Lecture 5

Ashwini Vaidya

15/4

NLTK

The natural language processing toolkit can be installed by typing

```
conda install nltk
```

Once this is installed, also download the corpora that are found in the NLTK book by typing the following:

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

Choose the option "Everything used in the NLTK book" and wait until the downloader finishes getting the texts.

The NLTK library is a module, which is a namespace containing a number of useful functions. When we import a module, python will internally find the file where this module resides, compile it and run it to access all the objects defined in it. That's why there will be a tiny pause when you type the following command to get nltk in the interactive mode:

```
import nltk
```

The NLTK book itself is a cookbook style collection of recipes to carry out basic tasks in natural language processing. Generally, NLP tasks follow a basic pipeline, tokenization, lemmatization, part of speech tagging, shallow parsing. Usually, this is followed by syntactic parsing and other tasks-semantic role labelling. We can take a look at how NLTK helps with this pipeline.

```
>>> import nltk
>>>from nltk import book
```

```
>>> f=open('Hin107.txt')
>>> raw= f.read()
>>> raw_u= raw.decode('utf-8')
>>> tokens=nltk.word_tokenize(raw_u)
>>> for word in tokens[0:20]:
...     print word
```

We then make use of the NLTK's built in functions such as FreqDist to find out the frequency distribution of certain words. ¿¿¿ w= ".decode('utf-8')

```
>>> fdist=FreqDist(tokens)
>>> fdist
FreqDist({u'\u0915\u0947': 72, u'\u092e\u0947\u0902': 54,  ...})
>>> w= ''.decode('utf-8')
>>> fdist[w]
26
```

FreqDist is an object from the NLTK module, which stores

classes and functions. We will often refer to them using their class names directly or via the object.attribute