ICA 4 POS Tagging

(September 30th)

In this activity, we will work on POS Tagging

(Same as ICA 1) From NLTK import gutenberg corpus and open up the book *Alice in Wonderland* by Lewis Carol.

No need for preprocessing the text, in fact we want it to be un-preprocessed.

Here is a resource that will help:
http://www.ling.helsinki.fi/kit/2009s/clt231/NLTK/book/ch05-CategorizingAndTaggingWords.html

Step 2: Run the sentences in the first three paragraphs of *Alice in Wonderland* through POS tagger using the below command.

```
nltk.pos_tag(text)
```

Below is a list of tags and what they represent.

Tag	Meaning	Examples
ADJ	adjective	*new, good, high, special, big, local*
ADV	adverb	*really, already, still, early, now*
CNJ	conjunction	*and, or, but, if, while, although*
DET	determiner	*the, a, some, most, every, no*
EX	existential	*there, there's*
FW	foreign word	*dolce, ersatz, esprit, quo, maitre*
MOD	modal verb	*will, can, would, may, must, should*
N	noun	*year, home, costs, time, education*
NP	proper noun	*Alison, Africa, April, Washington*
NUM	number	*twenty-four, fourth, 1991, 14:24*
PRO	pronoun	*he, their, her, its, my, I, us*
P	preposition	*on, of, at, with, by, into, under*
TO	the word *to*	*to*
UH	interjection	*ah, bang, ha, whee, hmpf, oops*
V	verb	*is, has, get, do, make, see, run*
VD	past tense	*said, took, told, made, asked*
VG	present participle	*making, going, playing, working*
VN	past participle	*given, taken, begun, sung*
WH	*wh* determiner	*who, which, when, what, where, how*

Step 3: Using the output from the previous step, list tags in order of decreasing frequency. Which are the top 5 most frequent tags?

Step 4: The POS tagging task can be accomplished using regular expressions. For example, we can assume that any word ending in “-ed” is a past participle of a verb.
Using a regular expression based POS tagger, with the following three regexes, determine at least three of the words from the output of Step 2 above which would be incorrectly tagged.

```
    (r'.*ing$', 'VBG'),              # gerunds
...  (r'.*ed$', 'VBD'),               # simple past
...  (r'.*s$', 'NNS'),                # plural nouns
```