CMPE 409 Machine Translation

Worksheet(Week-02)

1 Tokenizing Text into Sentences

```
para = "Hello World. It's good to see you. Thanks for buying this book."
from nltk.tokenize import sent_tokenize
print(sent_tokenize(para))

# check how many Sentences do you have

import nltk.data
tokenizer = nltk.data.load('tokenizers/punkt/PY3/english.pickle')
print(tokenizer.tokenize(para))

# check the result again

spanish_tokenizer = nltk.data.load('tokenizers/punkt/PY3/spanish.pickle')
print(spanish_tokenizer.tokenize('Hola amigo. Estoy bien.'))

#check the result again
```

2 Tokenizing Sentences into Words

```
from nltk.tokenize import word_tokenize
print( word_tokenize('Hello World.'))

#check the output

from nltk.tokenize import TreebankWordTokenizer
tokenizer = TreebankWordTokenizer()
print(tokenizer.tokenize('Hello World.'))

#check the output

print(word_tokenize("can't"))
#check the output
```

```
from nltk.tokenize import WordPunctTokenizer
tokenizer = WordPunctTokenizer()
print(tokenizer.tokenize("Can't is a contraction."))
#check the output
```

3 Training a Sentence Tokenizer

```
from nltk.tokenize import PunktSentenceTokenizer
from nltk.corpus import webtext
text = webtext.raw('overheard.txt')
sent_tokenizer = PunktSentenceTokenizer(text)
sents1 = sent_tokenizer.tokenize(text)
print(sents1[0])

#check the output


from nltk.tokenize import sent_tokenize
sents2 = sent_tokenize(text)
print(sents2[0])

#check the output
compare outputs of "sents2[0]" and "sents1[0]"

# now compare outputs of "sents1[678]" and "sents2[678]"

print(sents1[678])
print(sents2[678])

# explian differences
```

4 Filtering Stopwords in a Tokenized Sentence

```
from nltk.corpus import stopwords
english_stops = set(stopwords.words('english'))
words = ["Can't", 'is', 'a', 'contraction']
print([word for word in words if word not in english_stops])

#undestand the python script "word for word in words if word not in english_stops"
#check the output

# check "stopword" library with the following method
print(stopwords.fileids())

#check the output

#list the stop words of the "Dutch"
>>> stopwords.words('dutch')

#check the output


#list the stop words of the Turkish
```

5 Filtering Stopwords in a Turkish Text

- Define a Turkish paragraph
- Tokenize it into Sentences
- Tonenize Sentence into words
- Remove all stopwords from the paragraph and print out the result
- Count how many words do you have without stopwords.

6 Uploading

Show your work to your instructor and upload to learn

7 Resource

This worksheet is prepared from the following book

Jacob Perkins, **Python 3 Text Processing with NLTK 3 Cookbook, Packt Publishing**,ISBN: 9781782167853