Introduction to Text Mining and Natural Language Processing

Homework-1

Submitted by: Group 8

Deepak Kumar Malik,
Noemi Lucchi,
Tirdod Behbehani

Barcelona School of Economics
Master in Data Science for Decision Making
Master in Data Science for Methodology

February 5, 2025

Contents

1 Part 1: Scraping

1.1 Identify a (future) event that makes a lot of people come to Barcelona. Think about music festivals, local festivities etc.

We identified the 2025 Mobile World Congress as our event. The Mobile World Congress (MWC) is an annual event that was founded in 1994, and it is a leading mobile connectivity event which brings leading brands together alongside groundbreaking technologies and engaging content. The event is known for attracting many high ranking executives. We selected MWC as our event due to the presence of ultra-high-net-worth individuals and tech corporations, whose strong spending power and high demand for hotels during the event can drive up prices.

1.2 Think of the time periods to scrape and what second city to scrape. The second city will be your control group. Explain your choices in writing

The MWC is a four day event, occurring from Monday, March 3rd through Thursday, March 6th. For our second time frame, we selected the dates of Monday, February 24th through Thursday, February 27th. February 24th - February 27th is the week prior to our event, so it will likely match the same consumer as the event. Additionally, the days of the week directly match each other, so those trends will likely be parallel as well.

For our second city (control group), we selected Madrid. Madrid and Barcelona are Spain's two largest cities, and they exhibit many common traits. Both cities are economic and cultural hubs, international and multicultural cities, have similar climates and culinary scenes, among other similarities.

We initially considered Valencia as a control city but ultimately selected Madrid instead, as it is more similar to Barcelona in terms of commerce and tourism.

1.3 Design a careful scraping pipeline that follows the advises seen in class and TAs.

The designed web scraping pipeline is created using a headless Firefox browser—configured with functions like `ffx_preferences` and `start_up`—to efficiently collect data without a graphical interface, thus reducing resource usage and speeding up the process. It automatically handles pop-ups and cookie banners via `close_genius_popup` and simulates user interactions using `search_city_and_dates` to set search parameters such as city and date ranges accurately. The pipeline dynamically loads all hotel listings using `scroll_down` and `scrape_all_results`, managing infinite scrolling and "Load More" actions while avoiding duplicates, and further accelerates data enrichment by concurrently fetching hotel descriptions with `add_descriptions_to_hotels` using Python's `ThreadPoolExecutor`. This com-

prehensive approach ensures robust and reliable data collection in approximately 11 minutes. To use the pipeline, update the download folder and geckodriver path in the `main` function, modify the `city_date_ranges` with your desired cities, date ranges, and output CSV filenames, and then run the `main()` function; detailed explanations and the rationale behind each component are provided in the file **web_scraping_pipeline.ipynb**.

2 Part 2: Text Analysis

2.1 Pre-process the text by removing stop words and stemming. Customize your stopword list if needed.

In order to run our regressions for Part 3, we preprocess the text to improve its usability. Our general goal is to:

1. **Input and remove custom stopwords:**

   We first input a list of words that we believe are extraneous and do not add any value to our analysis. We define this list with personalized stopwords, such as `['el', 'la', 'con', 'se', ...]`. We will remove these custom stopwords from our descriptions.

2. **Convert text to lowercase:**

   This enables us to treat words like "Metro" and "metro" as the same word.

3. **Perform tokenization:**

   Tokenization allows us to split an entire description into a list of all the words in the description. This allows us to process each word individually.

4. **Apply stemming:**

   Stemming removes suffixes from words, which enables us to view words in their root form. For instance, stemming would allow us to treat "habitaciones" and "habitación" as just "habitacion".

5. **Store our preprocessed descriptions in a new column within our dataframe:**

   After preprocessing, we will store the cleaned descriptions in a new column in our dataframe for further analysis.

See part 2 of **text_analysis_DiD.ipynb** for a full overview of our natural language processing.

2.2 Create two word clouds before and after pre-processing for each city (a total of four). Comment on the changes in the word clouds.

In this section, we compare the word clouds generated from hotel descriptions before and after text preprocessing. The goal is to observe how removing stopwords and applying stemming affects the most frequent words appearing in the dataset.

