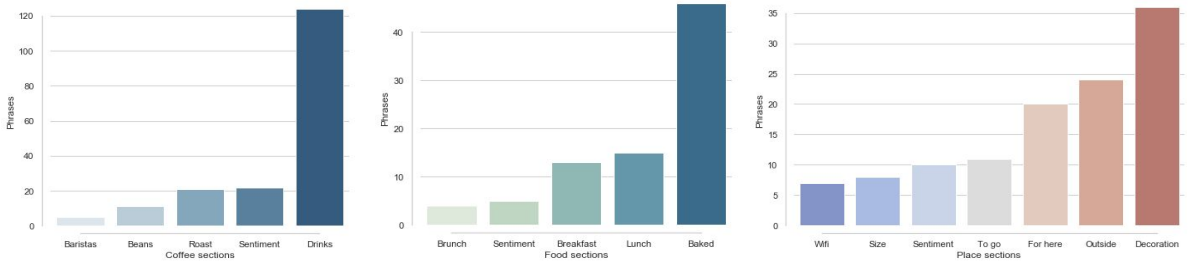

Fig. 3: Barplot of categories and sections built from the key-pairs dictionary

Sections for coffee are about the experience that involves the cup of coffee: types of drinks (including all varieties of coffee drinks, wine, and tea), roasting information (own roasters, origin of roast, types of roast offered to customers), beans, baristas (expertise of baristas) and sentiments (how do you feel about your drink). Then, 67% of coffee phrases are relatives to types of drinks, follows for feelings (12%) and roast features (11.4%). Baristas and Beans have lower predominance.

In the food phrases universe, baked items have the highest predominance (55%) and it makes sense because the current trends of coffee shops put more prominence in drinks and then in baked items, usually snacks and small elements. Lunch and Breakfast are the next priority with a predominance of 18% and 15%, respectively.

We considered as a place everything related to the physical space inside and outside of the coffee shop. Decoration in walls, musical elements and books are some examples of inside features. We discover some allusions to the features on the streets, sunsets and the scenario beyond the coffee shop itself but a component of the location of coffee. Additionally, there is information about coffee dynamic (for here, to go), availability of wifi and size of the shop. Decoration has majority predominance with 31%, follows by features out of the coffee shop (20%). There is special attention for parklets and what part of the city you could see from your seat in the coffee shop, for example. Sections For here and To go as one big section would be in the second predominance (26%). This last information is

extremely important to choose a shop because define an essential part of the type of experience that the customer will have.

2. Sentiment Analysis

Customers reviews are written from a perspective subjective, but it does not mean that all information there is completely subjective. How you could determine the subjectivity of reviews? Could you measure how much positive, negative or neutral is the information of customer reviews? In this section, we will catch two properties of review data from blogs and Yelp customers: polarity and subjectivity.

We are using the **sentiment** function from **TextBlob** library to study *polarity* and *subjectivity* of all dataset with the default analyzer, **PatternAnalyzer**, based on the **pattern** library). According to this function, polarity is between **-1 (negative result)** and **1 (positive result)** and subjectivity is between **0 (no subjective)** and **1 (absolutely subjective).**

Global analysis of reviews takes each one as a whole entity or **paragraph**. Then, they are splitting into **sentences** for discovering the intentionality of the parts of the message. Polarity on reviews according to the scores is displayed below.

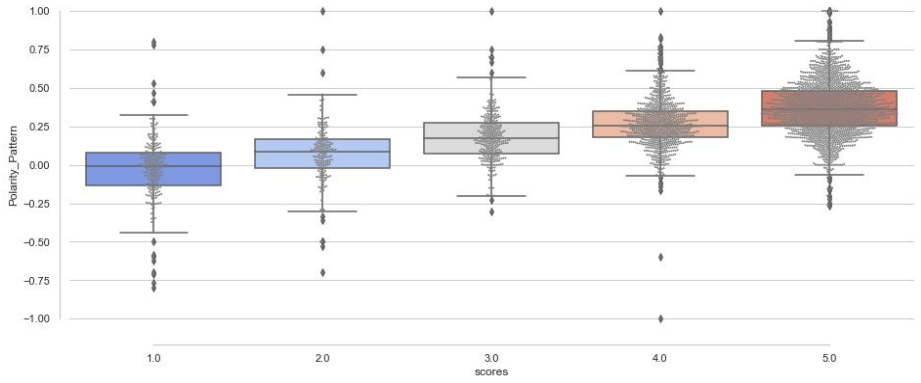

Fig. 4: Polarity pattern through the scores

The polarity mean is ascending through the rise of the score, as it can be expected. How much change the positivity and negativity of data after to split paragraphs into sentences? Must we expect higher or lower mean in polarity and subjectivity distributions? The following figure displays polarity and subjectivity histograms of reviews as a whole paragraph and split into sentences. The result in **polarity histograms is that the mean shifted toward the left when we split reviews**. It makes sense because we disintegrate the paragraph and determine the intentionality of every sentence.

