CS6714 Project Part 1

Adam Yi <z5231521@cse.unsw.edu.au>

1. TF-IDF index construction for Entities and Tokens

1. Use Spacy to tokenize and recognize entities
2. Save location offset of single-word entities to a set (for filter purposes, this makes checking if a token should be filtered out constant time)
3. Filter tokens by checking it’s not a stopword, not a punctuation, and location offset is not in the aforementioned filter set
4. Calculate TF (use collections.Counter for efficiency)
5. Calculate normalized TF
6. Calculate IDF

2. Split the Query into Entities and Tokens

1. Split query into words (just by “ ”, no special treatment)
2. Compute TF for words in query
3. Check all possible entities by selecting subset of words and combining them in order (achieved via itertools.combinations and “ ”.join)
4. For all subset of possible entities, combine them as a corpus and compute TF for word
5. Check that for any word, TF is strictly not larger than the query words TF.
6. Add the subset and its word complement set to result as a possible split.

3. Query Score Computation

1. Query split
2. For each split, compute TF-IDF for entities and tokens
3. Calculate score
4. Pick max and return