Figure 1: Raw Descriptions (Barcelona)

Figure 2: Raw Descriptions (Madrid)

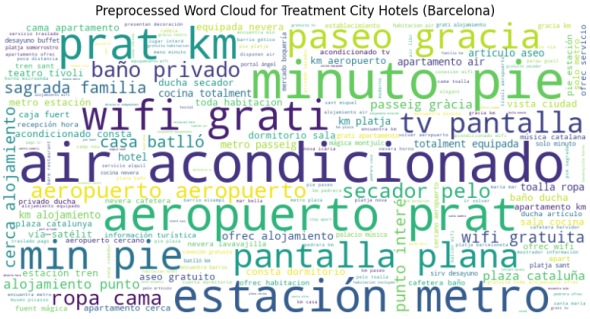

Figure 3: Preprocessed Descriptions (Barcelona)

Figure 4: Preprocessed Descriptions (Madrid)

After preprocessing, we observe notable changes in the word clouds:

- Many common but non-informative words (e.g., "Barcelona", "Madrid", "de", "la", "en", "el") are removed.
- Words representing hotel features, such as "estación metro", "air acondiocionado", "min pie", become more prominent.
- Some words are reduced to their root forms due to stemming (e.g., "habitaciones" and "habitación" both become "habitacion"). However, we can see that our stemming is not perfect, as "minuto pie' and "min pie" are treated as two separate bigrams in Figure 3.

Most importantly, we can clearly see that natural language processing allows us to better highlight and emphasize hotel features. We will use our preprocessed descriptions in our regressions in Part 3.

3 Part 3: Differences-in-differences

3.1 Write down a fixed effects regression equation that allows you to derive a difference-in-difference estimate of the effect of the event on prices. Think of controls to add, why is this relevant? Explain why you need a second city for this.

1. **Simple Diff in Diff**

$$\text{Price}_{it} = \beta_0 + \beta_1 \cdot \text{Treated}_i + \beta_2 \cdot \text{Time}_t + \beta_3 \cdot (\text{Treated\_Time}_{it}) + \epsilon_{it}$$

We regress the price on the two dummies that indicate the treatment status and the period (pre-treatment or post-treatment) and their interaction. We used this setting as a starting point, which we further improve based on the data availability and the more advanced techniques that help address possible biases.

2. **Diff in Diff with controls**

$$\begin{aligned}\text{Price}_{it} = &\beta_0 + \beta_1 \cdot \text{Treated}_i + \beta_2 \cdot \text{Time}_t \\ &+ \beta_3 \cdot (\text{Treated\_Time}_{it}) + \beta_4 \cdot \text{Rating}_{it} + \beta_5 \cdot \text{Number\_Ratings}_{it} + \epsilon_{it}\end{aligned}$$

We add Rating and the Number of Reviews as controls, given that these two features can determine some price discrimination, although there are many other factors that can lead to differences in price.

3. **Diff in Diff with fixed effects**

$$\begin{aligned}\text{Price}_{it} = &\beta_0 + \beta_1 \cdot \text{Time}_t + \beta_2 \cdot (\text{Treated\_Time}_{it}) \\ &+ \mu_i + \epsilon_{it}\end{aligned}$$

Instead of adding controls, we use the Fixed Effects method to address the omitted variable bias. This method captures all the factors that vary across units but are constant over time, representing a more robust way of controlling for omitted variables compared to the inclusion of Rating and Number of Reviews. Here, μ_i represents the individual fixed effect for the hotel i.

Generally speaking, adding controls helps reduce the omitted variable bias, preventing us from wrongly attributing the effect of other factors to the treatment. The omitted variable bias occurs when variables that influence the response variable, while being correlated with the regressors, are omitted. This leads to a violation of the orthogonality assumption, mathematically $\mathbb{E}[\epsilon \cdot X] = 0$, and consequently, the estimated coefficients do not represent the true causal effect.

In this case, Rating and Number of Reviews certainly influence the price but are only weakly correlated with the regressors, particularly the treatment dummy. Indeed, we have found that Barcelona systematically has higher ratings and a greater number of reviews. Given the low correlation between the controls and the regressors, while not including them would not lead to a biased estimation of the causal effect, including them increases the precision of the estimates.

Fixed Effects achieves the same goal — reducing the omitted variable bias — but it does so in a more comprehensive way. With these models, we capture the individual fixed effects that vary across units but are constant over time. The individual effect captures everything that is different between one hotel and another, helping to isolate the causal effect of interest when it is not possible to explicitly control for all these differences by including the actual variables.