Subjectivity analysis is interesting because the shape of the distributions is completely different: In paragraphs presence of data on the extremes is almost null and almost all data is concentrated in regular levels of subjectivity. In sentences, there is considerable information in the extremes, showing that we can find a strong presence of argumentative sentences.

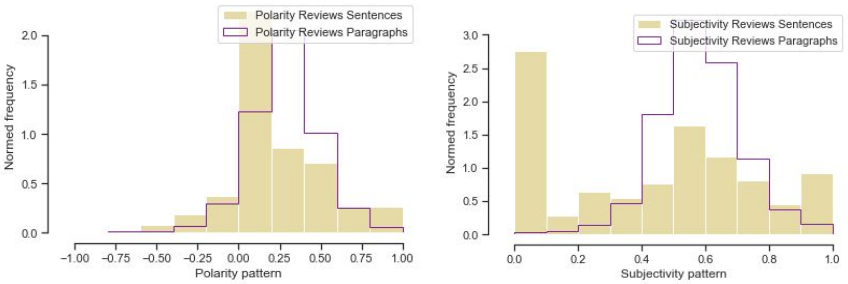

Fig. 5: Histograms of Polarity and Subjectivity to reviews on paragraphs and sentences

In this way, we can find **less positive and more argumentative messages that we could ignore in a global analysis** of reviews.

Initially, we had a dataset with almost three thousand and half of reviews. After the split, we have twenty thousand rows on the dataset.

a) Are people putting more emphasis and strong intention in positive or negative messages (polarity vs subjectivity)? How do we know that we take a representative random sample?

Pearson coefficient between polarity and subjectivity patterns for a random sample review is 0.467. We are wondering if, with a significance level of 0.05, this statistical metric is just a chance or **exists a strong correlation between both variables** (null hypothesis). A paired bootstrapping was applied to calculate the Pearson coefficient of every sample. After one thousand trials, p_value is 0.813, **concluding that we fail to reject the null hypothesis** and the relation founded is not a chance. A correlation of almost 0.5 suggests that most subjectivity data tend to be more positive too, that means, more emphasis on positive messages, as Figure 6 shows on the scatter plot below:

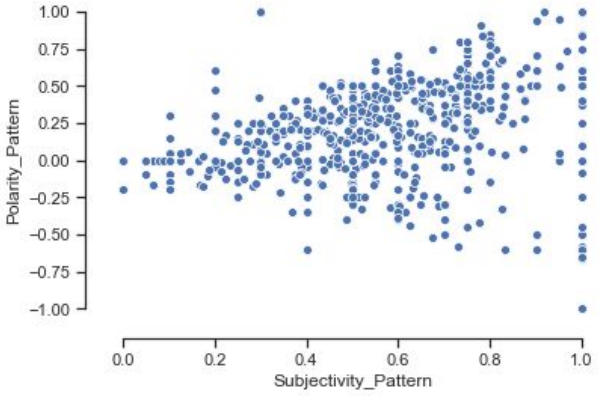

Fig. 6: Polarity and Subjectivity patterns

b) Polarity distributions of customer and blog reviews

Considering that we have two sources of information, can we find differences in polarity distributions of bloggers and customers reviews? Inspecting the boxplots of

customer and blog reviews, outliers were deleted using z score criteria. Figure 7 displays the cumulative polarity distribution for both. A t-test was applied between the two independent samples to measure whether the expected averages are equals, with a significance level of 0.05. The p-value founded for polarity distributions was 0.9483, suggesting that we **can't reject the null hypothesis of identical average score between customers reviews and blog reviews for polarity patterns**. Then, we conclude that there is no significant difference between both distributions.

