Milestone 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

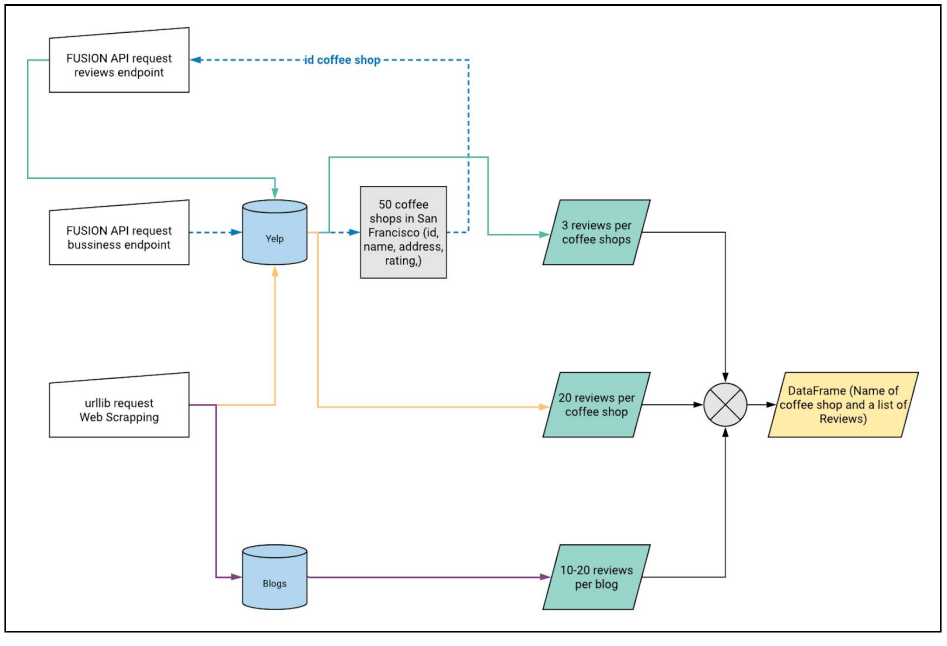

Fig. 1: Data flow from extraction to building of DataFrame

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 scrapping.

The second step is web scrapping 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 per shop. In this section, all HTML tags are removed to extract names and reviews.

b) Blogs Reviews Extraction

Process of data consists of web scrapping 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 key-words 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.

