# Project_Part2 Report

The purpose of project_part2 is to find the characteristics of better representative candidates entities and get a higher accuracy rate when predict.

- The first part is to get token and entity by using Spacy, and then calculate the corresponding TF and IDF.
- The second part is to calculate the features according to the given data set.
  1. Feature1: For each mention, loop its all candidate entity, which has some tokens in 'parsed_entity_pages', calculate the sum of each token's tf-idf.
  2. Feature2: For each candidate entity, calculate the sum of tf-idf in entities set.
  3. Feature3: Calculate the sum of tf-idf of each word in each mention.
  4. Feature4: The same number of words in each candidate entity and mention as a feature.
  5. Feature5: The tf-idf of the lower case of token.
  6. Feature6: The number of words in candidate entity.
  7. Feature7: The number of words in mention.
  8. Feature8: The difference between each candidate entity string length and mention string length.
- Then get the label of each candidate entity according to the "train.mentions" and "train.label" files. In each item, if the candidate entity and corresponding mention are same, and then set label to 1, otherwise set label to 0. And collect the length of each 'candidate_entities' set as train group.
- Next, according to the xgboost model to get the prediction results.