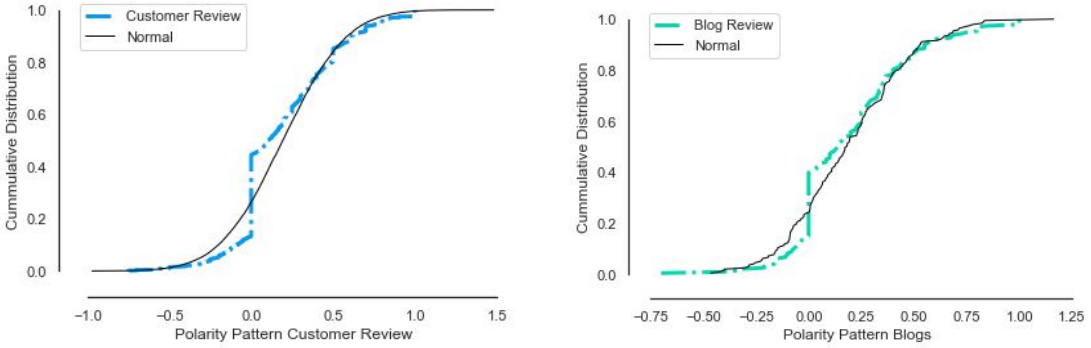

Fig 7: Cumulative distribution of Polarity for customer and blog reviews

c) Periods between positive and negative reviews

Finally, could we find any negative impact of bad scores/reviews on how long take to write the next review for a specific coffee shop?

The difference (in score and time) between consecutive reviews was calculated and we tabulated the difference between scores as *delta score* and the difference in days as the *period between reviews*. All coffee shops with a *period between reviews* more than 15 days were filtered. Following graph (Figure 8) displays the distributions of the most interesting categories. We are curious about the extremes *delta score* because they present the most significant differences. A *delta score* of *+4* represents a first user that put 1 star in his review followed by another that put 5 stars. Delta score of *-3* or *-4* represents the opposite. We are wondering if the period of positive *delta score* is longer than the inverse.

The mean for positive and negative *delta score* is 5.6 days and 4.92 days respectively. Observing distributions in Figure 8, we can see that a *positive delta* score of *+4* has longer periods between reviews that a negative *delta score* of *-4*.

After to compare all the possibles *delta scores* (Figure 9), we conclude that the positive *delta score* of *+4* takes more time than the rest of *delta scores* and it suggests that **the hypothesis of the negative impact can't be rejected**. We don't have evidence to know if the bad scores affect the fluency of people in coffee shops, because we only know, through the dates, when people are writing reviews, but the result at least says that the frequency of reviews could be affected.

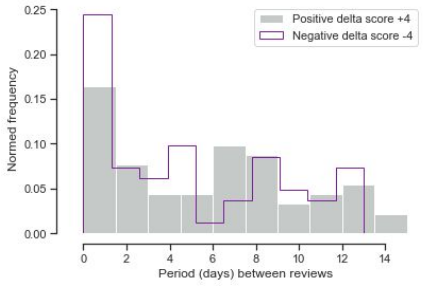

Fig. 8: Histograms of positive and negative delta score of ± 4

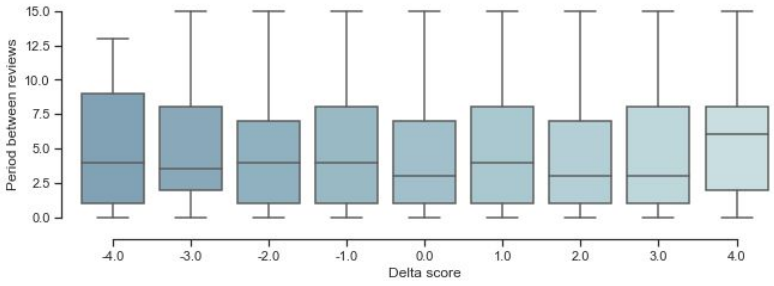

Fig. 9: Boxplots of delta score in all the possible range

IV. In-depth analysis using machine learning

What if we looking for the categories and sections of the coffee topics (previously mentioned) on reviews, compute the sentiments associated with those and predict how much pleasurable or disappointing the experiences are? Could we classify the coffee shops using the keywords of blog-reviews?

1. 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, 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. 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.

