

CS839 –Project Stage 1 Report

Team Members:

- Sreyas krishna Natarajan(snatarajan6@wisc.edu)
- Sugadev Chellakkannu(chellakkannu@wisc.edu)
- Vignesh Thirunavukkarasu(thirunavukka@wisc.edu)

Entity Type:

- We extracted the **Person names** (cricket players) from the articles of CricBuzz repository (webpage for the Cricket Sport).
- Person names are marked up in the documents using `<person>...</person>` tags
- Examples:
 1. `<person>Virat Kohli</person>` is the captain of India
 2. `<person>Hashim Amla</person>` has had enough of anchoring the innings and now he's going for the big shots.
 3. Now, `<person>David Warner</person>` walks over and offers a hug to `<person>Dilshan</person>`.

Documents

- **Overall:**
 - Number of Documents: 305
 - Number of Mentions: 1817
- **Set I:**
 - Number of Documents: 202
 - Number of Mentions: 1212
- **Set J:**
 - Number of Documents: 103
 - Number of Mentions: 605

Classifier:

- **Initial (*first time*):** The type of the classifier that was selected after performing cross validation on set I *the first time* was the Logistic Regression classifier.

Type of classifier: Logistic Regression

Precision: 0.8488998121913329

Recall: 0.6202333034870001

F1: 0.6627420418211681

- **Final (*after feature addition*):**

We increased our accuracy by adding additional features by analyzing the documents. We have **34 features in total**.

We got a Precision of 0.9+ and Recall of 0.6+ **without the rule-based postprocessing** step.

Type of classifier: Logistic Regression

Precision: 0.9276828371922088

Recall: 0.6506795739274696

F1: 0.7148101250898973

Precision, Recall, F1 on Set J (Test)

The Best classifier for our dataset was Logistic Regression:

Classifier: Logistic Regression

Precision: 0.9193543846640266

Recall: 0.6398560040653727

F1: 0.7080104161812235