# Problem Set 3 - Natural Language Processing (UET)

https://github.com/CoderHung/Problem-set-3-Natural-language-processing-class-UET

## Overview

This project explores five topics:

- **Science**
- **Politics**
- **Nature**
- **Sports**
- **Fashion**

This project uses the term-document matrix, TF-IDF, and t-SNE to visualize word vectors.

## Files

- **Term-Document Matrix**: `TD.csv`
- **TF-IDF**: `TF-IDF.csv`
- **t-SNE Visualization**: 150 random word vectors from TD matrix projected on a 2D plane

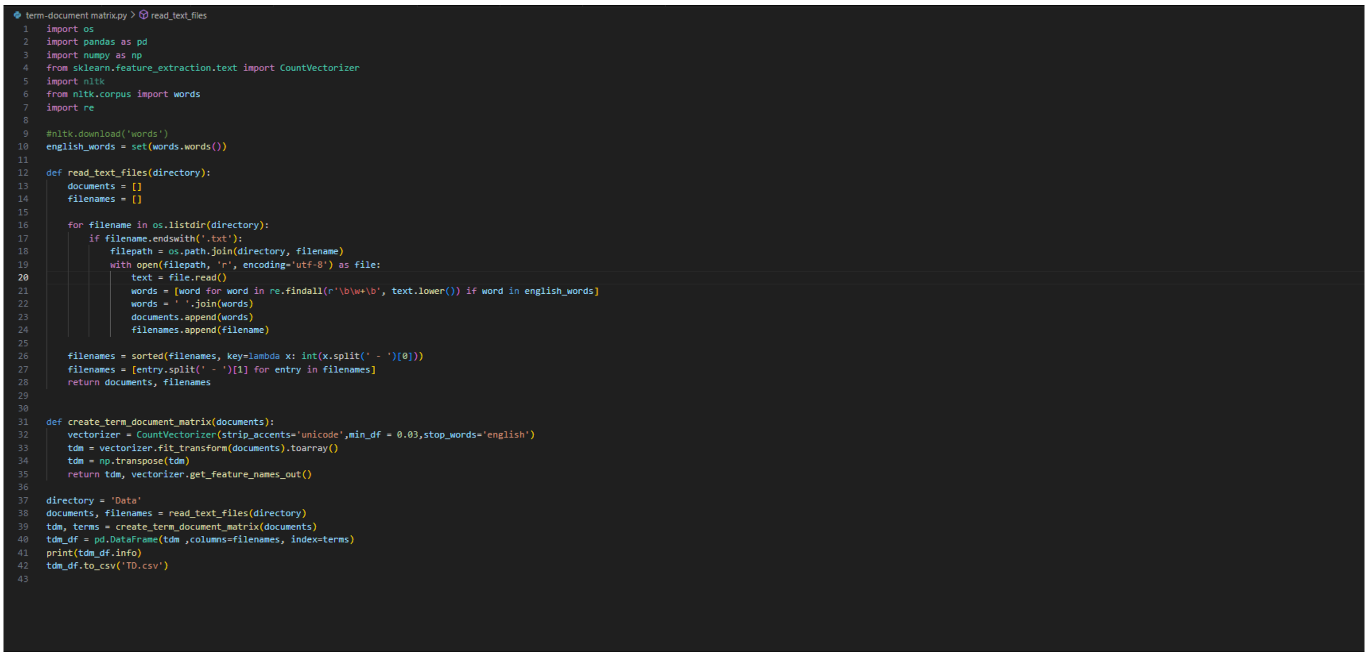
## Instructions

### Step 1

Unzip the dataset:

