CMPE 409 Machine Translation

Worksheet(Week-04)

1 Download NLTK

Download NLTK package with following instructions

```
>>> import nltk
>>> nltk.download()
```

2 Test corpus Path

```
print("Testing custom corpora path")

import os, os.path
path=os.path.expanduser("~/nltk_data") # it may be differenent on Windows
print("Do we have this path?")
print(os.path.exists(path))
print("The path is: ",path)

import nltk.data
print(path in nltk.data.path)

data= nltk.data.load("bilgi/murat/a.txt", format = "raw")
print("Content of the file are: ")
print(data)
# Try to understand outputs

## Creae two diffent corpora and print out their paths
```

3 Create WordList Corpus

Create corpus from text files

```
print("Testing WordList corpus")
from nltk.corpus.reader import WordListCorpusReader
reader= WordListCorpusReader("/home/murat/nltk_data/bilgi/murat",
        ["list","a.txt","y.txt"])

print(reader.words())
print("file names are: ",reader.fileids())

print("out put with raw....")
print(reader.raw())

### This code similar
print("Test with line_tokenize")

from nltk.tokenize import line_tokenize
print(line_tokenize(reader.raw()))

#### Test Enlish Word corpus##############

from nltk.corpus import words
print(words.fileids())
print(len(words.words('en-basic')))
print(len(words.words('en')))


# remove stop words
# remove repated words
# Create a wordlist corpus from 5 diffentf files such as: books, cars,
  animals,countries, sports etc.
```

4 Create a Corpus from Webs

```
 print ("collect from web files")

from urllib import request
url="https://www.bbc.com/news/uk-60708450"
#url="https://www.dunyabulteni.net/irak/irakta-secimler-surpriz-koalisyonlar-dogurd
response=request.urlopen(url)
raw=response.read().decode("utf8")
print(len(raw))

## See the outputs...


from bs4 import BeautifulSoup
text= BeautifulSoup(raw,"html.parser").get_text()
print(len(text))
print(text)

print("tokenization....")

print(line_tokenize(text))
print("words")

from nltk.tokenize import word_tokenize
token=word_tokenize(text)

print(token)
for i in token:
print(i)

## Analize outputs
```

5 Create Corpus from Text Files

```
print("Reading from text files")


print("Text reader")
from nltk.corpus.reader import PlaintextCorpusReader
corpus = PlaintextCorpusReader("/home/murat/nltk_data/bilgi/murat", '.*')
print(corpus.words())
print(corpus.fileids())
print(len(corpus.words()))

for i in  corpus.words():
print(i)
```

6 IBM Module-1

```
print ("test IBM model 1")
from nltk.translate import AlignedSent, Alignment, IBMModel1
print (" packages IBM model 1 imported")

bitext= []

bitext.append(AlignedSent(['klein','ist','das','haus'],
                          ['the','hause','is','small']))
bitext.append(AlignedSent(['das', 'haus', 'ist', 'ja', 'groß'],
                         ['the', 'house', 'is', 'big']))
bitext.append(AlignedSent(['das', 'buch', 'ist', 'ja', 'klein'],
                          ['the', 'book', 'is', 'small']))
bitext.append(AlignedSent(['das', 'haus'], ['the', 'house']))
bitext.append(AlignedSent(['das', 'buch'], ['the', 'book']))
bitext.append(AlignedSent(['ein', 'buch'], ['a', 'book']))

print(bitext)

myIBM = IBMModel1(bitext,5)

print("translate")
print(myIBM.translation_table['buch']['book'])
print(myIBM.translation_table['das']['the'])
```

```
print("test -one-by-one")

test_sentence= bitext[0]

print(test_sentence.words)
print(test_sentence.mots)

print(test_sentence.alignment)

print(" check the 3rd sentence")

test_sentence= bitext[3]
print(test_sentence.words)
print(test_sentence.mots)
print(test_sentence.alignment)

print(" check the 2nd sentence")

test_sentence= bitext[2]
print(test_sentence.words)
print(test_sentence.mots)
print(test_sentence.alignment)
```

7 Uploading

Show your work to your instructor and upload to learn

8 Resource

This worksheet is prepared from the following books:

- Jacob Perkins, **Python 3 Text Processing with NLTK 3 Cookbook, Packt Publishing**,ISBN: 9781782167853
- Steven Bird, Ewan Klein & Edward Loper, **Natural Language Processing with Python**, O'Reily, June, 2009