Summery classification

Mohammad hossein Rajabi

June 10, 2023

1 Source of Dataset

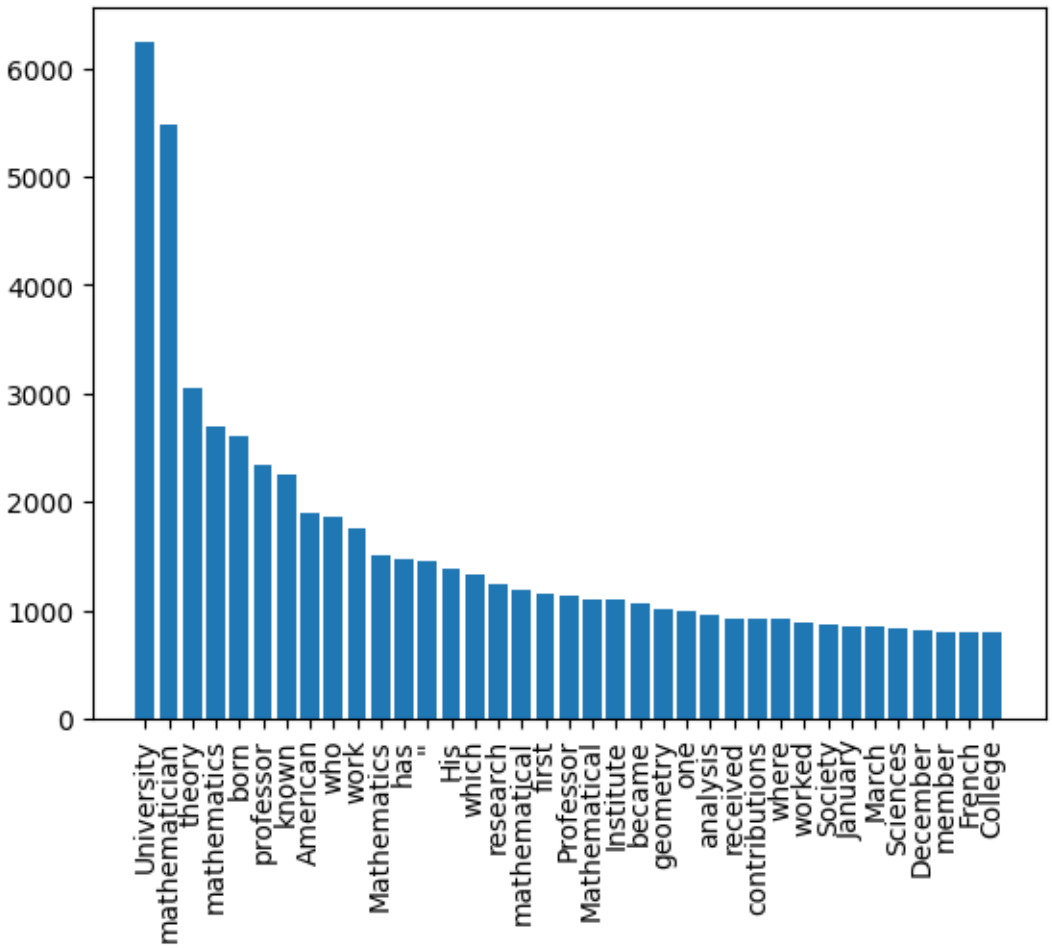
The following sources were used to collect this dataset:
mathematicians data : https://www.kaggle.com/datasets/joephilleo/mathematicians-on-wikipedia
politicians data: http://everypolitician.org/united-states-of-america/house/download.html

2 Data Collection method

To collect the main dataset, we extracted the names of the people using the above csv files and obtained a summary for each of them using the wikipedia api and saved them in a csv file along with their tags.
(tag , name , summary)

3 data statistics

Number of sentences :The file has 21966 sentences
Number of word : The file has 557425 words
Unique word count : The files have 49784 unique words
Common words between tags: files have 10166 common words
Histogram of the number of occurrences of each unique word in order from high to low frequency :



4 Github

In the link below, you can access all the datasets and written codes.
https://github.com/rajabi78/NLP$_{P}roject/tree/main$