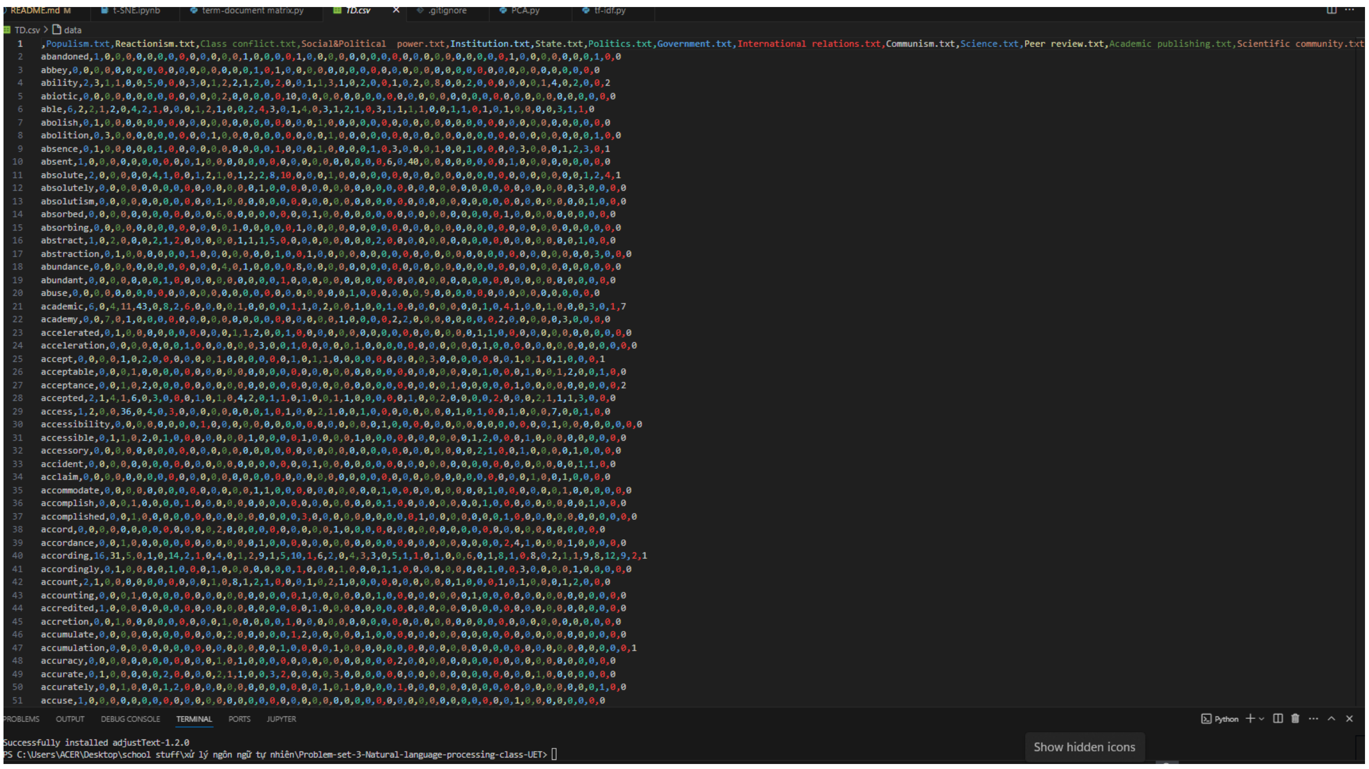
### Step 2

run term-document matrix.py to generate the term-document matrix:



term-document matrix.py > read_text_files

```python
import os
import pandas as pd
import numpy as np
from sklearn.feature_extraction.text import CountVectorizer
import nltk
from nltk.corpus import words
import re

#nltk.download('words')
english_words = set(words.words())

def read_text_files(directory):
    documents = []
    filenames = []

    for filename in os.listdir(directory):
        if filename.endswith('.txt'):
            filepath = os.path.join(directory, filename)
            with open(filepath, 'r', encoding='utf-8') as file:
                text = file.read()
                words = [word for word in re.findall(r'\b\w+\b', text.lower()) if word in english_words]
                words = ' '.join(words)
                documents.append(words)
                filenames.append(filename)

    filenames = sorted(filenames, key=lambda x: int(x.split(' - ')[0]))
    filenames = [entry.split(' - ')[1] for entry in filenames]
    return documents, filenames


def create_term_document_matrix(documents):
    vectorizer = CountVectorizer(strip_accents='unicode',min_df = 0.03,stop_words='english')
    tdm = vectorizer.fit_transform(documents).toarray()
    tdm = np.transpose(tdm)
    return tdm, vectorizer.get_feature_names_out()

directory = 'Data'
documents, filenames = read_text_files(directory)
tdm, terms = create_term_document_matrix(documents)
tdm_df = pd.DataFrame(tdm ,columns=filenames, index=terms)
print(tdm_df.info)
tdm_df.to_csv('TD.csv')
```

