NLP Report A-1

Kaustav Vats | 2016048

Assumptions

1. Numbers won't be counted as a words and it's ok to remove them.
2. Kaustav.Vats both are different words. This split can be performed by word tokenizer.
3. For word count i can remove all punctuation, numbers etc.
4. Allowed to update file name in code.

Pre-Processing

1. Removing extra spaces from string using strip.
2. *Removing numeric points like "1. Bla bla" -> "Bla Bla". A number followed by full stop was counted as a sentence. So i removed this pattern using regex.*
3. *Removing symbols from the text except **['.', '!', '?', '@']** ('@' to avoid splitting into two words), because these are punctuation sentences.*
4. Removing dot from words like email.id.net and converting to [email, id, net] as separate words.
5. Converted all sentences to lowercase.

Steps

1. Calculating header by checking the first occurence of "\n\n".
2. Did sentence and word pre processing for calculating number of sentences and words.
3. Created regex for finding email address.
4. For checking the first and last words I used word tokenization and checked first and last word of the string.
5. To count occurence of words i word tokenized whole document and checked if each word is equal to specified word or not.