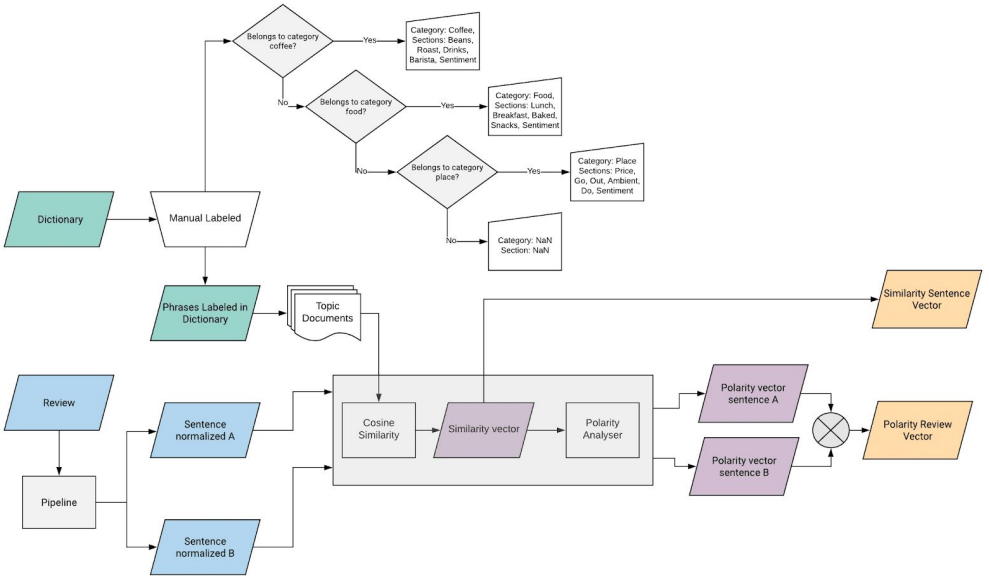

Fig. 10: Pipeline building features vectors

2. Supervised Learning

a) Features selection

A statistic chi-squared test with confidence level of 5% 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 2: Selection of features

b) 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.

c) XG Boost

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 tunning the model using 1000 estimators, 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 apply 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).

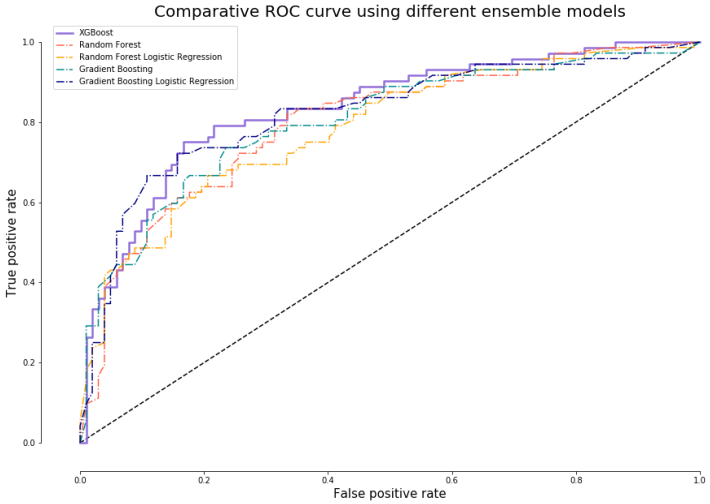

Fig. 11: Comparative ROC of XGBoost, Random Forest and Gradient Boosting combined with Logistic Regression.

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 3: Performance of the different models

The Confusion Matrix is:

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

Table 4: Confusion Matrix for XGBoost model fitting

3. 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.

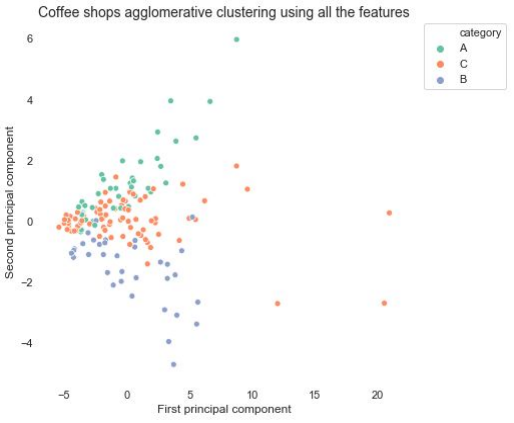

Fig. 12: Agglomerative Clustering of coffee shops using the 16 features

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 above.

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).