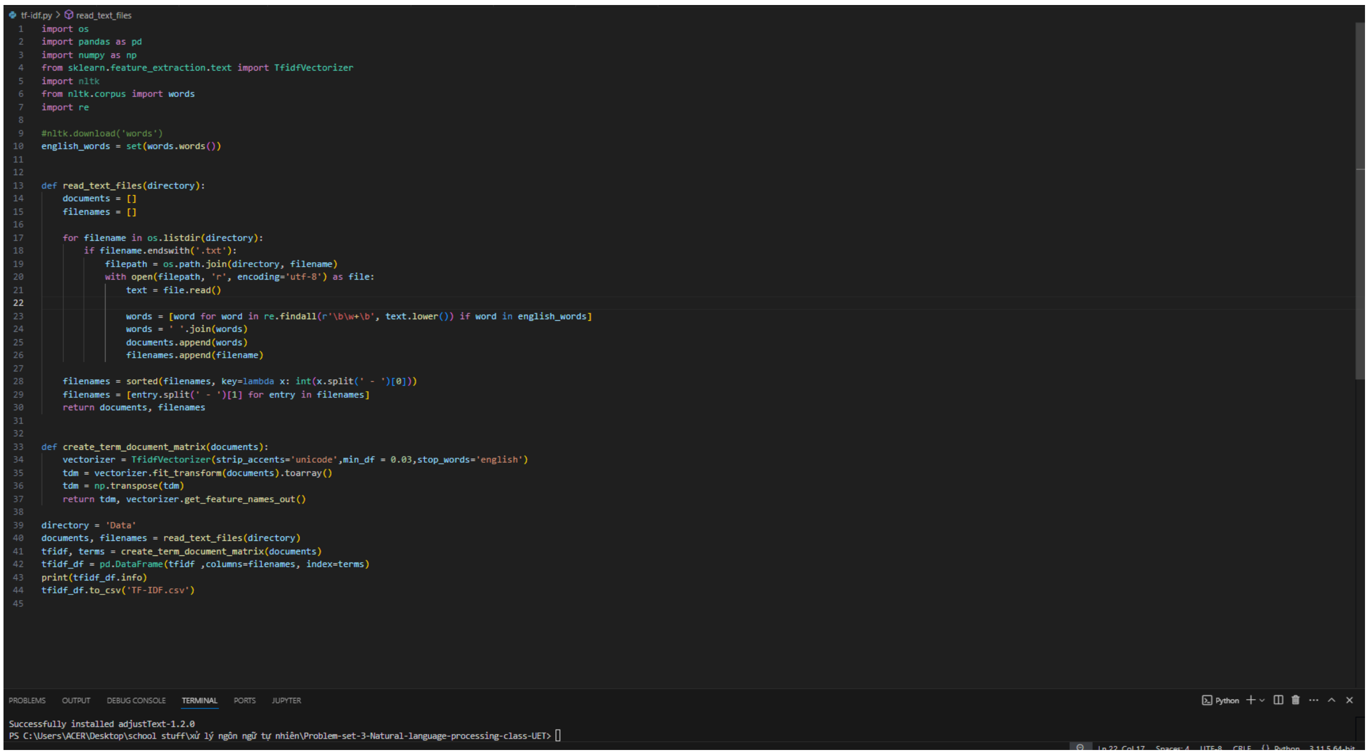


```
,Populism.txt,Reactionism.txt,Class conflict.txt,Social&Political  power.txt,Institution.txt,State.txt,Politics.txt,Government.txt,International relations.txt,Communism.txt,Science.txt,Peer review.txt,Academic publishing.txt,Scientific community.txt
abandoned,1,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,1,0,0
abbey,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
ability,2,3,1,1,0,0,5,0,0,0,3,0,1,2,2,1,2,0,2,0,0,1,1,3,1,0,2,0,0,1,0,2,0,8,0,0,2,0,0,0,0,0,0,1,4,0,2,0,0,2
abiotic,0,0,0,0,0,0,0,0,0,0,0,0,0,2,0,0,0,0,0,10,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
able,6,2,2,1,2,0,4,2,1,0,0,0,1,2,1,0,0,2,4,3,0,1,4,0,3,1,2,1,0,3,1,1,1,1,0,0,1,1,0,1,0,1,0,0,0,0,3,1,1,0
abolish,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
abolition,0,3,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0
absence,0,1,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,1,0,0,0,1,0,0,0,0,1,0,3,0,0,0,1,0,0,1,0,0,0,0,3,0,0,0,1,2,3,0,1
absent,1,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,6,0,40,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0
absolute,2,0,0,0,0,0,4,1,0,0,1,2,1,0,1,2,2,8,10,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,2,4,1
absolutely,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,3,0,0,0,0
absolutism,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0
absorbed,0,0,0,0,0,0,0,0,0,0,0,0,6,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0
absorbing,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
abstract,1,0,2,0,0,0,2,1,2,0,0,0,0,0,1,1,1,5,0,0,0,0,0,0,0,0,0,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0
abstraction,0,1,0,0,0,0,0,0,1,0,0,0,0,0,0,0,1,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,3,0,0,0
abundance,0,0,0,0,0,0,0,0,0,0,0,0,4,0,1,0,0,0,0,8,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
abundant,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
abuse,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,9,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
academic,6,0,4,11,43,0,8,2,6,0,0,0,0,1,0,0,0,0,1,1,0,2,0,0,1,0,0,1,0,0,0,0,0,0,0,0,1,0,4,1,0,0,1,0,0,0,3,0,1,7
academy,0,0,7,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,2,2,0,0,0,0,0,0,0,0,2,0,0,0,0,0,3,0,0,0,0,0
accelerated,0,1,0,0,0,0,0,0,0,0,0,0,1,1,2,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0
acceleration,0,0,0,0,0,0,0,1,0,0,0,0,0,0,3,0,0,1,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0
accept,0,0,0,0,1,0,2,0,0,0,0,0,0,1,0,0,0,0,0,0,1,0,1,1,0,0,0,0,0,0,0,0,0,3,0,0,0,0,0,0,0,1,0,1,0,1,0,0,0,1
