Announcements – Assignments

- **Homework 01**
  - Due tomorrow night

- **Readings:**
  - Reading 01 – due last night, get it in ASAP if haven’t
  - Reading 02 – link course site, due Sunday

- **Week 2 Tutorials:**
  - 2.1 – Tokenization, lemmatization, stopwords, etc
    - Based on today’s lecture
  - 2.2 – Exploring dictionary-based methods
    - Based on Wednesday’s lecture
Gauri will be lecturing on Regular Expressions and holding open hours during course time
This week’s focus: Words, words, words
Why focus on words?

- Words suggest meaning
- If we can identify words, we can count them
- If we can count words, we can quantify (aspects of) a text that contains those words.
- If we can quantify a text, we can compute with it.
  - Answer quantitative questions about text
- Caveat:
  - Quantifying a text isn't the same thing as being correct about what that text means, nor is meaning solely a function of word counts(!).

Matthew Wilkens - https://mattwilkens.com/
What is a word?
Outline

- Tokenization
- Lemmatization
- Stemming
- Stopwords
- Part of Speech
- Dependency Parsing
- Named Entities
“The process of identifying the words in the input sequence of characters, mainly by separating the punctuation marks but also by identifying contractions, abbreviations, and so forth”

Chapter 5
Basic Text Processing In: Text Mining: A Guidebook for the Social Sciences
“Mr. Smith doesn’t like apples.”

How many tokens are in the sentence?
“Mr. Smith doesn’t like apples.”

“The process of identifying the words in the input sequence of characters, mainly by separating the punctuation marks but also by identifying contractions, abbreviations, and so forth”
“Mr. Smith doesn’t like apples.”

“The process of identifying the words in the input sequence of characters, mainly by separating the punctuation marks but also by identifying contractions, abbreviations, and so forth”
“Mr. Smith doesn’t like apples.”

“The process of identifying the words in the input sequence of characters, mainly by separating the punctuation marks but also by identifying contractions, abbreviations, and so forth”
“Mr. Smith doesn’t like apples.”

Mr. Smith does n’t like apples.
**Type vs Token**

- **Type**: An element of the vocabulary
- **Token**: an instance of a type in the text

- \( N \) = number of tokens
- \( \mathcal{V} \) = vocabulary, i.e. set of tokens
- \( |\mathcal{V}| \) = size of Vocabulary
Type vs Token

- **Type**: An element of the vocabulary

- **Token**: an instance of a type in the text

“We refuse to believe that there are insufficient funds in the great vaults of opportunity of this nation. And so we've come to cash this check, a check that will give us upon demand the riches of freedom and the security of justice”

- Q: How many types, tokens?
Lemmatization & Stemming
Lemmatization

“reduces the inflectional forms of a word to its root form”

Boys -> boy
Children -> child
Am, are, is -> be

Chapter 5
Basic Text Processing In: Text Mining: A Guidebook for the Social Sciences
I have a dream that one day even the state of Mississippi, a state sweltering with the heat of injustice, sweltering with the heat of oppression will be transformed into an oasis of freedom and justice.

With this faith we will be able to transform the jangling discords of our nation into a beautiful symphony of brotherhood.
“applies a set of rules to an input word to remove suffixes and prefixes and obtain its stem, which will now be shared with other related words.”

Chapter 5
Basic Text Processing In: Text Mining:
A Guidebook for the Social Sciences

“more radical way to reduce variation”

Chapter 2
Dirk Hovy textbook
An algorithm for suffix stripping

M.F. Porter

Computer Laboratory, Corn Exchange Street, Cambridge
1. INTRODUCTION

Removing suffixes from words by automatic means is an operation which is especially useful in the field of information retrieval. In a typical IR environment, one has a collection of documents, each described by the words in the document title and possibly the words in the document abstract. Ignoring the issue of precisely where the words originate, we can say that a document is represented by a vector of words, or terms. Terms with a common stem will usually have similar meanings, for example:

CONNECT
CONNECTED
CONNECTING
CONNECTION
CONNECTIONS

Frequently, the performance of an IR system will be improved if term groups such as this are conflated into a single term. This may be done by removal of the various suffixes -ED, -ING, -ION, -IONS to leave the single stem ___________. In addition, the suffix stripping process will reduce the total number of terms in the IR system, and hence reduce the size and complexity of the data in the system, which is always advantageous.
“For each language, it defines a number of suffixes (i.e., word endings) and the order in which they should be removed or replaced. By repeatedly applying these actions, we reduce all words to their stems.”

