資訊檢索與文字探勘導論 作業二

資管三 陳彥廷

2023 年 9 月 30 日

Environment

Jupyter Notebook

Programming Language

I use pyenv-win and **Python 3.11.4**.

Execution Procedure

1. Setup your python environment.
2. Install **nltk**、**numpy**.
3. Execute the code in **Jupyter Notebook**.

Explanation

preprocessing

1. Tokenization: I use **text.spilt()** to tokenize the text.
2. Lowercasing everything: I use **text.lower()** to lowercase the text.
3. Stopword removal: I get the stopword list from Link and remove the stopword from the text.

4. Remove punctuation: I use **re.sub()** to remove some punctuations.
5. Stemming using Porter's algorithm: I use **PorterStemmer** from **nltk.stem** to stem the text.
6. Stopword removal: I use stopword list again to remove the stopword from the text.

TF-IDF

1. Tf: I calculate the term frequency from scratch.
2. Idf: I calculate the inverse document frequency from scratch.
3. Unit vector: I use **numpy.linalg.norm()** to calculate the unit vector.

Cosine Similarity

1. I calculate the inner product from scratch.
2. I use **numpy.linalg.norm()** to calculate the vector.
3. divide the inner product by the product of the vector.

Result

The cosine similarity between document 1 and document 2 is **0.196**.