acceptable,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,1,0,0,1,2,0,0,1,0,0
acceptance,0,0,1,0,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,2
accepted,2,1,4,1,6,0,3,0,0,0,1,0,1,0,4,2,0,1,1,0,1,0,0,1,1,0,0,0,0,0,1,0,0,2,0,0,0,0,2,0,0,0,2,1,1,1,3,0,0,0
access,1,2,0,0,36,0,4,0,3,0,0,0,0,0,0,0,0,1,0,1,0,0,2,1,0,0,1,0,0,0,0,0,0,0,0,1,0,1,0,0,1,0,0,0,7,0,0,1,0,0
accessibility,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0
accessible,0,1,1,0,2,0,1,0,0,0,0,0,0,0,1,0,0,0,0,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,1,2,0,0,0,1,0,0,0,0,0,0,0,0,0
accessory,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,2,1,0,0,1,0,0,0,0,1,0,0,0,0
accident,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0
acclaim,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,1,0,0,0,0
accommodate,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,1,0,0,0,0,0,0
accomplish,0,0,0,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,1,0,0,0
accomplished,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,3,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0
accord,0,0,0,0,0,0,0,0,0,0,0,0,0,2,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
accordance,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,2,4,1,0,0,0,1,0,0,0,0,0,0
according,16,31,5,0,1,0,14,2,1,0,4,0,1,2,9,1,5,10,1,6,2,0,4,3,3,0,5,1,1,0,1,0,0,6,0,1,8,1,0,8,0,2,1,1,9,8,12,9,2,1
accordingly,0,1,0,0,0,0,1,0,0,0,1,0,0,0,0,0,0,0,1,0,0,0,1,0,0,0,1,1,0,0,0,0,0,0,0,0,1,0,0,3,0,0,0,0,1,0,0,0,0,0
account,2,1,0,0,0,0,0,0,0,0,0,0,1,0,8,1,2,1,0,0,0,1,0,2,1,0,0,0,0,0,0,0,0,0,0,1,0,0,0,1,0,1,0,0,0,1,2,0,0,0
accounting,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0
accredited,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
accretion,0,0,1,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
accumulate,0,0,0,0,0,0,0,0,0,0,0,0,2,0,0,0,0,0,1,2,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
accumulation,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1
accuracy,0,0,0,0,0,0,0,0,0,0,0,0,1,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
accurate,0,1,0,0,0,0,0,2,0,0,0,0,2,1,1,0,0,3,2,0,0,0,0,3,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0
accurately,0,0,1,0,0,0,1,2,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0
accuse,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0
```

```
Successfully installed adjustText-1.2.0
PS C:\Users\ACER\Desktop\school stuff\xử lý ngôn ngữ tự nhiên\Problem-set-3-Natural-language-processing-class-UET> 
```