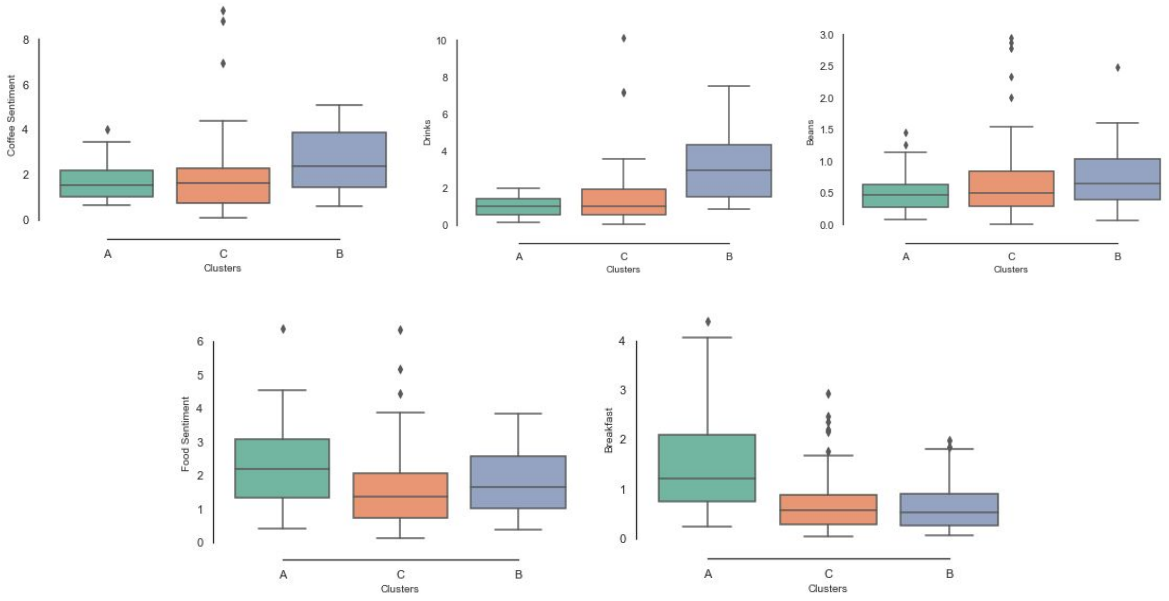

Fig. 13: Representation of features values in boxplots split into the resultant clusters of Hierarchical Clustering

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.

a) 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).

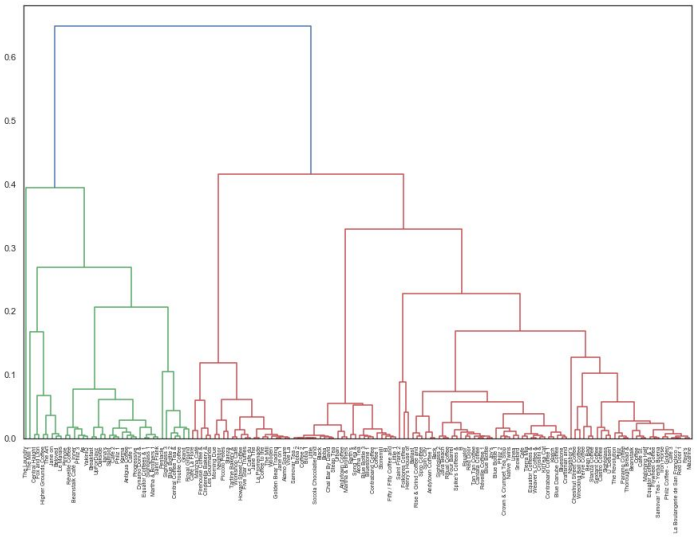

Fig. 14. Dendrogram of coffee shops clustering by coffee features

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.

