COMP6714_Project_Part1

1. __init__ function

Four variables tf_tokens, tf_entities, idf_tokens and idf_entities are respectively initialized to defaultdict which are used to store the term frequencies and inverse document frequencies for tokens and entities.

2. index_documents function

- Traverse each document, deal with it with spacy, and then format tf_entities, tf_tokens dicts as required respectively.
- And pay attention to restrictions for tokens such as is_stop, is_punct and single-token entity which should be filtered.
- Then traverse tf_entities, tf_tokens and use the given mathematical formulas to get results of idf_entities and idf_tokens.

3. split_query function

- Firstly, select the eligible entities in doe and save them as a list.
- Secondly, get all subsets of these eligible entities.
- Thirdly, select valid subsets.
- Finally, according to the matching entity subsets, the corresponding token and entity combinations of a query are obtained.

4. max_score_query function

- Traverse each query obtained from the previous step, the corresponding token and entity sets are obtained respectively.
- For the entities set, TF-IDF of each entity is calculated by the corresponding TF and IDF calculation methods, and the accumulation is saved by entities_score. Similarly, TF-IDF corresponding to each token is calculated, and the accumulation is saved with tokens_score.
- Then entities_score and tokens_score are respectively given corresponding weights and added to combined_score. Save all scores to a list.
- Sort the score_list and get the max score query, and then return it.