Data Science - Project Stage 1

Person Name Extraction in Baseball match reviews

Team Members:

- Karan Dharni (dharni@wisc.edu)
- Sudarshan Maru (maru@wisc.edu)
- Amogh Joshi (asjoshi4@wisc.edu)

Entity Type: Person Name extraction

Dataset: Baseball (MLB) match reviews/commentaries

Markup Style:

<>John Smith</> Dr.<>Tom Bradley</> <>Jay-Prakash Singh</>

Total number of Documents: 300

Total labeled names (markups): 1238 (Set I: 783, Set J: 455)

Number of Documents in Set I: 200

Number of Documents in Set J: 100

Classifier M: Decision Tree (Selected initially - On set I**)**

Classifier	Precision	Recall	F1 score
Decision Tree	**0.821**	**0.723**	**0.769**

Classifier X: SVM (Selected after debugging False Positives and False Negatives)

NO rule-base post-processing: Achieved desired precision, recall **without** any post-processing or whitelists/blacklists

Final Results:

On Training set (Set I)

Classifier	Precision	Recall	F1 score
SVM	**0.934**	**0.741**	**0.826**

Test set (Set J)

Classifier	Precision	Recall	F1 score
SVM	**0.775**	**0.750**	**0.762**

Discussion:

Selected Dataset: We have chosen the data from MLB match reviews from people on social media sites.

The data is unstructured and has no specific format. There are many grammatical, syntactic and semantic errors which makes it harder to learn correct labels.

Example: John is referred also as “Johny”, “Benjamin” as “Ben”, “Benjy” etc.

Names have article prefixes which is incorrect.

Tokens: We have created tokens of length 1, 2 and 3 as possible candidates for names.

Features selected: We have used **13** features. All features are Boolean (0 /1).

Feature 1: Word Length: Whether a token has more than 2 words or less.

Example: **Alpha Beta Gamma** Feature_value = 1

Example: **Alpha** Feature_value = 0

Feature 2: All Upper Case: Whether a token has all upper-case letters.

Example: **JOHN NASH** Feature_value = 1

Example: **He has** Feature_value = 0

Feature 3: Prefix Word starts with an Upper Case:

Example: Alpha **Beta Gamma** Feature_value = 1

Example: is **Beta Gamma** Feature_value = 0

Feature 4: Suffix Word starts with an Upper Case:

Example: **Alpha Beta** Gamma Feature_value = 1

Example: **Alpha Beta** will Feature_value = 0

Feature 5: If middle word starts with an Upper Case:

Example: **Steve C. Liu** Feature_value = 1

Example: **Jack and Jill** Feature_value = 0

Feature 6: If prefix of a token contains a comma:

Example: Also**, Tom** Feature_value = 1

Example: and **Thomas** Feature_value = 0

Feature 7: If token contains period:

Example: **Steve Thomas** Feature_value = 1

Example: name is **Richard. Anthony** has Feature_value = 0

Feature 8: If token contains a number:

Example: **Ricky3Ponting** Feature_value = 1

Example: **Vince** has 3 cars Feature_value = 0

Feature 9: If token contains email address (‘@’, ‘.edu’ etc):

Example: address is **russell@wisc.edu** Feature_value = 1

Example: **Don Baylor** said Feature_value = 0

Feature 10: If token contains a greetings prefix word (‘thanks’, ‘cheers’, ‘regards’):

Example: Regards, **Anthony** Feature_value = 1

Example: Thanks! **Jay** Feature_value = 0

Feature 11: If suffix is token ends with a ‘s

Example: **John**’s house is big Feature_value = 1

Example: **Regardless** of Feature_value = 0

Feature 12: If suffix contains word (‘writes, ‘wrote’):

Example: **Karan** writes Feature_value = 1

Example: He **can** win Feature_value = 0

Feature 13: If prefix token contains an article (‘a, ‘an’, ‘the’):

Example: the **Yankees** Feature_value = 1

Example: is Jordan Feature_value = 0