Exercise Homework 2

February 12, 2025

This homework can be made a lot easier by choosing a corpus we have seen in class (e.g. AP corpus on Spain or the security council paragraphs) and apply an easy target. However, we will give some points for creativity in the pipeline/question design so make sure to come up with a coherent and creative question. Please read all the instructions carefully before starting.

1. Get together in groups as randomized here. (Link to Google Sheet)
2. Download the material for the homework. Prove work that you did by attaching your code.
3. Due date is 25th of February.
4. **Submission must be done as follows:**
   - One single .zip file per group.
   - File name should be on the format *group#_surname1_surname2.zip*; with all the surnames of the members included.
   - Include codes and pdfs and upload to google class (only a single submission per group).
   - In case you have problems coordinating work with your teammates or IT problems let us know as soon as possible.

Mini Research Project

We will practice project pipelining in the example of a simple pipeline to develop a dictionary generation. You need a corpus and an idea of what type of vocabulary you want to track in this corpus using a dictionary method. You need to identify a sample of text for the type of vocabulary you want to track so keep this in mind. Best to read all questions first and then think about corpus-vocabulary-question together.

When running out of time: The answers to questions 5. and 8. can be easily adapted to be more or less involved/ambitious. When you have very little time come up with a fast question ex-post, i.e. after implementing some vocabulary and looking for broad patterns in the country you generated. Failure to generate good dictionaries or interesting findings is totally fine if you try to explain in your answer to question 8.

Part 1: Getting Main Text Data

Get a dataset from somewhere. You can pick one from the class corpora, scrape something, get something from Kaggle or other available databases like Google books. Before deciding what to take, read requirements/tasks below.

1. Make sure the text is diverse enough that it is covering different topics. Longer documents will work better but there is also ways in which documents like paragraphs and tweets can be made to work.

2. Spend some time cleaning the corpus and making sure you understand the structure of the text and the metadata. Demonstrate this by producing summaries of the text metadata (by category, over time) and show some examples of the text.

Part 2: Develop Methodology

3. Make sure you understand the use of methods like Tf-idf to generate dictionaries from example texts. Pay attention to detail like the weight given to different parts of dictionaries in Gentzkow and Shapiro (2010) or Hassan et al (2019). Read the articles Gentzkow and Shapiro (2010), Hassan et al (2019) and Garcia-Uribe (2024) if you want to understand the context better.

4. Decide on what type of content you want to track and how. It is important to note that Gentzkow and Shapiro (2010) use labels of party affiliations to generate dictionaries. The other two references use types of text to generate labels. Follow the second strategy and identify types of text you want to use to generate your dictionaries. Be creative on where to get the type of text from and make sure you have at least two types of text. In Hassan et al (2019) there are just two types P and N, in Garcia-Uribe et al (2024) there are four/five types. Your method will need to adapt to this.

5. Make your expectations explicit. Why would one care to track the type of vocabulary in your corpus? What interesting questions could be answered with this?

Part 3: Implementation

Implement your method:

6. Generate dictionaries and make a nice table showing these dictionaries for human interpretation. Explain why you think these dictionaries capture what you want to capture. Are your expectations fulfilled?

7. Show something interesting using the dictionary counts in the corpus - at minimum show dictionary count distributions across the entire corpus and by metadata subgroups.
8. Answer the questions you set out to answer.

References

Gentzkow, Matthew, and Jesse M. Shapiro. "Media Bias and Reputation." Journal of Political Economy 114, no. 2 (2006): 280–316.

García-Uribe, Sandra, Hannes Mueller, and Carlos Sanz. "Economic Uncertainty and Divisive Politics: Evidence from the dos Españas." The Journal of Economic History 84, no. 1 (2024): 40–73.

Hassan, Tarek A., Stephan Hollander, Laurence van Lent, and Ahmed Tahoun. "Firm-Level Political Risk: Measurement and Effects." The Quarterly Journal of Economics 134, no. 4 (2019): 2135–2202.