CSCI 8450 Chapter 3b: Assignment

Mohan Sai Ambati

1. wh words in text2 : 1869
   wh words in text7 : 631

2. The readability scores (ARI) for the genres in brown corpus is as follows:

Section	ARI
adventure	4.0841684990890705
belles_lettres	10.987652885621749
editorial	9.471025332953673
fiction	4.9104735321302115
government	12.08430349501021
hobbies	8.922356393630267
humor	7.887805248319808
learned	11.926007043317348
lore	10.254756197101155
mystery	3.8335518942055167
news	10.176684595052684
religion	10.203109907301261
reviews	10.769699888473433
romance	4.34922419804213
science_fiction	4.978058336905399

   This shows that genres adventure, fiction, mystery, romance, science fiction are easily readable whereas genres like belles letters, government, learned, lore, news, religion, reviews are difficult to read.

3. The differences between Porter Stemmer and Lancaster Stemmer is as follows:
   words in lancaster not in porter : {'twic', 'wer', 'reply', 'his', 'oh', 'going', 'in', 'cam', 'lucky', 'on', 'neighb', 'bef', 'farm', 'ar'}
   words in porter not in lancaster : {'Oh', 'were', 'twice', 'came', 'repli', 'befor', 'one', 'lucki', 'go', 'neighbor', 'farmer', 'In', 'are', 'hi'}
   It seems like Lancaster has mapped all the words to proper stems than porter. Porter missed words like 'twice','came'. But porter performed very well in finding stem of words like 'farmer' to 'farm'. But, it is difficult to find the exact difference with this small text.

4. I have used word_tokenize() to count the average word length and for splitting text into sentences I used sent_tokenize() which inbuilt uses an punkt sentence tokenizer instance which is trained with *"nltk_data/tokenizers/punkt/English.pickle".*
   I have added a flag to print the sentences after tokenizing.
   I have calculated readability based on these two parameters. The readability is given below,
   Reading difficulty of rural.txt : 12.61676279331764
   Reading difficulty of science.txt : 12.773526577547678
   It is clear that both have almost equal readability scores but rural.txt is little easily readable when compared to science.txt.