Information Retrieval and Web Analytics

Semester 2 - 2022

Practical Sheet 02

Write python code to do the followings.

1). Lets assume you have the following corpus.

- Doc 1: breakthrough drug for schizophrenia
- Doc 2: new schizophrenia drug
- Doc 3: new approach for treatment of schizophrenia
- Doc 4: new hopes for schizophrenia patients

a). Write the code to build the inverted index for the above corpus.

b). Write suitable code to do the following retrieval tasks.

I. schizophrenia AND drug

II. for AND NOT(drug OR approach)

NLTK

NLTK is a platform which support to build Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries.

Installing NLTK

NLTK requires Python versions 3.5, 3.6, 3.7, 3.8, or 3.9

Mac/Unix

- Install NLTK: run pip install --user -U nltk
- Install Numpy (optional): run pip install --user -U numpy
- Test installation: run python then type import nltk

Windows

- Install Numpy (optional): https://www.scipy.org/scipylib/download.html
  pip install numpy
- Install NLTK: http://pypi.python.org/pypi/nltk
  Pip install nltk

Further information: https://www.nltk.org/install.html

a) Remove stop words in the given string using nltk library

```
quote = "Pythoners are very intelligent and work very pythonly and n
ow they are pythoning their way to success."
```

- Include ‘intelligent, ‘work’ as stopwords and print the new word list after removing stopwords

b) Use stemming text processing for the given sentence

2). Write the code to build the positional index for the given set of documents (positional folder)