Problem Set 4

Q1. Description of Class of String

1. [a-zA-Z]+

- Matches one or more of of a range of characters = [a-zA-Z]+

2. [A-Z][a-z]\*

- Matches at least one range of characters [A-Z] and up to two range of characters [A-Z][a-z] but [a-z]\* can be zero or more of a range of characters.

3. p[aeiou]{,2}t

- Always has ‘p’ as first character followed by a set of characters [aeiou] with no more than two repeats {,2} and ends with ‘t’ as last character

4. \d+(\.\d+)?

- one or more digits followed by a full stop and one or more digits enclosed by parentheses indicating the scope of the ‘?’ operator indicating there is zero or one of what’s encoled within the parentheses

5. ([^aeiou][aeiou][^aeiou])\*

- Matches some pattern *aeiou* at the start of a string, then matches a set of characters *aeiou* and finally matches some pattern *aeiou* at the start of a string again. Zero or more of what’s inside the parenthesis.

6. \w+|[^\w\s]+

- one or more alphanumeric character including the underscore OR one or more of the set of characters containing an alphanumeric character including the underscore followed by whitespace

Q2. Stemmer Differences

The Porter stemmer included the capitalized letters whereas the Lancaster stemmer normalizes the words and makes it all lower case. It’s clear that Lancaster stemmer does a better job at normalizing the word and it doesn’t strip off the affixes correctly. The Porter stemmer incorrectly removes too many characters sometimes and fails to lemmatize some of the words.

Q3. Brown Readability

After using the program on the different genres of the brown corpus it is clear that some of the genres which are expected to be harder to understand have a higher ARI. For example, government has the highest ARI probably because it requires a lot of terms which are very specific to its field. Then when we look at lore and learned they have roughly the same ARI probably which is not expected. I believe there is something wrong with my code because the ARI numbers outputted from the program are out of the range (0-16). Also, a lot of the numbers are very similar.