Gaurav Bishnoi

CSCI 5832: Natural Language Processing

Programming Assignment 1

Date: Jan 28, 2016

Word:
Regular Expression: **/␣|\.\n|!\n|\?\n|\"\n/**

To search word, I look for space between words and for special case i.e. last word of paragraph I search for period, exclamation mark, question mark or quotation marks followed by newline character. I am using words separated by hyphen and apostrophe as a single word. The logic is that every matched pattern of space counts for the word it follow. Therefore, to count the last word of paragraph we need to specify these exceptions.

Sentence:
Regular Expression 1: **/\.␣+[A-Z]|\.␣\"|\.\n|\.\"\n|\.\"|\?␣[A-Z]|\?\n|[a-zA-Z]!+|[a-zA-Z]!+\n|\.\)␣+|\.\)\n/**

Regular Expression 2: **/Mr\\.|␣mr\\.|␣mrs\\.|Mrs\\.|Dr\\.|␣dr\\.|␣[A-Z]\\.␣|Jr\\.␣|␣jr\\.␣|Sr\\.␣|␣sr\\.␣|St\\.␣|␣st\\./**

To search number of sentences in the article, I subtract results of expression 2 from expression 1. Expression 1 searches for all the suspected patterns for sentence. Presence of 'period' indicates the end of sentence. But this symbol is also used for abbreviations. So, I designed some patterns to look for end of sentences. For example, the first letter of new sentence is always (approximately) uppercase. So, I put this as one of the patterns to look for while searching for sentences. If the period signifies the end of last sentence of paragraph, it is followed by 'newline' character. This is another pattern to search for.

Another way a new sentence can start with is quotation marks. Or the sentence could end in quotation marks or in brackets. These patterns are also included in the search expression. Other characters that can mark the end of sentence are exclamation mark or question mark. And like for period, they may also represent the end of last sentence of paragraph. So, they are dealt in similar fashion as period. Sometimes, there are multiple exclamation marks but they can represent the end of just one sentence. It is necessary to precede an exclamation mark in pattern with a lowercase or uppercase letter.

In expression 2, some common expressions that can falsely affect the search for sentence are specified.

Paragraphs:
Regular Expression: **/[^\r\n]+((\r|\n|\r\n)[^\r\n]+)*/**

Paragraphs are marked by certain types of newline characters. These are optional i.e. sometimes only '\r' is sufficient or '\n' or maybe '\r\n'. This is basic search and sufficient to search for all types of formats between paragraphs. These combinations of newline characters must follow characters other than themselves.

Programming Technique:
I have used 'sys' and 're' modules for this assignment. The program takes the name of file to be read as an argument using 'sys' module. At first, the desired regular expression is compiled using 're' module. Then, all the instances of pattern are searched through the document. This returns a List of all the matched instances. Number of occurrences of the above pattern are determined by calculating the length of List. For Regular Expressions I have to use extra escape characters because of Python. For example, for newline character I used '\\n' instead of '\n'. One backslash is to escape in 're' module and anther one for Python.

Python Commands used:
1. re.compile("Regular Expression of desired pattern")
2. findall(filename)
3. len(ListName)

Results:
Test Article (Final):
Number of words: 650
Number of Sentences: 30
Number of Paragraphs: 10

Development Article (Initial):
Number of words: 603
Number of Sentences: 22
Number of Paragraphs: 8

This program runs as per the instructions of Professor. It takes the filename from command line and operated on that particular file.