Project_part2 Report

In this file, there are 6 functions:

disambiguate_mentions(train_mentions, train_labels, dev_mentions, men_docs, parsed_entity_pages),
preprocessing(men_docs, parsed_entity_pages),
doc_tfidf(wiki_term, doc_tf, doc_idf, mentions),
missing_words(wiki_words, doc_words, mentions),
length_diff(mentions),
mention_candidate_words_sim(mentions).

Spacy, re, numpy, xgboost and math.log were imported in this file.

Implementation:

1. *disambiguate_mentions(train_mentions, train_labels, dev_mentions, men_docs, parsed_entity_pages)*
   This function aggregates all the functions that create features and finally construct the model by training and use the model to predict the result.

2. *preprocessing(men_docs, parsed_entity_pages)*
   Take the mention_documents and parsed_entity_pages and tokenize them into lowercase lemmas. For parsed_entity_pages, the words which were adj., adv. and det. were removed, for mention_documents, stop words and punctuations were removed. Besides, term frequency and idf for mention_documents were calculated in this function.

3. *doc_tfidf(wiki_term, doc_tf, doc_idf, mentions)*
   In this function, the tf-idf of specific tokens from parsed_entity_pages in mention_documents were calculated. Only the first 25 terms in parsed_entity_pages were chosen.

4. *missing_words(wiki_words, doc_words, mentions)*
   The proportion of missing words in mention_documents comparing to parsed_entity_pages was calculated. If a word occurs in the mention_documents but not in parsed_entity_pages, that word is regarded as missing word.

5. *length_diff(mentions)*
   For each mention, the differences between the length of that mention and the lengths of its candidate entities were compared and the proportion of the difference to the total length of mention was calculated.

6. *mention_candidate_words_sim(mentions)*
   In this function, the similarity of mentions and their candidate entities were calculated. For each single word in one candidate entity, if the word was in its mention word, its position in mention was found and the following words were checked. So in this case, if two words differ too much, their similarity will be low.

In the training, 4 features were used (document_tf-idf, missing word, length_difference, similarity), the parameters were: {'objective': 'rank:pairwise', 'max_depth': 7, 'eta': 0.05, 'lambda': 100, 'min_child_weight': 0.01, 'subsample': 0.5}.