An Introduction to Machine Learning and Natural Language Processing Tools

Presented by: Nick Rizzolo, Mark Sammons, Vivek Srikumar, and James Clarke

Advisor: Dan Roth
Motivation: the News
Motivation: the News
Motivation: the News
Motivation: the News
Motivation: the News
News readers help with organization
News readers help with organization

I can search for Rod Blagojevich.
News readers help with organization

I can search for Rod Blagojevich.

What about people like Rod Blagojevich?
Maybe Politicians or Athletes or Corporate Moguls…
What we need:

\[ f(\text{Obama}) = \text{“politician”} \]
\[ f(\text{Jeter}) = \text{“athlete”} \]
\[ f(\text{Bill Gates}) = \text{“CEO”} \]
Where to get it: Machine Learning

Feature Functions

Learning Algorithm

\[ f \]
Where to get it: Machine Learning

Data

- "politician"
- "athlete"
- "CEO"

Feature Functions

Learning Algorithm

\( f \)
So, what are “feature functions”?

- Take same input as \( f \)
- Indicate some property of the input a.k.a., a feature
So, what are “feature functions”? 

- Take same input as \( f \) 
- Indicate some property of the input a.k.a., a feature 

**Typical NLP feature functions**

- **Binary**
  - Appearance of a given word
  - Appearance of two words consecutively a.k.a., a bigram
  - Appearance of a word with a given part of speech
  - Appearance of a named entity (e.g. “Barack Obama”)

- **Real**
  - Counts of binary features
  - TFIDF (a statistical measure of a document)
In This Tutorial, We Will…

- Introduce **Learning Based Java (LBJ)**
  - A modeling language for learning and inference

- Introduce our state-of-the-art **NLP tools**
  - Describe their functionality
  - Demonstrate their use in new classifiers
In This Tutorial, We Will…

- **Day 1:**
  - Simple Classifiers
    - 20 Newsgroups
    - Spam detection
    - Language identification

- **Day 2:**
  - The “Fame” Classifier
  - Write a program that:
    - Takes a collection of news articles
    - Outputs lists of people organized by what they’re famous for