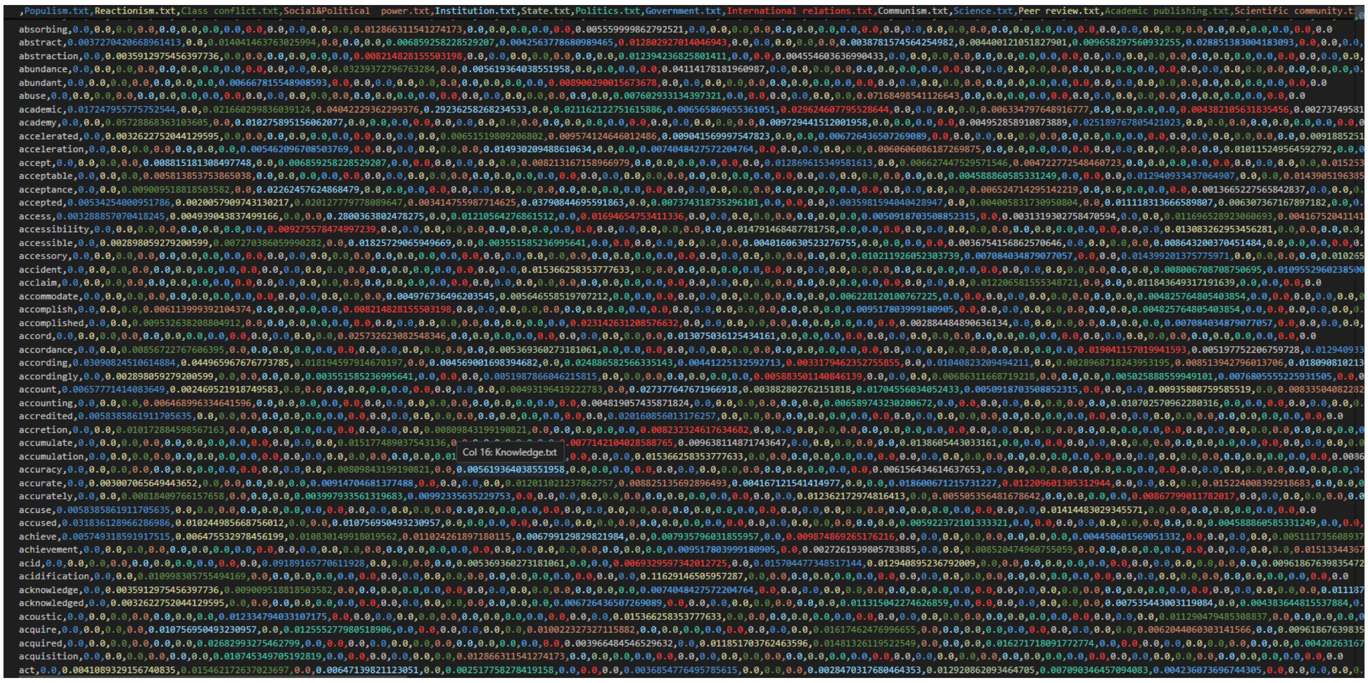
## Step 3

### run tf-idf.py to compute TF-IDF matrix:



```python
import os
import pandas as pd
import numpy as np
from sklearn.feature_extraction.text import TfidfVectorizer
import nltk
from nltk.corpus import words
import re

#nltk.download('words')
english_words = set(words.words())


def read_text_files(directory):
    documents = []
    filenames = []

    for filename in os.listdir(directory):
        if filename.endswith('.txt'):
            filepath = os.path.join(directory, filename)
            with open(filepath, 'r', encoding='utf-8') as file:
                text = file.read()

                words = [word for word in re.findall(r'\b\w+\b', text.lower()) if word in english_words]
                words = ' '.join(words)
                documents.append(words)
                filenames.append(filename)

    filenames = sorted(filenames, key=lambda x: int(x.split(' - ')[0]))
    filenames = [entry.split(' - ')[1] for entry in filenames]
    return documents, filenames


def create_term_document_matrix(documents):
    vectorizer = TfidfVectorizer(strip_accents='unicode',min_df = 0.03,stop_words='english')
    tdm = vectorizer.fit_transform(documents).toarray()
    tdm = np.transpose(tdm)
    return tdm, vectorizer.get_feature_names_out()

directory = 'Data'
documents, filenames = read_text_files(directory)
tfidf, terms = create_term_document_matrix(documents)
tfidf_df = pd.DataFrame(tfidf ,columns=filenames, index=terms)
print(tfidf_df.info)
tfidf_df.to_csv('TF-IDF.csv')
```

```
Successfully installed adjustText-1.2.0
PS C:\Users\ACER\Desktop\school stuff\xử lý ngôn ngữ tự nhiên\Problem-set-3-Natural-language-processing-class-UET> 
```