The Simple Diff in Diff model estimates that the Mobile World Congress led to an increase in hotel prices by 961€ over the four-day time period. When adding controls, this effect decreases to 923€, and with the Fixed Effects model, it further drops to 916€. These results align with the theoretical framework discussed earlier:

- The higher estimate obtained with the Simple Diff in Diff suggests the presence of omitted variables, whose influence is partially captured in the treatment effect.
- Adding Rating and Number of Reviews as controls reduces the estimate, as these variables explain part of the price variation that would otherwise be attributed to the treatment.
- The Fixed Effects model further reduces the estimate, as it accounts for all time-invariant differences across hotels, providing a more precise identification of the causal effect.

3.2 How would you use text features from the description as controls? Think about the text in the descriptions you scraped. How would this help? Why would terms like "Barcelona" not help?

- **Diff in Diff with text features as controls**

$$\begin{aligned}\text{Price}_{it} = \beta_0 &+ \beta_1 \cdot \text{Treated}_i + \beta_2 \cdot \text{Time}_t \\ &+ \beta_3 \cdot (\text{Treated\_Time}_{it}) + \beta_4 \cdot \text{control\_words}_{it} + \epsilon_{it}\end{aligned}$$

We selected control words that can help discriminate between cheaper and more expensive hotels. Examples include 'hotel' (since hotels are generally more expensive than hostels or apartments), 'zona', 'ciudad', 'pie estación', 'restaurant' (as proximity to specific areas can determine higher prices), 'balcón', 'recepción', 'servicio', 'tv pantalla', 'terrazza', 'vista', and 'apartamento air', 'piscina', 'desayuno buffet' (these features and amenities increase the price).

After selecting these control words, we created a dummy variable that equals 1 if a hotel's description contains any of these words and 0 otherwise. The coefficient associated with this dummy variable indicates that, on average, hotels featuring these words in their descriptions are €86 more expensive than those that do not.

However, the causal effect estimated for the treatment remains at 962€. This is likely because, although these words influence the price, they are not systematically correlated with the other regressors. Specifically, there are no significant systematic differences between Barcelona and Madrid or between the pre-treatment and post-treatment periods in terms of the presence of these words. Therefore, their inclusion as controls leaves the estimated causal effect as in the Simple Diff in Diff.

Words like "Barcelona" would not be helpful as they are common to all treated units, making them irrelevant for extracting information that differentiates between hotels. Such words do not contribute to identifying additional variables, beyond the treatment, that may explain price variation. The information conveyed by the word "Barcelona" is already captured by the condition "Treated = 1," rendering it redundant and potentially multi-collinear.

3.3 Now suppose we want to decompose the treatment effect by hotel quality. Can you use the text description here? How would you use them to study heterogenous treatment effects? Write down a regression equation.

To estimate heterogeneous treatment effects and decompose the treatment effect by hotel quality, we employed the same set of words previously specified in the regression. These words were chosen because they typically identify high-quality hotels, which are often more expensive. Based on this set of words, we created a dummy variable that equals 1 if a hotel's description contains any of these words and 0 otherwise. This dummy variable serves as a proxy for high-quality or "luxury" hotels. To capture the variation in treatment effects across hotel quality, we interacted this dummy with the term representing the causal effect — the interaction between the treatment dummy and the time dummy.

- **Difference-in-Differences with heterogenous treatment effects**

$$\begin{aligned}\text{Price}_{it} = &\beta_0 + \beta_1 \cdot \text{Treated}_i + \beta_2 \cdot \text{Time}_t \\ &+ \beta_3 \cdot (\text{Treated\_Time}_{it}) + \beta_4 \cdot \text{control\_words}_{it} \\ &+ \beta_5 \cdot (\text{Treated\_Time}_{it} \cdot \text{control\_words}_{it}) + \epsilon_{it}\end{aligned}$$

This regression estimates a causal effect of 733€, with a β_5 equal to 243€. This implies that, on average:

- When control_words = 0, the event leads to an increase in prices by 733€.

- For high-quality hotels, identified by control_words = 1, the increase is larger:

$$733€ + 243€ = 976€$$

This finding suggests that the event has a differential impact on prices depending on hotel quality, with high-quality hotels experiencing a greater price increase. This result aligns with the hypothesis that higher-quality hotels, often catering to more affluent customers, can pass through a larger share of the event-driven demand shock to their pricing.
In this way, the model successfully captures heterogeneity in treatment effects, assuming that the event's impact varies across units and is more pronounced for more expensive hotels.

Lastly, when combining the two effects - the baseline increase in average price due to the event and the additional price increase experienced by high-quality hotels — we obtain a total effect that aligns with the initial estimate derived from the simple Diff in Diff model. The consistency of the results across different model specifications confirms the robustness of the estimated causal effect.

A more detailed data analysis can be found in part 3 of **text_analysis_DiD.ipynb**.