**Sentence Segmentation**

For this sentence segmentation program, I use regex only to split the document into sentences. To compare with the standard library’s performance I apply NLTK on the given input and find my program’s performance very close to it.

Design: Using regex and python’s built in function “split” I segmented the sentences. After reading the file, I removed ‘\n’ characters. For building the regex formula, I considered couple of cases. So, my regex can handle the sentences that end with either by “.”, “?” or “!”. If my regex considers only the segmentation based on “.” it may suffer from decimal points in the number. So, I modified the regex so that it can handle decimal points also. Another important scenario I have to handle is acronyms like Mr., P.h.D, etc. So, my regex formula also handled this type of scenario. I also considered different types of punctuations like “,”, “;”, “:”, “-” etc.

Performance Comparison: To compare the performance, I applied NLTK on the input.txt and found 2123 sentences. My program found 2114 sentences which is close to the standard one. I manually checked some sentences to check the quality of the taks. It looks satisfactory to me.

Limitations: There are some limitations of my program. First , it can’t handle sentences containing couple of dots following a small letter where the dots may use for expressing the flow of the speech. For example: *“It is not enough to stand on this record because we are dealing with the most ruthless, fanatical... leaders that the world has ever seen.”*  My program splitted this into two sentences where NLTK would consider as a single sentences. My program can’t handle dot inside a quotation mark. For example: *The Communists say, "Come with us; look what we've done." And we've been in - on the whole, uninterested.* As the dot is inside of the quoted marks, my program couldn’t split this. Although my program handled the case of acronym like “Mr.” in the middle of the sentence, there is a problem of segmenting this type of scenario like: *And yet they were polled by the U.S.I.A. The point of all this is, this is a struggle in which we're engaged.* During the handling the case of acronym, I overlooked the case that sentence may end with the acronym. So it can’t split the above sentence.

Improvement: I think it is difficult to handle all the cases with regex only. I need to build some heuristics to handle some cases. For example, my heuristic will consider the following word’s first character (Capital or Small) after couple of dots to check whether it is a continuation of the sentence. This heuristic may also solve the last problem I have discussed above. I would like to build another heuristic that may handle dot inside a quotation mark.