CMPE 346- NATURAL LANGUAGE PROCESSING

3rd Week Exercise

Essential Terminal Commands

mkdir: abbreviation of make directory, usage mkdir SomeFolder
cd: stands for change directory, usage cd SomeFolder
ls: lists all contents of the current directory, usage; ls, ls *.py etc.
pwd: path of the current working directory
mv: move file or change name, usage myprogram.py SomeFolder

Pip - Package Manager for Python Environment

pip: is the primary command which has several options that are employed for our purposes.
e.g. pip install numpy
This command will install numpy library of python to our computers.

Virtual Environment

A virtual environment helps developers create an isolated working directory to prevent the problems that might occur because of the different version of libraries or python.

You firstly need to install the Virtual Environment

```
pip install virtualenv
```

After that, you can create a virtual environment as following:

```
Virtualenv –p /usr/bin/python3 my_virtualenv
```

or

```
virtualenv –p /path/to/mypython3.6 /path/to/myvirtualenv
```

And then, you need to activate

```
source virtualenv_name/bin/activate
```

After you have done working within this environment, you might prefer deactivating, using deactivate command.

NLTK(NATURAL LANGUAGE TOOLKIT)

NLTK is a Python package that provides a set of different algorithms that are used to analyze textual data to find patterns and statistics for natural language processing purposes.

We could create a Python programming application which uses NLTK package to make some text analytics for practical purposes.

Our program will tokenize the text bot word-by-word and sentence-by-sentence.

We could also learn that number of occurrences of the word in our text, and the most common word in this text.

We could also eliminate the stopwords in our text. Stopwords are considered as noises (unnecessarily repeated words that are not helpful to understand texts)

Example Program:

```
import nltk # to import the library
from nltk.probability import FreqDist
from nltk.corpus import stopwords


text="""Hello Mr. Smith, how are you doing today?
        The weather is great, and city is awesome.
        The sky is pinkish-blue. You shouldn't eat cardboard"""

tokenized = nltk.word_tokenize(text)
tokenized_sentence = nltk.sent_tokenize(text)
distribution = FreqDist(tokenized)
stop_words=set(stopwords.words("english"))
filtered_sentences=[]
for w in tokenized_sentence:
  if w not in stop_words:
    filtered_sentences.append(w)

print (tokenized)
print(tokenized_sentence)
print(distribution)
print(distribution.most_common(2))
print(stop_words)
print("Tokenized Sentence:",tokenized_sentence)
print("Filterd Sentence:",filtered_sentences)
```