Analysis of the Greek Parliament Proceedings 1989 – 2020

M.Sc. on Data and Web Science
Technologies for Big Data Analytics
Assignment 2021

DESCRIPTION

In this assignment we are going to perform techniques for big data analysis over a dataset that has been created by using the Greek Parliament Proceedings. The dataset was generated by crawling the corresponding website of the Greek Parliament

https://www.hellenicparliament.gr/Praktika/Synedriaseis-Olomeleias

This dataset includes 1,280,918 speeches (rows) of Greek Parliament members with a total volume of 2.30 GB, that were exported from 5,355 parliamentary sitting record files. They extend chronologically from early July 1989 up to late July 2020. The dataset consists of a .csv file in UTF-8 encoding and includes several columns related to the member, the party, the date of speech etc. More information about the data can be found in the following Github repo:

https://github.com/iMEdD-Lab/Greek_Parliament_Proceedings

Although you may think that the dataset is not that large, it is fine for our purposes since the techniques that we will apply can be also used in larger datasets. In particular, we will use Apache Spark with the Scala programming language and the available libraries, to develop algorithmic techniques to analyze the dataset in multiple ways.

TASKS TO IMPLEMENT

Below there is a list of tasks that you should handle.

Task1. Given all speeches (for all years) we need to detect the different topics (i.e., thematic areas). For each topic we need also to know the most representative keywords. Moreover, we need to see how the topics change across years.

Task2. Given all speeches we need to detect pairwise similarities between parliament members. In particular, we need to find a way to extract a feature vector for every member and then perform pairwise similarities to be able to detect the top-k pairs with the highest degree of similarity (k is a parameter).

Task3. For each member and also for each party we need to detect how the most important keywords evolve across years.

Task4. The financial crisis (assuming that it started around 2009, 2010) was a major “checkpoint” in history, not only for Greece but for other countries as well. Can we detect any significant deviation (per member, per party or in general) with respect to the speeches before and after the crisis?

Task5. Taking into account all speeches, we need to detect if we can group them in meaningful clusters. Then we may check about the participation of each member in each cluster and also the participation of each party in the cluster.

Task6. Here you should define your own analytical task to apply. This task should be interesting with non-trivial results. Each team should come up with an additional task, ideally different from the tasks of other teams.

GUIDELINES

In cases where the original dataset seems too large, you may use a sample of the data. You should form teams with 2 or 3 persons. In the class we are 28 persons, so ideally we should have around 10 teams. (*Note: The bibliographic assignment will also be implemented by the same teams, due to a large number of students. More details about the bibliographic assignment will be available on e-learning*). The project deliverables are: 1) the source code with comments, 2) the technical report summarizing your work and 3) the slides for your presentation. Each team will have around 20-25 minutes to present the work in class. This project takes 40% of the total grade. The presentation of the projects will be performed during the last lectures of the course, normally the week 31/1/2022 – 4/2/2022 depending on availability and the situation with respect to the COVID-19 pandemic. With this project you have the chance to work with an interesting dataset and elaborate on the techniques that we discuss in class. Also, **teamwork is very important** for your future career. Last but not least, have fun with the project. Try to uncover interesting things!

Good luck,
a.p.