CS 839 Data Science

Project Stage 1

＜Group 14＞

1. The names of all team members.
   - Wen-Fu Lee (wlee256@wisc.edu)
   - Yuan-Ting Hsieh (yhsieh28@wisc.edu)
   - Yahn-Chung Chen (chen666@wisc.edu)
2. The entity type that you have decided to extract, give a few examples of mentions of this entity type.
   - We decide to extract person name in movie reviews. Ex. “Now that Chris and his girlfriend”, we want to extract Chris, "Based on the novel by Stephen King", we want to extract Stephen King.
3. The total number of mentions that you have marked up.
   - 1631
4. The number of documents in set I, the number of mentions in set I.
   - 221 documents, number of mentions is 1117
5. The number of documents in set J, the number of mentions in set J.
   - 103 documents, number of mentions is 514
6. The type of the classifier that you selected after performing cross validation on set I *the first time*, and the precision, recall, F1 of this classifier (on set I). This classifier is referred to as classifier M in the description above.
   - Type of the classifier: random forest
   - Precision: 0.828162
   - Recall: 0.621307
   - F1: 0.709974
7. The type of the classifier that you have finally settled on *before* the rule-based postprocessing step, and the precision, recall, F1 of this classifier (on set J). This classifier is referred to as classifier X in the description above.
   - Type of the classifier: random forest
   - Precision: 0.845606
   - Recall: 0.692607
   - F1: 0.761497
8. If you have done any rule-based post-processing, then give examples of rules that you have used, and describe where can we find all the rules (e.g., is it in the code directory somewhere?).
   - We have constructed a list of stop words, which no only contains typical stop words like I, my, mine, we also add movie specific ones like Production. Because we know that it can’t be a person’s name. Ex. If our classifier thinks ‘Sony Production’ is positive example, then we will take it off in the post process phase.
   - We have also kept a list of prefix, such as Dr., lawyer, CEO, and those can’t be names either. So we will remove those positive examples if they contain prefix.

    - The final rule is that, we know some celebrities in movie/film area. So if we see exact matches of those names in our false-predicted examples, we will change them to positive.
    - The codes can be found in the function **post_processing** of the **classifier_training.py** in the code folder.

9. Report the precision, recall, F1 of classifier Y (see description above) on set J. This is the final classifier (plus rule-based post-processing if you have done any).
    - Precision: 0.902062
    - Recall: 0.680934
    - F1: 0.776053
10. If you have not reached precision of at least 90% and recall of at least 60%, provide a discussion on why, and what else can you possibly do to improve the accuracy.
    - Our performance reaches the requirement.