COMP 6714 Project Part I I

Student ID: z5197332
Student Name: Jiling Yang

Introduction

Using tf-idf values to solve Named Entity Linking (NEL). It aims at using the $XGBoost$ classifier to build learning-to-rank model to find unique identities to the specific mention identified in the text.

Method

Using tf-idf score to compute relationship between the given mentions and candidate entities.
The key of tf_tokens dictionary is the token's name. Its value contains a dictionary which contains documents titles as key and the total number of occurrences of the token in those documents as values.

Progress

1. Generating Features and labels for Training Data
2. Constructing the same features for dev_data
3. Transforming data for XGBoost
4. Modelling training and doing prediction
5. Calculating prediction scores of each testing group

Features

Feature 1 is the sum of tokens tf-idf values in parsed_entity_pages document for each entity in candidate_entities list. The condition for calculating the sum is to sum all tokens when its entity-tag is 'O' but pos-tag is not 'NOUN' or when its entity-tag is 'PROPN' but pos-tag is 'B-GPE' or 'I-ORG'. The purpose for these conditions is to reduce calculate tf-idf values and avoid overfitting.

Feature 2 is the total number of tokens in mention.

Feature 3 is the sum of tf-idf values of all tokens in mention, delimited by space.

Feature 4 is the total number of capitalized tokens in mention, delimited by space.

Feature 5 is the total number of tokens in each entity within the candidate_entities list, delimited by underscore.

Feature 6 is the sum of tf-idf values of tokens within candidate_entities list, delimited by underscore.

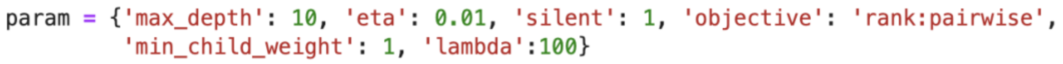
Feature 7 is the total number of capitalized tokens in each entity within the candidate_entities list.

Feature 8 is the total number of same tokens in entity and mentions with lower case.

Feature 9 is the total number of same tokens in entity and mentions.

Feature 10 is the rate between the total number of same tokens in entity and mentions with lower case and the length of entity within the candidate_entities list.

Parameters in $XGBoost$



```
param = {'max_depth': 10, 'eta': 0.01, 'silent': 1, 'objective': 'rank:pairwise',
         'min_child_weight': 1, 'lambda':100}
```

Discussion and improvement

The first version of coding calculates tf-idf values of all tokens in $parsed\_candidate\_entities$ document. However, it leads to overfitting to data set. After modification, I several conditions were set to avoid overfitting.