COMP6714

Information Retrieve and Web Search

Project Part 2

Jingshi Yang

z5110579

1.TF-IDF construction for Tokens

In this step, I iterate over all mention documents, then use spacy to get all single entities by using an array to store the index of ent.start and ent.end of the entity, if the difference between the two values is 1, 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 for each token by using given formula.

2. Features

In function disambiguate_mentions, I find three features for both training and test dataset. The first feature is tf-idf score for all tokens of a candidate entity. The second feature is similarity between each candidate and the mention, which is a percentage showing how many words in mention are also in the candidate. The third feature is the length diff between each candidate entity and the mention. By using the three features, I can get accuracy 83% on dev1 and accuracy 61% on dev2.

3. Features evaluation

I tried to combine some of my features in different ways, then use different subsets of features to compare their own accuracy. Features I have tested are cosine document similarity, mention in candidate entity's text, candidate entity in mention, length diff between candidate and mention, whether the mention's name starts the candidate entity and so on. I can get the accuracy over 90% on dev1 but meanwhile, the accuracy on dev2 is only 57%. Hence finally I decide to use the model with less overfitting.