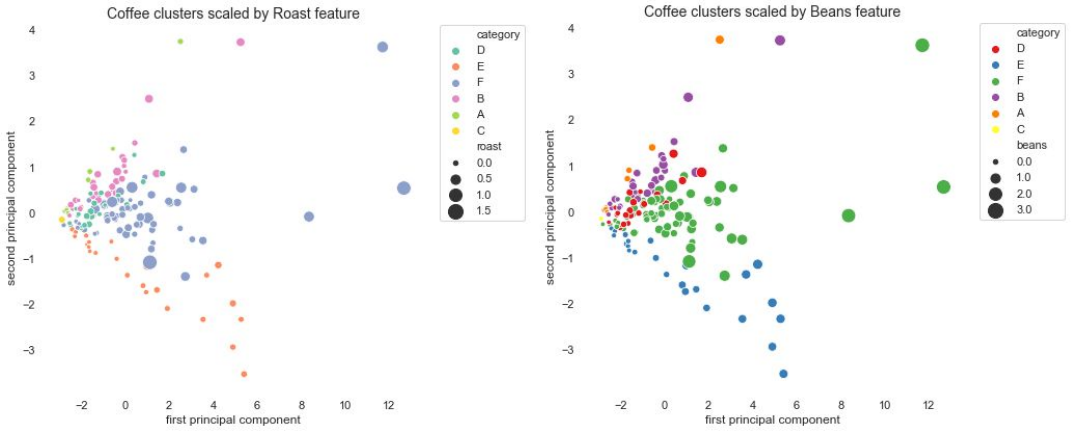

Fig. 15. Clusters using the Coffee Category with the size of points adjusted to different features

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.

b) 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.

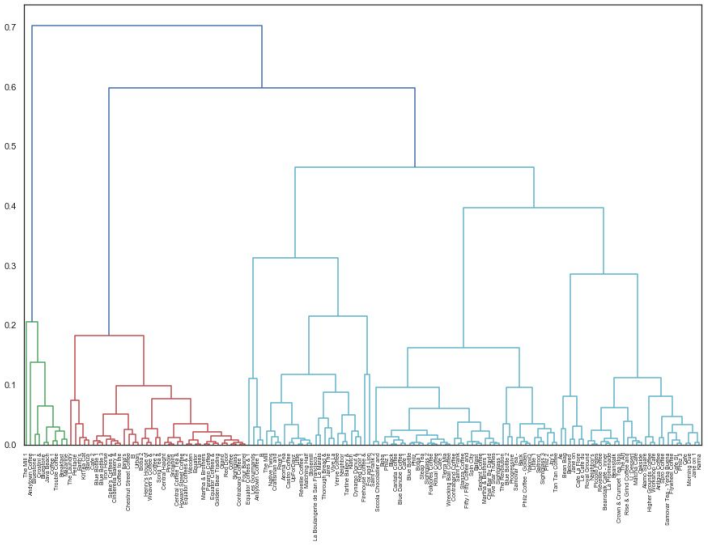

Fig. 16. Dendrogram of coffee shops clustering by food features

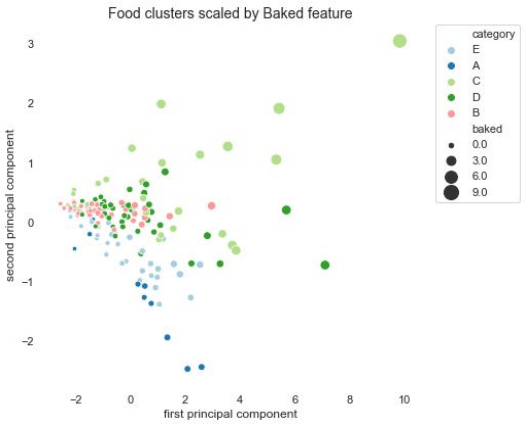

Fig. 17. Clusters using the Food Category with the size of points adjusted to Baked feature

c) 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.

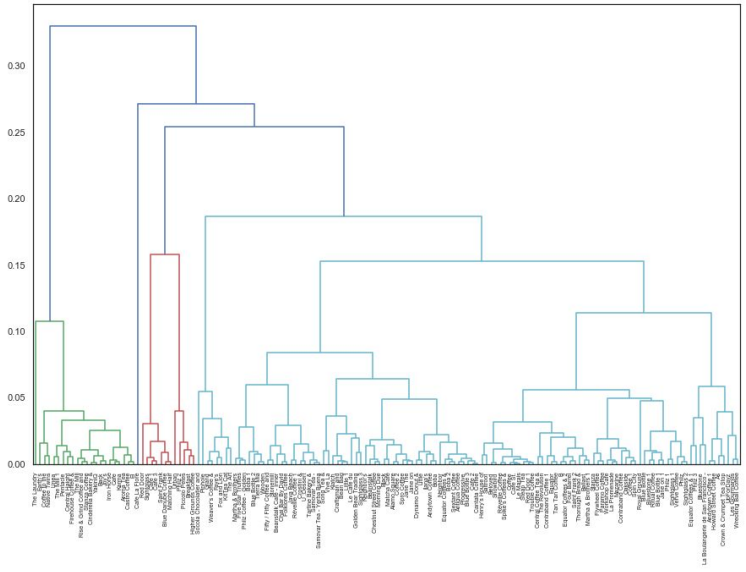

Fig. 18. Dendrogram of coffee shops clustering by food features

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).

