Project part1 report

In this part of project, I split it into 3 function which named index_document, split_query and max_score_query. Method index_document would count the term frequency(TF) of each token and entity present in document set. Furthermore, it computes the IDF and store them for the computation of max score

Method split_query would find all subset of DoE which is combined from query token with increasing order. Moreover, it return the rest token of query.

Method max_score_query compute the TF-IDF score of token and entity, then sum them with a weight. This produce the final score which mean the TF-IDF score of Query with Doc_i.

◆ Part1

The core function of this method is to count the TF and compute the IDF. For those purpose, it gets the list of tokens and entities via spacy library and then travels the property for count the frequency of term

.

While traveling the terms, firstly program travels the entities list. We can recognize mulit-entity by the properties of entity object named start and end, the start and stop index of this entity. After travel all of entities, program visit all of tokens that are not entity via property named ent_iob which could recognize if the token is included in a entity.

Finally,the function compute the IDF via TF counted above.

◆ Part2

Step1: Call function combinations from package itertools for getting all term subset of Q. We can set the parameter r of function combinations to the max length of entity in DoE. This could reduce the time spent in producing subset of DoE.

Step2: Take the intersection of DoE and the subset got above. The new set is the new DoE

Step3: Call function combinations again for getting the entity subset and eliminate the subset whose token count exceed the corresponding token count in Q.

Step4: The processed collection of subset is the final result.

◆ Part3

According to the formula in first part, compute the normalized TF and then get S_{i1} and S_{i2}. We could get the number of docs containing entity e via the length of corresponding value in TF-IDF index dictionary.

At first of this part, initialize the max score to be 0. As long as the score is higher than the max score, set max score to be the current score,