# CS 839 Project Stage 1 Report

Zeping Ren, Stephen Sheen, Shiqi Yang

## 1 Entity Type

We extracted location names from Winter Olympics news from different news sources (US, UK and Canada). This includes continent names, country names, state names, county names, and city names.

Examples: "Africa", "United States", "Washington", "Jeongseon", and "Sochi".

## 2 Training and Test Set Information

We labeled 315 documents with 1479 mentions. Then, we shuffled the documents and selected the first 200 documents to be set $I$ and the next 100 documents to be Set $J$. The leftover 15 documents were not considered.

Set $I$:

- Number of documents: 200
- Number of mentions: 958

Set $J$:

- Number of documents: 100
- Number of mentions: 485

## 3 Model Information

### 3.1 Word selection

We extract the strings of length 1, 2 and 3 words first. Then remove the phrases that has small-case words (so that we reduce the number of negative examples) or contains stop words (e.g. "a", "the", "should", etc.).

### 3.2 Features

According to our discussion, we decided to use the hash value of each string as the main feature, because in Winter Olympic news, there are some location names that are very much frequently used across the news (e.g. Pyeongchang, Korea, and countries on the top of the medal rankings). Additional features include the word the occurs right before the examples (for example, prepositions such as "in", "from" and verbs such as "defeat" are likely to be followed by a place name), and after the examples (words followed by "said" might as well be a person name instead of a place name).

### 3.3 Results

Out of the suggested models, we chose SVM as it had the best performance. Shown below are the reported values on classifier $M$ before we performed debugging.

- Precision: 0.97
- Recall: 0.67

- F1: 0.79

We also tried other models, some of which also achieved decent metrics. For example, random forest has a recall of 0.87, but since the precision was still lower than 90 even after we tried to debug, we decided not to use this model.

After debugging (mostly fixed label-tagging errors; we also used the RBF kernel on our SVM model), we obtained the classifier $X$ with the following metrics:

- Precision: 0.99
- Recall: 0.76
- F1: 0.86

Since we did not have any rule-based post-processing steps, this is the final classifier $Y$. The precision, recall and F1 values on set $J$ are as follows:

- Precision: 0.99
- Recall: 0.76
- F1: 0.86