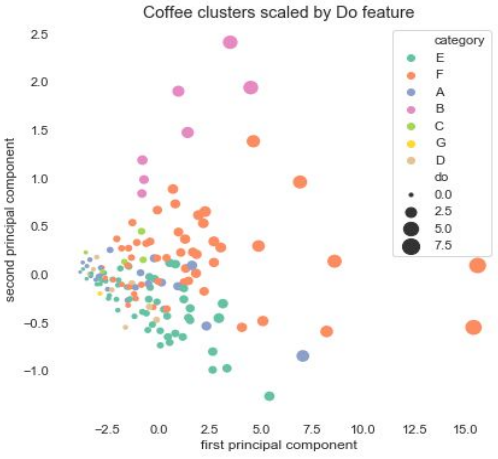

Fig. 19. Clusters using the Place Category with the size of points adjusted to Do feature

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.

V. Conclusions

Different techniques of NLP were applied in this research to extract information from coffee blogs and customer reviews of San Francisco coffee shops in Yelp about the main topics and keywords using TF-IDF, analysis of similarity, polarity and subjectivity patterns.

All paragraphs were split into sentences to measure polarity and subjectivity in each one, after to discover that we could loss emisor intentions using a whole review as a entity. Histograms of polarity and subjectivity patterns proved that less positive and more argumentative messages were found splitting the information. Additionally, we found out more emphasis and strong intention in positive messages, tested with a paired bootstrapping of random samples of sentences applied to calculate the Pearson coefficient between polarity and subjectivity patterns. Another interesting discovery was that the positive *delta score* of +4 takes more time than the rest of *delta scores* and it suggests that the frequency of reviews in a coffee shop could be affected after that someone wrote a negative review assigning 1 score.

We built a dictionary with three principal categories: coffee, food, place and subcategories trying to find topics and patterns in blogs and reviews. A brief inspection of key phrases obtained from blogs indicates that the variety of drinks, baked items (pastries), decoration, vibe and description related to the outside of coffee shop are more frequently mentioned. On the other hand, according to a statistic chi-squared test with a significance level of 5%, the topics more vinculated with the score of customer reviews are sentiments associated with drinks, place, food, breakfast, baked elements, baristas, available of seats, what to do on coffee shops and prices. Topics less relationed are quality and variety of beans, roasting, little snacks and lunch options.

Some machine learning techniques were used from supervised and unsupervised perspectives. Random Forest, Gradient Boosting and an extension of the last, XGBoost were used to predict sentiment patterns in customer reviews. XGBoost got the best performance, with an AUC of 83.4% and an accuracy of 77%. Agglomerative Clustering, dendrograms and PCA as visualization tool, were applied to find clusters with coffee shops distinguishing styles, what to do there and how much interested is the people in talking about beans, roasting, drinks, places and food.

VI. References

1. Yelp API (FUSION) and web scraping from Yelp
2. "*28 of San Francisco's Essential Coffee Shops: Hot spots for your morning cup*" (Ellen Fort and Caleb Pershan). Available here.

3. *"17 of San Francisco's Most Unique Coffee Shops* (Katie Bush). Available here.
4. *"The Best SF Coffee Shops For Getting Work Done"* (Taylor Abrams, Frida Garza, and Will Kamensky). Available here.
5. Coffee gives me superpowers (Ryoko Iwata). Published on April 7, 2015
6. DRIFT San Francisco (A. Goldberg, Velasco, Lee, E. Goldberg and Spicer). Published on July 15, 2018
7. XGBoost: The Excalibur for Everyone (Raghu Raj Rai). Towards Data Science. Available here
8. A Beginner's guide to XGBoost (George Seif). Towards Data Science. Available here
9. Traditional Methods for Text Data (Dipanjan Sarkar). Towards Data Science. Available here
10. Practical Statistics for Data Scientist (Peter Bruce and Andrew Bruce). O'REILLY, 2017.
11. Text Classification is Your New Secret Weapon (Adam Geitgey, Medium). Available here
12. A Practitioner's Guide to Natural Language Processing (Part I) - Processing and Understanding Text (Dipanjan Sarkar). Towards Data Science. Available here