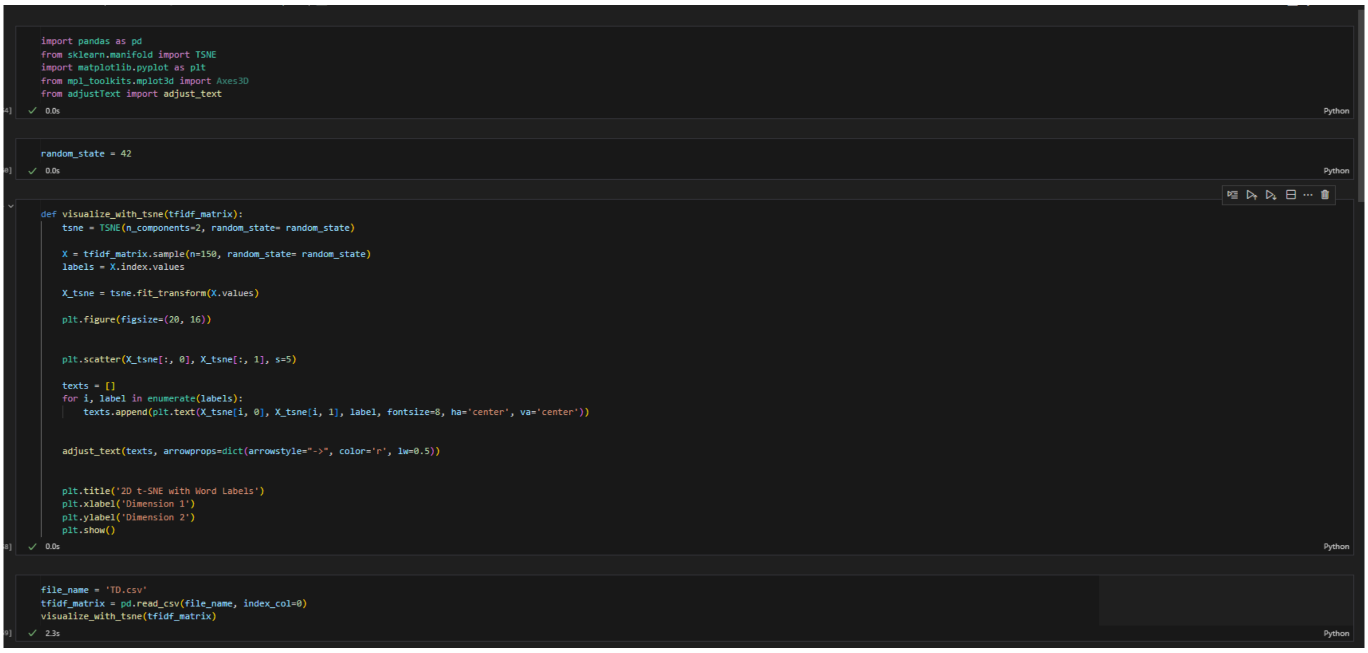


```
,Populism.txt,Reactionism.txt,Class conflict.txt,Social&Political  power.txt,Institution.txt,State.txt,Politics.txt,Government.txt,International relations.txt,Communism.txt,Science.txt,Peer review.txt,Academic publishing.txt,Scientific community.t
absorbing,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.012866311541274173,0.0,0.0,0.0,0.0,0.0,0.005559999862792521,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0
abstract,0.0037270420668961413,0.0,0.014041463763025994,0.0,0.0,0.0,0.006859258228529207,0.0042563778680989465,0.012802927014046943,0.0,0.0,0.0,0.0,0.0,0.0038781574564254982,0.004400121051827901,0.009658297560932255,0.028851383004183093,0.0,0.0,0.0,
abstraction,0.0,0.0035912975456397736,0.0,0.0,0.0,0.0,0.0,0.0,0.008214828155503198,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.012394236825801411,0.0,0.0,0.004554603636990433,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
abundance,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.03239372796763284,0.0,0.005619364038551958,0.0,0.0,0.0,0.0,0.04114178181960987,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
abundant,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.006667815548908593,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.008900290015673678,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0
abuse,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.007602933134397321,0.0,0.0,0.0,0.0,0.0,0.0,0.07168498541126643,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0
academic,0.017247955775752544,0.0,0.021660299836039124,0.04042229362299376,0.29236258268234533,0.0,0.021162122751615886,0.006565869655361051,0.029624607795528644,0.0,0.0,0.0,0.0,0.006334797648916777,0.0,0.0,0.0,0.0,0.004382105631835456,0.00273749581
academy,0.0,0.0,0.05728868363103605,0.0,0.010275895156062077,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.009729441512001958,0.0,0.0,0.0,0.0,0.004952858910873889,0.025189767805421023,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
accelerated,0.0,0.0032622752044129595,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.00651519809206802,0.009574124646012486,0.009041569997547823,0.0,0.0,0.006726436507269089,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0091885251
acceleration,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.005462096708503769,0.0,0.0,0.0,0.0,0.0,0.0,0.014930209488610634,0.0,0.0,0.0074048427572204764,0.0,0.0,0.0,0.0,0.0,0.006060606086187269875,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.010115249564592792,0.0,0
accept,0.0,0.0,0.0,0.0,0.008815181308497748,0.0,0.006859258228529207,0.0,0.0,0.0,0.0,0.0,0.0,0.008213167158966979,0.0,0.0,0.0,0.0,0.0,0.0,0.012869615349581613,0.0,0.006627447529571546,0.004722772548460723,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.015253
acceptable,0.0,0.0,0.0,0.005813853753865038,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.004588860585331249,0.0,0.0,0.0,0.012940933437064907,0.0,0.0,0.0143905196385
acceptance,0.0,0.0,0.009009518818503582,0.0,0.02262457624868479,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.006524714295142219,0.0,0.0,0.0,0.0,0.0,0.0013665227565842837,0.0,0.0,0.
accepted,0.00534254000951786,0.0020057909743130217,0.020127779778089647,0.003414755987714625,0.03790844695591863,0.0,0.007374318735296101,0.0,0.0,0.0,0.0035981594040428947,0.0,0.004005831730950804,0.0,0.011118313666589807,0.006307367167897182,0.0,0.
access,0.003288857070418245,0.004939043837499166,0.0,0.0,0.2800363802478275,0.0,0.01210564276861512,0.0,0.01694654753411336,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0050918703508852315,0.0,0.0031319302758470594,0.0,0.0,0.011696528923060693,0.00416752041141
accessibility,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.009275578474997239,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.014791468487781758,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.013083262953456281,0.0,0.0,0.0,0.0,
accessible,0.0,0.002898059279200599,0.007270386059990282,0.0,0.01825729065949669,0.0,0.003551585236995641,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0040160630523276755,0.0,0.0,0.0,0.0,0.0036754156862570646,0.0,0.0,0.0,0.0,0.008643200370451484,0.0,0.0,0.0,0.0,0.
accessory,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.010211926052303739,0.007084034879077057,0.0,0.0,0.014399201375775971,0.0,0.0,0.0,0.0,0.010265
accident,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.015366258353777633,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.008006708708750695,0.0109552960238500
acclaim,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.012206581555348721,0.0,0.0,0.011843649317191639,0.0,0.0,0.0,0.0
accommodate,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.004976736496203545,0.005646558519707212,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.006228120100767225,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.004825764805403854,0.0,0.0,0.0,0.0,0.0,0
accomplish,0.0,0.0,0.0,0.006113999392104374,0.0,0.0,0.0,0.0,0.008214828155503198,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.009517803999180905,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.004825764805403854,0.0,0.0,0.0,0.0,0.0,0.
accomplished,0.0,0.0,0.009532638208804912,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.023142631208576632,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.002884484890636134,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.007084034879077057,0.0,0.0,0.0,0.0,
accord,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.025732623082548346,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.013075036125434161,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0
