COMP6714 Information Retrieval and Web Search

Project 1 Part 2

Han YANG - z5140181

In my solution, there are 3 main parts to do the language processing, i.e., Named Entity Disambiguation, assigning unique identities to the mention identified in the text.

First part - Preprocessing and Index Construction:

In this part, I tokenized mentions' documents and constructed the inverted index for both mention document (men_docs.pickle) and candidate entities'description (parsed_candidate_entities.pickle), same as project-part1.

Second part – Generate Features:

In this part, I generated 4 features for this learning-to-rank model:

1. TF-IDF score of each candidate entity in the index of mention document, in this feature, I treated entity as query and computed total score for all tokens appeared in both entity and mention document. It extracted the relevance of entity among mention document.
2. TF-IDF score of mention document in parsed entity index, to construct this feature, I treated the whole mention document as query and computed the total score for all tokens appeared in both document and parsed entity description. This feature reflected the relevance of mention document among entity description.
3. The rate of length of entity and mention, this feature reflected the similarity between mention and candidate entity in some cases.
4. The number of common words in mention document and entity description, to construct this feature I counted the common words and normalized it by the number of unique words in mention document.

Third part – Train and Test:

In this part, I used XGBoost to train and test the model, the parameters are {'max_depth':8, 'eta':0.05, 'n_estimators':5000, 'objective':'rank:pairwise', 'min_child_weight':0.01, 'lambda':100}, num_boost_round=4900.