Methods in Computational Linguistics I

Andrew Rosenberg
August 27, 2010
Today

• Get to know each other
• Get to know the Python Environment
• Overview of the course
• Bureaucracy
  • Syllabus
  • Course Policies
  • Course Website
Who am I?

- Assistant Professor at Queens College
- Computer Science Department
  - Natural Language Processing
- Speech and Intonation
- Machine Learning
- Worked at Google and IBM
Who are you?

• Where in the graduate school process are you?

• Do you research? What?

• What interests you about language?

• What is your programming, math, statistics, linguistics background?

• What do you hope to get from this class?
Computational Linguistics

- Processing language computationally
- Empirical testing of linguistic hypotheses
- Statistical modeling of language phenomena.
CL in this course

- Computational Linguistics brings together computer science, statistics and linguistics.
- This class:
  - Light on statistics.
  - Light on linguistics.
  - Big on computer science.
Major Victories of Computational Linguistics

• Computational Linguistics provides Linguistics a method for **testing** hypotheses.

• Spell checking

• Google (and many others processing language on the web)

• Spoken Dialog Systems

• Spam/Phishing detection
Doing Computational Linguistics

• What are the major steps?
• What are our major tools?
Python

- Python is an “interpretive”, “procedural” programming language.
- Shallow learning curve.
- Great support for computational linguistics.
- Easy string manipulation
- NLTK - Natural Language Tool Kit
Installing Python

- Make sure you can do this on your home computer.
- This is important.
- Download Python.
- Download NLTK and all necessary libraries.
Welcome to the Python Interpreter

• Math Demo - Math in Python
• Numerical Precision
• SyntaxError
Math Functions

• importing libraries
• Exponentiation: \texttt{pow(base, power)}
• Logarithm: \texttt{log(x, base)}
• Square root: \texttt{sqrt(x)}
• Trigonometry functions
  • \texttt{sin(x)}, \texttt{cos(x)}, \texttt{tan(x)}
• SyntaxError
Strings

• Strings Demo
  • What’s a string?
  • Name splitting
  • TypeError
String functions

- Casing: upper() lower()
- find and replace
- formatting
- we’ll come back to these frequently
Download and Install NLTK

Corpus Demo
Course Website

• And syllabus
Major CL tasks

- Part of Speech Tagging
- Parsing
- Word Net
- Named Entity Recognition
- Information Retrieval
- Sentiment Analysis
- Document Clustering
- Topic Segmentation
- Authoring
- Machine Translation
- Summarization
- Information Extraction
- Spoken Dialog Systems
- Natural Language Generation
- Word Sense Disambiguation
Part of Speech Tagging

• Task: Given a string of words, identify the parts of speech for each word.

A man walks into a bar.
Part of Speech Tagging

• Surface level syntax.
• Primary operation
  • Parsing
  • Word Sense Disambiguation
  • Semantic Role labeling
• Segmentation
  • Discourse, Topic, Sentence
How do we do it?

• Learn from Data.
• Annotated Data:
  A man walks into a bar.
• Unlabeled Data:
  A man walks home.
The pitcher issued four walks.
### Part of speech tagging

<table>
<thead>
<tr>
<th></th>
<th>Det</th>
<th>Noun</th>
<th>Verb</th>
<th>Prep</th>
<th>Adj</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.9</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>man</td>
<td>0.0</td>
<td>0.6</td>
<td>0.2</td>
<td>0.0</td>
<td>0.2</td>
</tr>
<tr>
<td>walks</td>
<td>0.0</td>
<td>0.2</td>
<td>0.8</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>into</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
</tr>
<tr>
<td>bar</td>
<td>0.0</td>
<td>0.7</td>
<td>0.3</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>
Limitations

• Unseen tokens
• Uncommon interpretations
• Long term dependencies
Next Time

- Introduction to the Natural Language Toolkit (NLTK)