**Basic Language Processing with NLTK**

In this post, we explore some basic text processing using the Natural Language Toolkit (NLTK).

We will be grabbing the most popular nouns from a list of text documents.

We will need to start by downloading a couple of NLTK packages for language processing.

punkt is used for tokenising sentences and averaged\_perceptron\_tagger is used for tagging words with their parts of speech (POS). We also need to set the add this directory to the NLTK data path.

In [1]:

**import** **os**

**import** **nltk**

# Create NLTK data directory

NLTK\_DATA\_DIR = './nltk\_data'

**if** **not** os.path.exists(NLTK\_DATA\_DIR):

os.makedirs(NLTK\_DATA\_DIR)

nltk.data.path.append(NLTK\_DATA\_DIR)

# Download packages and store in directory above

nltk.download('punkt', download\_dir=NLTK\_DATA\_DIR)

nltk.download('averaged\_perceptron\_tagger', download\_dir=NLTK\_DATA\_DIR)

Out[1]:

True

The documents to be analysed here are in the docs folder, we see that there are 6 text documents.

In [2]:

DOCS\_DIR = './docs'

files = os.listdir(DOCS\_DIR)

files

Out[2]:

['doc1.txt', 'doc2.txt', 'doc3.txt', 'doc4.txt', 'doc5.txt', 'doc6.txt']

For now we will analyse the first document doc1.txt

In [3]:

**import** **io**

path = os.path.join(DOCS\_DIR, 'doc1.txt')

**with** io.open(path, encoding='utf-8') **as** f:

text\_from\_file = ' '.join(f.read().splitlines())

text\_from\_file now has the whole corpus of text from doc1.txt.

If we tokenise this block of text by sentences, we are left with a list of sentences.

In [4]:

# 5th sentence

sentences = nltk.sent\_tokenize(text\_from\_file)

fifth\_sentence = sentences[4]

**print**(fifth\_sentence)

In the face of despair, you believe there can be hope.

We can also tokenise the words within the sentence

In [5]:

# 3rd word from 5th sentence

words\_in\_fifth\_sentence = nltk.word\_tokenize(fifth\_sentence)

third\_word\_fifth\_sentence = words\_in\_fifth\_sentence[2]

third\_word\_fifth\_sentence

Out[5]:

u'face'

Now we will tag the parts of speech of this sentence

In [6]:

# Part of speech tagging

nltk.pos\_tag(words\_in\_fifth\_sentence)

Out[6]:

[(u'In', 'IN'),

(u'the', 'DT'),

(u'face', 'NN'),

(u'of', 'IN'),

(u'despair', 'NN'),

(u',', ','),

(u'you', 'PRP'),

(u'believe', 'VBP'),

(u'there', 'EX'),

(u'can', 'MD'),

(u'be', 'VB'),

(u'hope', 'VBN'),

(u'.', '.')]

We see that the words are now tuples of the given word and a tag for the part of speech.

In the list above, the following abbreviations can be observed:

* IN – Preposition
* DT – Determiner
* NN – Noun
* PRP – Personal pronoun
* EX – Existential there
* MD – Modal
* VB – Verb base form
* VBN – Verb past participle

Our aim is to extract the most popular nouns from all the sentences across all the documents.

In [7]:

# Get all sentences from all documents

all\_text = ''

**for** file\_name **in** files:

path = os.path.join(DOCS\_DIR, file\_name)

**with** io.open(path, encoding='utf-8') **as** f:

text\_from\_file = ' '.join(f.read().splitlines())

all\_text += text\_from\_file

sentences = nltk.sent\_tokenize(all\_text)

# tag speech

words = []

**for** sentence **in** sentences:

words.extend(nltk.word\_tokenize(sentence))

tagged\_words = nltk.pos\_tag(words)

# Count nouns

**from** **collections** **import** defaultdict

freqs = defaultdict(int)

nouns = [word[0].lower() **for** word **in** tagged\_words **if** word[1] == 'NN']

**for** noun **in** nouns:

**if** noun **in** freqs:

freqs[noun] += 1

**else**:

freqs[noun] = 1

# Get the 25 most common words

**import** **operator**

sorted\_by\_most\_common = sorted(freqs.items(), key=operator.itemgetter(1), reverse=True)

**for** word, count **in** sorted\_by\_most\_common[:25]:

**print** "{} was found {} times.".format(word.capitalize(), count)

Country was found 60 times.

Time was found 58 times.

Government was found 45 times.

War was found 40 times.

Promise was found 33 times.

Today was found 29 times.

World was found 29 times.

Care was found 27 times.

Health was found 23 times.

Work was found 23 times.

Way was found 21 times.

Nation was found 21 times.

Corruption was found 20 times.

Job was found 19 times.

Party was found 19 times.

Year was found 18 times.

Security was found 18 times.

Economy was found 17 times.

Part was found 17 times.

Progress was found 17 times.

Change was found 15 times.

Generation was found 15 times.

Man was found 14 times.

Threat was found 14 times.

Life was found 14 times.