NLTK - Natural Language ToolKit

In [1]:
```
import nltk
```

In [2]:
```
nltk.download('punkt')
nltk.download('words')
nltk.download('stopwords')
nltk.download('averaged_perceptron_tagger')
nltk.download('reuters')
nltk.download('names')
nltk.download('movie_reviews')
nltk.download('tagsets')
```

```
[nltk_data] Downloading package punkt to /Users/skalathur/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package words to /Users/skalathur/nltk_data...
[nltk_data]   Package words is already up-to-date!
[nltk_data] Downloading package stopwords to
[nltk_data]     /Users/skalathur/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package averaged_perceptron_tagger to
[nltk_data]     /Users/skalathur/nltk_data...
[nltk_data]   Package averaged_perceptron_tagger is already up-to-
[nltk_data]       date!
[nltk_data] Downloading package reuters to
[nltk_data]     /Users/skalathur/nltk_data...
[nltk_data]   Package reuters is already up-to-date!
[nltk_data] Downloading package names to /Users/skalathur/nltk_data...
[nltk_data]   Package names is already up-to-date!
[nltk_data] Downloading package movie_reviews to
[nltk_data]     /Users/skalathur/nltk_data...
[nltk_data]   Unzipping corpora/movie_reviews.zip.
[nltk_data] Downloading package tagsets to
[nltk_data]     /Users/skalathur/nltk_data...
[nltk_data]   Package tagsets is already up-to-date!
```

Out[2]: True

In [3]:
```
sample_text = '''I am a student from the University of Alabama. I
was born in Ontario, Canada and I am a huge fan of the United States.
I am going to get a degree in Philosophy to improve my chances of
becoming a Philosophy professor. I have been working towards this goal
for 4 years. I am currently enrolled in a PhD program. It is very difficult,
but I am confident that it will be a good decision'''

print(sample_text)
```

```
I am a student from the University of Alabama. I
was born in Ontario, Canada and I am a huge fan of the United States.
I am going to get a degree in Philosophy to improve my chances of
becoming a Philosophy professor. I have been working towards this goal
for 4 years. I am currently enrolled in a PhD program. It is very difficult,
but I am confident that it will be a good decision
```

In [ ]:

In [4]:
```
sample_word_tokens = nltk.word_tokenize(sample_text)
print(sample_word_tokens)
```

```
['I', 'am', 'a', 'student', 'from', 'the', 'University', 'of', 'Alabama', '.', 'I', 'was', 'born', 'in', 'Ont
ario', ',', 'Canada', 'and', 'I', 'am', 'a', 'huge', 'fan', 'of', 'the', 'United', 'States', '.', 'I', 'am',
'going', 'to', 'get', 'a', 'degree', 'in', 'Philosophy', 'to', 'improve', 'my', 'chances', 'of', 'becoming',
'a', 'Philosophy', 'professor', '.', 'I', 'have', 'been', 'working', 'towards', 'this', 'goal', 'for', '4', '
years', '.', 'I', 'am', 'currently', 'enrolled', 'in', 'a', 'PhD', 'program', '.', 'It', 'is', 'very', 'diffi
cult', ',', 'but', 'I', 'am', 'confident', 'that', 'it', 'will', 'be', 'a', 'good', 'decision']
```

In [5]:
```
tagged = nltk.pos_tag(sample_word_tokens)
tagged[0:10]
```

Out[5]:
```
[('I', 'PRP'),
 ('am', 'VBP'),
 ('a', 'DT'),
 ('student', 'NN'),
 ('from', 'IN'),
 ('the', 'DT'),
 ('University', 'NNP'),
 ('of', 'IN'),
 ('Alabama', 'NNP'),
 ('.', '.')]
```

In [ ]:

In [6]:
```
sample_sent_tokens = nltk.sent_tokenize(sample_text)
print(sample_sent_tokens)
```

```
['I am a student from the University of Alabama.', 'I \nwas born in Ontario, Canada and I am a huge fan of th
e United States.', 'I am going to get a degree in Philosophy to improve my chances of \nbecoming a Philosophy
professor.', 'I have been working towards this goal\nfor 4 years.', 'I am currently enrolled in a PhD progra
m.', 'It is very difficult, \nbut I am confident that it will be a good decision']
```

In [7]:
```
from nltk.corpus import stopwords
```

In [8]:
```
stop_words = stopwords.words('english')
print(stop_words)
```

```
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've", "you'll", "you'd", '
your', 'yours', 'yourself', 'yourselves', 'he', 'him', 'his', 'himself', 'she', "she's", 'her', 'hers', 'hers
elf', 'it', "it's", 'its', 'itself', 'they', 'them', 'their', 'theirs', 'themselves', 'what', 'which', 'who',
'whom', 'this', 'that', "that'll", 'these', 'those', 'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being',
'have', 'has', 'had', 'having', 'do', 'does', 'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or', 'be
cause', 'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with', 'about', 'against', 'between', 'into', 'thro
ugh', 'during', 'before', 'after', 'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'o
ver', 'under', 'again', 'further', 'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'an
y', 'both', 'each', 'few', 'more', 'most', 'other', 'some', 'such', 'no', 'nor', 'not', 'only', 'own', 'same
', 'so', 'than', 'too', 'very', 's', 't', 'can', 'will', 'just', 'don', "don't", 'should', "should've", 'now
', 'd', 'll', 'm', 'o', 're', 've', 'y', 'ain', 'aren', "aren't", 'couldn', "couldn't", 'didn', "didn't", 'do
esn', "doesn't", 'hadn', "hadn't", 'hasn', "hasn't", 'haven', "haven't", 'isn', "isn't", 'ma', 'mightn', "mig
htn't", 'mustn', "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn', "shouldn't", 'wasn', "wasn't", '
weren', "weren't", 'won', "won't", 'wouldn', "wouldn't"]
```

In [9]:
```
word_tokens = [word for word in sample_word_tokens if word.lower() not in stop_words]
print(word_tokens)
```

```
['student', 'University', 'Alabama', '.', 'born', 'Ontario', ',', 'Canada', 'huge', 'fan', 'United', 'States
', '.', 'going', 'get', 'degree', 'Philosophy', 'improve', 'chances', 'becoming', 'Philosophy', 'professor',
'.', 'working', 'towards', 'goal', '4', 'years', '.', 'currently', 'enrolled', 'PhD', 'program', '.', 'diffic
ult', ',', 'confident', 'good', 'decision']
```

In [ ]:
```

```

In [10]:
```
nltk.help.brown_tagset('NN')
```

```
NN: noun, singular, common
    failure burden court fire appointment awarding compensation Mayor
    interim committee fact effect airport management surveillance jail
    doctor intern extern night weekend duty legislation Tax Office ...
```

In [11]:
```
nltk.help.brown_tagset('RB')
```

```
RB: adverb
    only often generally also nevertheless upon together back newly no
    likely meanwhile near then heavily there apparently yet outright fully
    aside consistently specifically formally ever just ...
```

In [12]:
```
nltk.help.brown_tagset('IN')
```

```
IN: preposition
    of in for by considering to on among at through with under into
    regarding than since despite according per before toward against as
    after during including between without except upon out over ...
```

In [ ]:
```

```