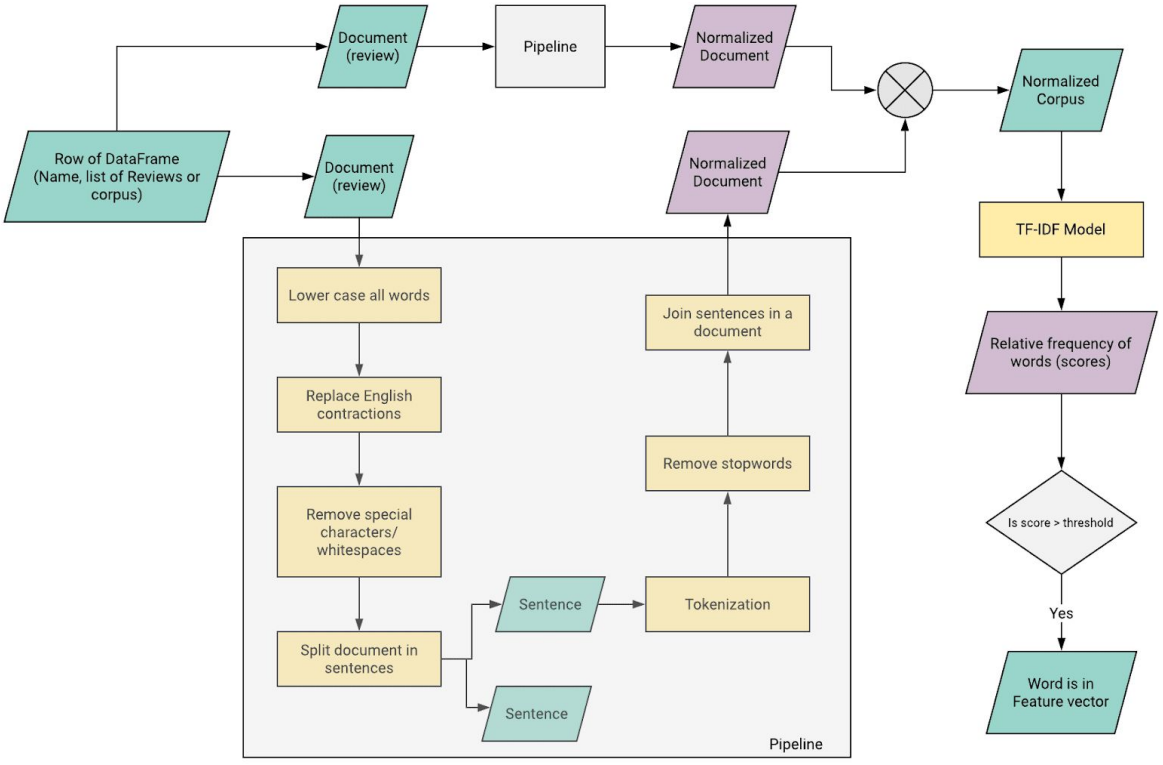

Fig. 2: NLP pipeline of text data

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.

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

a) 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.

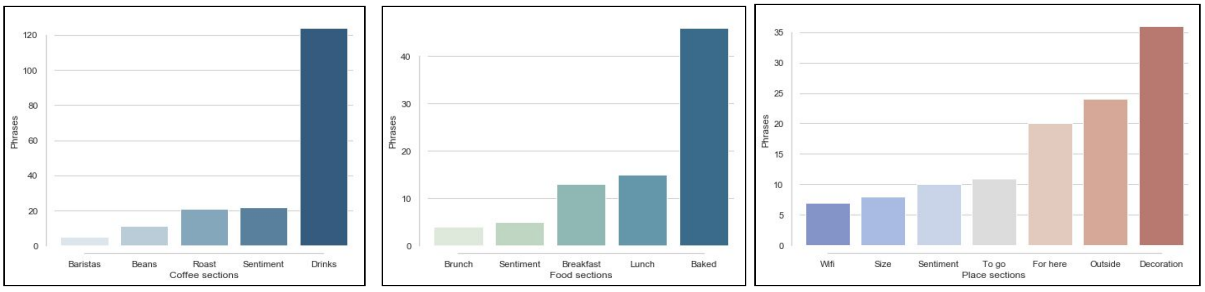

Fig. 3: Barplot of categories and sections built from the key-pairs dictionary

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.

b) 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.

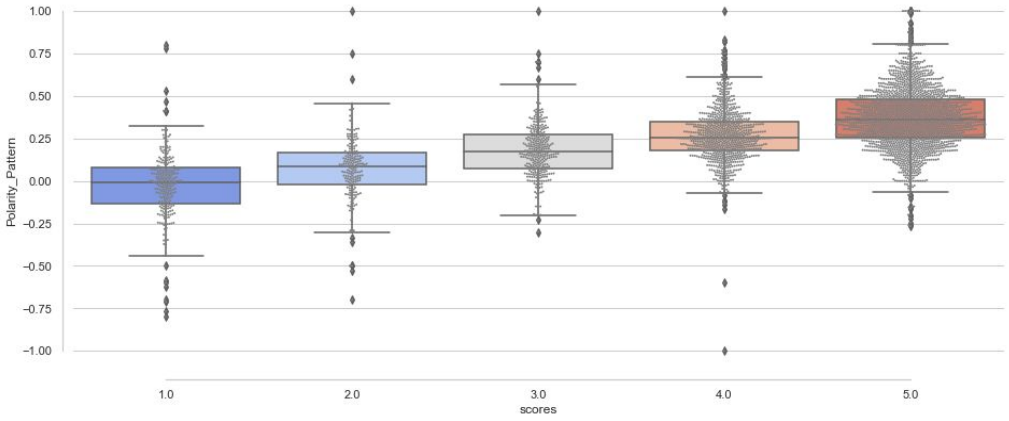

Fig. 4: Polarity pattern through the scores

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

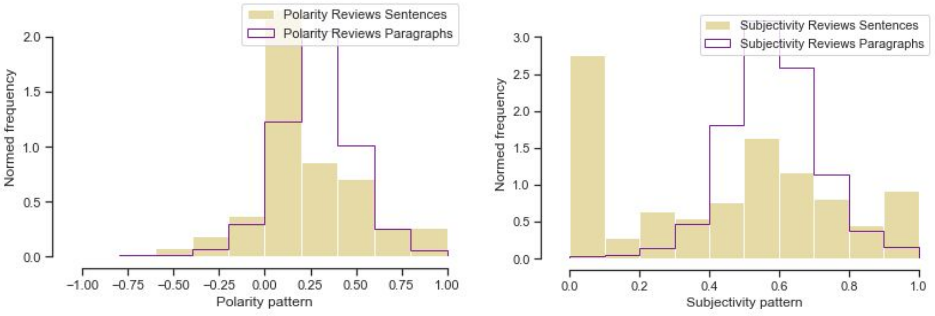

Fig. 5: Histograms of Polarity and Subjectivity to reviews on paragraphs and sentences