accordance,0.0,0.0,0.008567227676066395,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.005369360273181061,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.019041157019941593,0.005197752206759728,0.012940933
according,0.03090824510614884,0.044965967676773785,0.018194597914670197,0.0,0.004569001698394682,0.0,0.024886582566335143,0.00441225132592713,0.0033179462352755855,0.0,0.01040823209494211,0.0,0.0028968718243953195,0.008513942796013706,0.018090810213
accordingly,0.0,0.002898059279200599,0.0,0.0,0.0,0.0,0.003551585236995641,0.0,0.0,0.0,0.0051987866046215815,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.005883501140846139,0.0,0.0,0.0,0.00686311668719218,0.0,0.0,0.0,0.005025888559949101,0.007680555225931505,0.0,0
account,0.00657771414083649,0.002469521918749583,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.00493196419222783,0.0,0.027377647671966918,0.003882802762151818,0.01704556034052433,0.0050918703508852315,0.0,0.0,0.0,0.009358087595855519,0.0,0.00833504082282
accounting,0.0,0.0,0.0,0.006468996334641596,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.004819057435871824,0.0,0.0,0.0,0.0,0.0,0.0,0.006589743230200672,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.010702570962280316,0.0,0.0,0.0,0.0,0.0,0.0,0.
accredited,0.0058385861911705635,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.020160856013176257,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0
accretion,0.0,0.0,0.010172884598567163,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.00809843199190821,0.0,0.0,0.0,0.0,0.0,0.008232324617634682,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0
accumulate,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.015177489037543136,0[illegible].0077142104028588765,0.009638114871743647,0.0,0.0,0.0,0.0,0.0,0.0138605443033161,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
accumulation,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.01[illegible],0.0,0.0,0.0,0.015366258353777633,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0086
accuracy,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.00809843199190821,0.0,0.005619364038551958,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.006156434614637653,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
accurate,0.0,0.003007065649443652,0.0,0.0,0.0,0.0,0.0,0.00914704681377488,0.0,0.0,0.0,0.0,0.012011021237862757,0.008825135692896493,0.004167121541414977,0.0,0.0,0.018600671215731227,0.012209601305312944,0.0,0.0,0.0,0.0,0.015224008392918683,0.0,0.0,0
accurately,0.0,0.0,0.00818409766157658,0.0,0.0,0.0,0.003997933561319683,0.00992335635229753,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.012362172974816413,0.0,0.005505356481678642,0.0,0.0,0.0,0.0,0.00867799011782017,0.0,0.0,0.0,0.0,0.0,0.0
accuse,0.0058385861911705635,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.01414483029345571,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0
accused,0.031836128966286986,0.010244985668756012,0.0,0.0,0.010756950493230957,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.005922372101333321,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.004588860585331249,0.0,0.0,
achieve,0.005749318591917515,0.006475532978456199,0.010830149918019562,0.011024261897180115,0.006799129829821984,0.0,0.007935796031855957,0.0,0.009874869265176216,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.004450601569051332,0.0,0.0,0.0,0.0,0.005111735608937
achievement,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.009517803999180905,0.0,0.0027261939805783885,0.0,0.0,0.008520474960755059,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.01513344367
acid,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.09189165770611928,0.0,0.0,0.0,0.0,0.005369360273181061,0.0,0.0,0.0069329597342012725,0.0,0.015704477348517144,0.012940895236792009,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.00961867639835472
acidification,0.0,0.0,0.010998305755494169,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.11629146505957287,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0
acknowledge,0.0,0.0035912975456397736,0.009009518818503582,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0074048427572204764,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.011137
acknowledged,0.0,0.0032622752044129595,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.006726436507269089,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.011315042274626859,0.0,0.0,0.0,0.0,0.0,0.0,0.007535443003119084,0.0,0.004383644815537884,0.
acoustic,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.012334794033107175,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.015366258353777633,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.011290479485308837,0.0,0.0,0.0,0.0,0
acquire,0.0,0.0,0.0,0.0,0.010756950493230957,0.0,0.012555277980518906,0.0,0.0,0.0,0.0,0.0,0.0,0.010022327327115882,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.01617462476996655,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0062044060303141566,0.0,0.00961867639835
acquired,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.02682993275462799,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.003966484546529632,0.0,0.011851703762463596,0.01481326119522549,0.0,0.0,0.0,0.016271718091772774,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.00420263167
acquisition,0.0,0.0,0.0,0.0,0.0,0.0,0.010745349705192819,0.0,0.0,0.0,0.0,0.0,0.012866311541274173,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0
act,0.0,0.0041089329156740835,0.015462172637023697,0.0,0.00647139821123051,0.0,0.002517758278419158,0.0,0.0,0.0,0.0036854776495785615,0.0,0.0,0.0,0.0028470317680464353,0.012920862093464705,0.007090346457094083,0.004236073696744305,0.0,0.0,0.0,0.0,0.
```

