- (0% but required) If it is not blatantly obvious, please indicate where in your source code the indexing occurs, and where in your source code the evaluation occurs.

The indexing occurs from line 118 to 138 in my source code and for each query evaluation, there is a helper method with the same name.

- (10%) Description of the system, design tradeoffs, questions you had and how you resolved them, etc. It should be clear how you processed queries and how a new (but largely similar) query would be handled.

In this project, we are asked to find either playId or sceneId for each query, so we need a simple way to transform a docid into its playId or sceneId. Therefore, I use two dictionaries in python to overcome this problem by assigning docid as keys and playId or sceneId as values. Apart from this, I use many helper functions to help me deal with complex problems, so many duplicate codes are avoided and since those functions return result in docid format, I can easily transform them no matter if the query is asking for scene or play.

- (10%) List the software libraries you used, and for what purpose.
    - The implementation must be your own work.

I only used json and re from Python's built-in library. One for processing json files and one for split() using regular expression.

- (5%) Why might counts be misleading features for comparing different scenes? How could you fix this?

Since there is no guarantee that each scene is the same length as the others, only taking into account the number of existence is misleading. For example, comparing the existence of a word in a scene with 50 words and a scene with 50000 words is not telling us anything. To fix this, we can calculate the number of existence of the word divided by the total number of words in the scene and compare this value among different scenes.

- (5%) Which query or queries took the longest to execute? Why might this be the case?

I think the query for phrase1 took the longest to execute, because it has a relatively longer phrase and searching for phrase takes longer time to execute than searching for word.

- (5%) What is the average length of a scene? What are the shortest and longest scenes? What are the shortest and longest plays?

The average length of a scene is 1199.5561497326203. The shortest scene is 'antony_and_cleopatra:2.8' which has a length of 47, and the longest scene is 'loves_labors_lost:4.1' which has a length of 7988. The shortest play has a length of 16415 which is the play 'comedy_of_errors', and the longest play has a length of 32867 which is 'hamlet'.

- (5%) Create a plot of the first query: One series is the count of ("thee" or "thou") vs. `sceneNum` and another is the count of "you" vs. the `sceneNum`.