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

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:

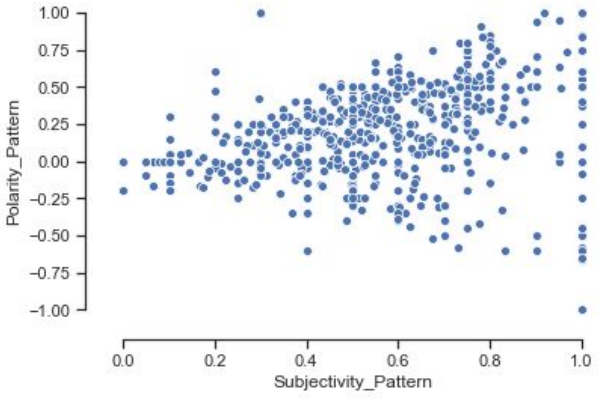

Fig. 6: Polarity and Subjectivity patterns

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.

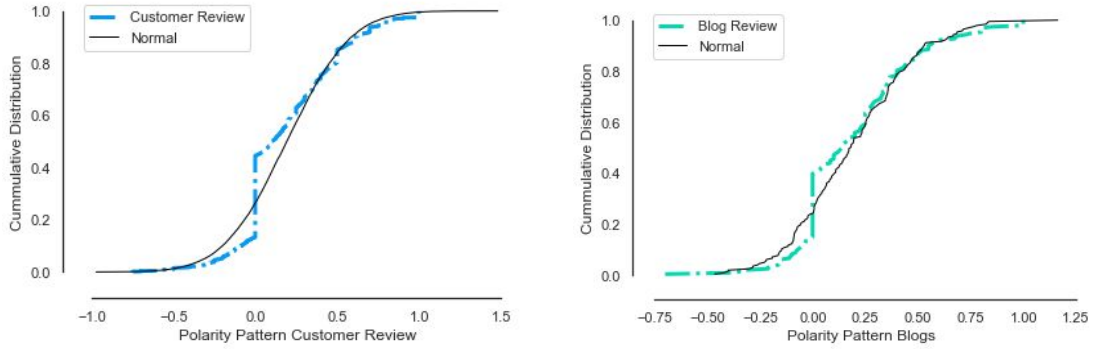

Fig 7: Cumulative distribution of Polarity for customer and blog reviews

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.

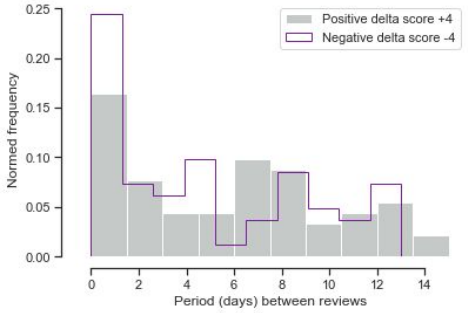

Fig. 8: Histograms of positive and negative delta score of ± 4

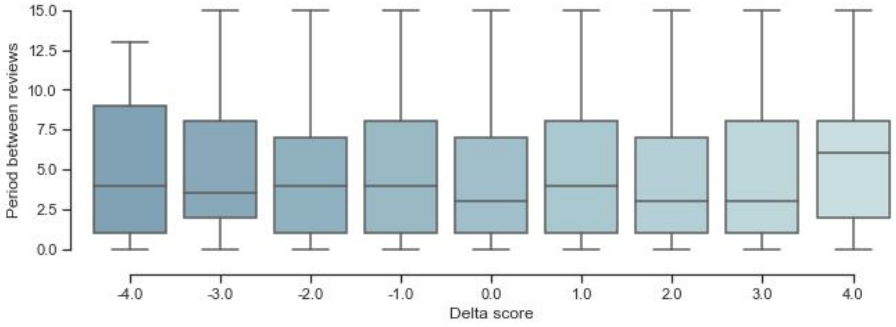

Fig. 9: Boxplots of delta score in all the possible range

IV. Next steps

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 key-words of blog-reviews?