## Step 4

Visualize word vectors using t-SNE:

*(To get different results, modify the `random_state` variable in the t-SNE script.)*

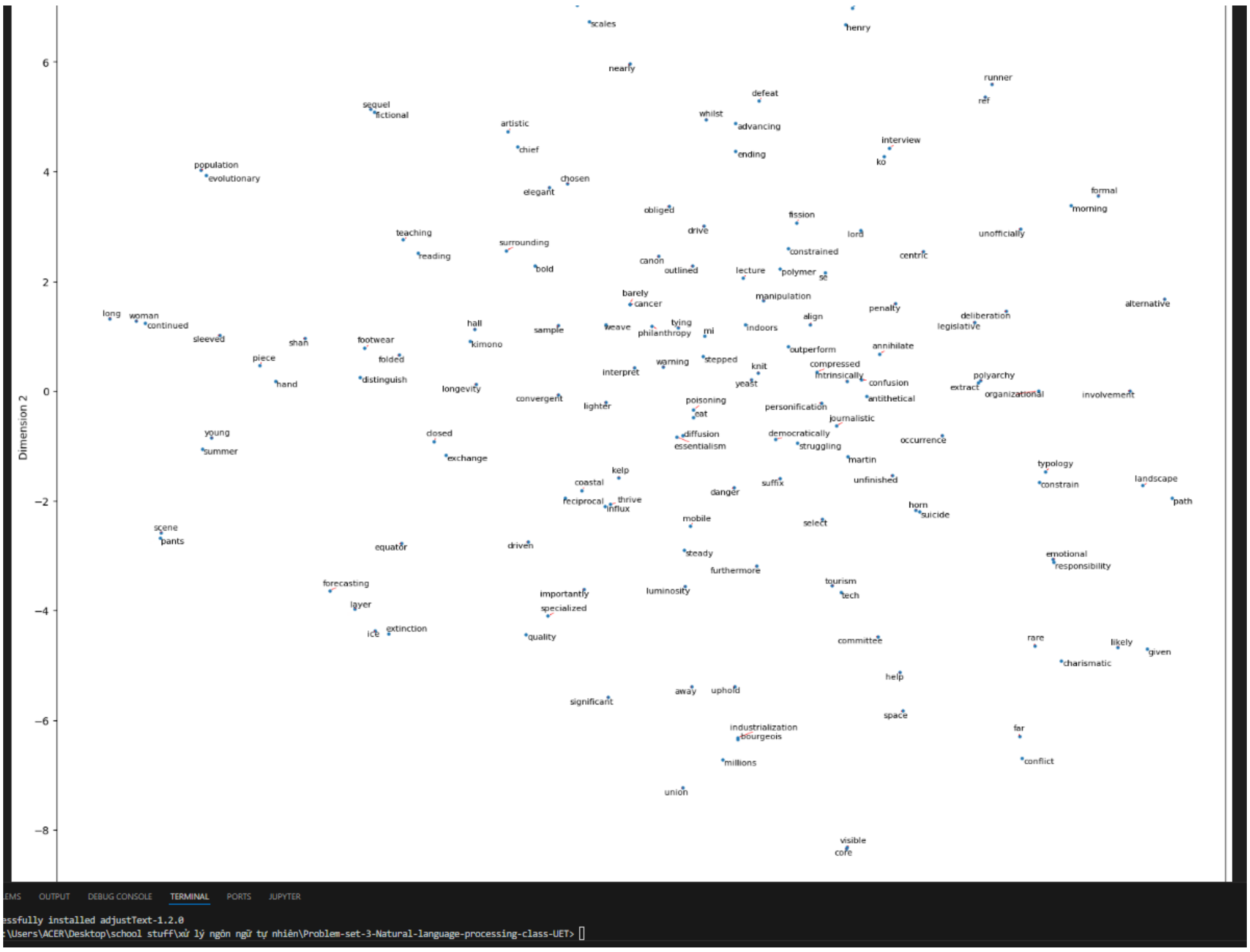


```python
import pandas as pd
from sklearn.manifold import TSNE
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D
from adjustText import adjust_text
```

```python
random_state = 42
```

```python
def visualize_with_tsne(tfidf_matrix):
    tsne = TSNE(n_components=2, random_state= random_state)

    X = tfidf_matrix.sample(n=150, random_state= random_state)
    labels = X.index.values

    X_tsne = tsne.fit_transform(X.values)

    plt.figure(figsize=(20, 16))


    plt.scatter(X_tsne[:, 0], X_tsne[:, 1], s=5)

    texts = []
    for i, label in enumerate(labels):
        texts.append(plt.text(X_tsne[i, 0], X_tsne[i, 1], label, fontsize=8, ha='center', va='center'))


    adjust_text(texts, arrowprops=dict(arrowstyle="->", color='r', lw=0.5))


    plt.title('2D t-SNE with Word Labels')
    plt.xlabel('Dimension 1')
    plt.ylabel('Dimension 2')
    plt.show()
```

```python
file_name = 'TD.csv'
tfidf_matrix = pd.read_csv(file_name, index_col=0)
visualize_with_tsne(tfidf_matrix)
```



```
essfully installed adjustText-1.2.0
:\Users\ACER\Desktop\school stuff\xử lý ngôn ngữ tự nhiên\Problem-set-3-Natural-language-processing-class-UET> 
```