1. wh words in text2 : 1869

wh words in text7 : 631

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

A screenshot of a cell phone

Description generated with very high confidence

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.

1. 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.

1. 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.