# **README for Political Comments Analysis Script**

## **Prerequisites**

### **Software and Libraries**

* Python: Ensure you have Python installed.
* Libraries: This script requires several libraries including zstandard, orjson, csv, logging, pandas, networkx, nltk, matplotlib, and gdown.

### **Installation of Libraries**

Use the following command to install all necessary libraries:

!pip install datetime zstandard os orjson csv logging pandas networkx nltk matplotlib gdown

For NLTK’s vader\_lexicon, run:

import nltk

nltk.download('vader\_lexicon')

## **Configuration**

### **Input File**

* Place your Zstandard-compressed JSON file (e.g., politics\_comments.zst) in the same directory as the script.

### **Script Parameters**

* Input File Name: Specify the name of your input file within the script.
* Fields: Define the fields to extract from each JSON object in the script.
* Date Range: Set the start\_year and end\_year variables to filter the comments.

## **Execution**

Run the script in a Python environment. The script performs the following operations:

1. **Reading and Decoding**: It reads and decodes chunks from the Zstandard-compressed file.
2. **Filtering Comments**: Filters comments based on the specified date range.
3. **Writing to CSV**: Outputs the filtered data to CSV files, one for each year in the specified range.
4. **Sentiment Analysis**: Performs sentiment analysis on comments using NLTK’s SentimentIntensityAnalyzer.
5. **Data Processing**: Processes data in pandas DataFrames and performs various manipulations including splitting IDs and creating connections dataframes.
6. **Graph Creation**: Builds directed graphs from the connections data using NetworkX for interaction analysis.

## **Other Features**

* **Google Colab Integration**: The script includes functions to download and process data using Google Colab, utilizing Google Drive for data storage.
* **Data Visualization**: Utilizes matplotlib and networkx for visualizing network graphs.
* **Memory Management**: Includes garbage collection to handle large datasets efficiently.

## **Usage Tips**

* Ensure adequate system resources due to the script’s intensive data processing.
* Modify the start\_year and end\_year parameters as needed for different datasets.

---------------------------------------------------------------------------------------------------------------------------

Run secondpart.ipynb script which perform SNA metrics and community detection on graphs from 2007 to 2022

Finally, before run textanalys.ipynb change the directory\_path variable in this notebook and enter the path to your folder. Also, make sure you have en\_core\_web\_sm installed on your machine (otherwise run this line in command prompt: python -m spacy download en\_core\_web\_sm)

This notebook perform keywords analysis, polarization assessment, emotion detection and NER analysis.