COMP 6714
Project Part 1
Yu Han
Z5219071

There are three main functions in this program, first function is used for building a dictionary for the documents to record the frequency of each term; second function is used for splitting the query to all possible combinations of entities and tokens; last function is used for calculate the score of each combination and output the combination who has the maximum score.

For the first function, it constructed an entity dictionary at first which includes the entity terms and the frequency of each term, then it traversed all documents to record the tokens. If the word not in entity dictionary and not stop words or punctuations, this word will be record into the token dictionary. Finally, the entity dictionary and tokens dictionary will be sent to the formula and output the score of each term, which named idf tokens and idf entities in this program.

The function for split the query is a little bit complicated. First it compared the query with the DoE to check which entities are valid in this query. And then it used a recursion to output all possible combinations of entities and tokens.

The last stage is used the score calculation formula to output the score of each combination. The higher similarity of combination and document, the higher score for this combination. Finally, the function output the combination with the highest score and format the combination as a tuple (max score, query with max score).