COMP6714

Information Retrieve and Web Search

Project Part 1

Jingshi Yang

z5110579

1.TF-IDF construction for Entities and Tokens

In this step, I iterate over all documents, then use spacy to get all entities and add them into self.tf_entities, meanwhile, I use an array to store the index of single entities by using ent.start and ent.end, if the difference between the two values is one, that means the entity is a single entity. After that, I add all tokens which are not stop words, not punctuation, not belonging to single entity array into self.tokens.

After this step, I can calculate idf_tokens and idf_entities for each token/entity by using given formula.

2. Split Query

In the second step, I follow the instruction given, first I get the maximum length of the entity among all entities in DoE, then use itertools.combinations to find all combinations and set the parameter as maximum length, this can reduce the time of creating combinations. Then I check all combinations I get in the first step whether they are in DoE and add all validated entities combinations into a new array. By using this method, I can ensure the entities in each subset are increasing order.

Next, I get combinations of these validated entities, and filter subsets by checking whether the token count exceed the corresponding token count in Q. After I finish that, I will get a final entities subset. Finally, for each subset, get the corresponding tokens for each subset.

3. Find max score

After I get index construction and some possible query splits. I can get TF-IDF score for each possible query split by using given formulas. Each time I use max_score and max_index to record the maximum score and the index of the query split with maximum score. Finally output the result.