Homework #1

Introduction of Text Mining
NTU, Fall 2023

Outline

- Programming Assignment 1
- Tools
- Submission
- Homework policy

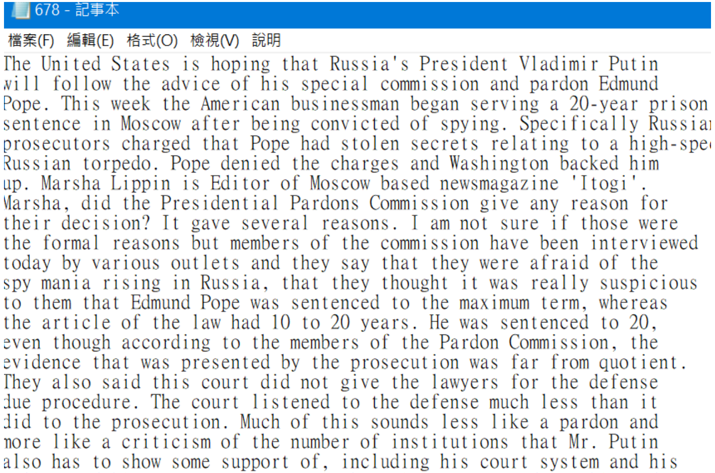
Outline

- **Programming Assignment 1**
- Tools
- Submission
- Homework policy

Programming Assignment 1

1. Write a program to convert a set of documents into TF-IDF vectors

- Text collection: 1095 news documents
- 1.txt ~ 1095.txt
- **Use `TfidfVectorizer`**
  - Lowercase everything
  - Filter out English stopwords



The United States is hoping that Russia's President Vladimir Putin will follow the advice of his special commission and pardon Edmund Pope. This week the American businessman began serving a 20-year prison sentence in Moscow after being convicted of spying. Specifically Russian prosecutors charged that Pope had stolen secrets relating to a high-spe Russian torpedo. Pope denied the charges and Washington backed him up. Marsha Lippin is Editor of Moscow based newsmagazine 'Itogi'. Marsha, did the Presidential Pardons Commission give any reason for their decision? It gave several reasons. I am not sure if those were the formal reasons but members of the commission have been interviewed today by various outlets and they say that they were afraid of the spy mania rising in Russia, that they thought it was really suspicious to them that Edmund Pope was sentenced to the maximum term, whereas the article of the law had 10 to 20 years. He was sentenced to 20, even though according to the members of the Pardon Commission, the evidence that was presented by the prosecution was far from quotient. They also said this court did not give the lawyers for the defense due procedure. The court listened to the defense much less than it did to the prosecution. Much of this sounds less like a pardon and more like a criticism of the number of institutions that Mr. Putin also has to show some support of, including his court system and his

Programming Assignment 1

2. Save each document's TF-IDF vector as a plain text file

  - 1.vec, 2.vec, … 1095.vec

3. Load the TF-IDF vectors of documents 1 and 2, and calculate their cosine similarity.

- Please zip the vector files of documents 1 and 2, your code, and a report (indicating the similarity of doc 1 and doc 2) to TA.
- You have 2 weeks to do the work, deadline is **2023/10/8**.

Outline

- Programming Assignment 1
- **Tools**
- Submission
- Homework policy

Tool for Dataset

- **Download the dataset**
    - Manually download the dataset here

      https://cool.ntu.edu.tw/courses/28998/files/4425061?wrap=1

Tool - Google Colab Tutorial

- **We suggest everyone use Google Colab to finish HW and provide some resources for someone who is not familiar with Google Colab.**
    - Introduction to Google Colab
    - Google Colab Tutorial
    - How to connect to your drive with google colab

Outline

- Programming Assignment 1
- Tools
- **Submission**
- Homework policy

Submission

- Deadline: **2023/10/8 (Sun.) 23:59 (GMT+8)**
- Your submission should include the following files:
  - D11725003
  - ├ report.pdf
  - ├ pa1.*(.py, .ipynb)
  - ├ output/
  - ├ 1.vec
  - ├ 2.vec
- Ensure your code can be executed successfully on **google colab** before your submission.
- You shall **NOT** hardcode any path in your python files, while the dataset given would be the absolute path to the directory.

Outline

- Programming Assignment 1
- Tools
- Submission
- **Homework policy**

How to find help

- Google!
- Post your question to NTU COOL
- Contact TAs by e-mail: d11725003@ntu.edu.tw