Chapter 2
Dirk Hovy textbook

This was not the map we found in Billy Bones’s chest, but an accurate copy, complete in all things-names and heights and soundings-with the single exception of the red crosses and the written notes.

Frequency of Rubio’s terms in 2016 Miami debate
“set of ignorable words that occur often, but not contribute much to our task, so it can be beneficial to remove.”

Chapter 2
Dirk Hovy textbook
Part of Speech
Categorize words based on their grammatical properties

Part-of-speech tagging:
- Process of identifying the grammatical category of tokens in a corpus
<table>
<thead>
<tr>
<th>Tag</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADJ</td>
<td>Adjective: noun modifiers describing properties</td>
<td>red, young, awesome</td>
</tr>
<tr>
<td>ADV</td>
<td>Adverb: verb modifiers of time, place, manner</td>
<td>very, slowly, home, yesterday</td>
</tr>
<tr>
<td>NOUN</td>
<td>words for persons, places, things, etc.</td>
<td>algorithm, cat, mango, beauty</td>
</tr>
<tr>
<td>VERB</td>
<td>words for actions and processes</td>
<td>draw, provide, go</td>
</tr>
<tr>
<td>PROPN</td>
<td>Proper noun: name of a person, organization, place, etc.</td>
<td>Regina, IBM, Colorado</td>
</tr>
<tr>
<td>INTJ</td>
<td>Interjection: exclamation, greeting, yes/no response, etc.</td>
<td>oh, um, yes, hello</td>
</tr>
<tr>
<td>ADP</td>
<td>Adposition (Preposition/Postposition): marks a noun’s spacial, temporal, or other relation</td>
<td>in, on, by under</td>
</tr>
<tr>
<td>AUX</td>
<td>Auxiliary: helping verb marking tense, aspect, mood, etc.,</td>
<td>can, may, should, are</td>
</tr>
<tr>
<td>CCONJ</td>
<td>Coordinating Conjunction: joins two phrases/clauses</td>
<td>and, or, but</td>
</tr>
<tr>
<td>DET</td>
<td>Determiner: marks noun phrase properties</td>
<td>a, an, the, this</td>
</tr>
<tr>
<td>NUM</td>
<td>Numeral</td>
<td>one, two, first, second</td>
</tr>
<tr>
<td>PART</td>
<td>Particle: a preposition-like form used together with a verb</td>
<td>up, down, on, off, in, out, at, by</td>
</tr>
<tr>
<td>PRON</td>
<td>Pronoun: a shorthand for referring to an entity or event</td>
<td>she, who, I, others</td>
</tr>
<tr>
<td>SCONJ</td>
<td>Subordinating Conjunction: joins a main clause with a subordinate clause such as a sentential complement</td>
<td>that, which</td>
</tr>
<tr>
<td>PUNCT</td>
<td>Punctuation</td>
<td>; , ()</td>
</tr>
<tr>
<td>SYM</td>
<td>Symbols like $ or emoji</td>
<td>$, %</td>
</tr>
<tr>
<td>X</td>
<td>Other</td>
<td>asdf, qwfg</td>
</tr>
<tr>
<td>Tag</td>
<td>Meaning</td>
<td>English Examples</td>
</tr>
<tr>
<td>-----</td>
<td>---------</td>
<td>-----------------</td>
</tr>
<tr>
<td>ADJ</td>
<td>adjective</td>
<td>new, good, high, special, big, local</td>
</tr>
<tr>
<td>ADP</td>
<td>adposition</td>
<td>on, of, at, with, by, into, under</td>
</tr>
<tr>
<td>ADV</td>
<td>adverb</td>
<td>really, already, still, early, now</td>
</tr>
<tr>
<td>CONJ</td>
<td>conjunction</td>
<td>and, or, but, if, while, although</td>
</tr>
<tr>
<td>DET</td>
<td>determiner, article</td>
<td>the, a, some, most, every, no, which</td>
</tr>
<tr>
<td>NOUN</td>
<td>noun</td>
<td>year, home, costs, time, Africa</td>
</tr>
<tr>
<td>NUM</td>
<td>numeral</td>
<td>twenty-four, fourth, 1991, 14:24</td>
</tr>
<tr>
<td>PRT</td>
<td>particle</td>
<td>at, on, out, over per, that, up, with</td>
</tr>
<tr>
<td>PRON</td>
<td>pronoun</td>
<td>he, their, her, its, my, I, us</td>
</tr>
<tr>
<td>VERB</td>
<td>verb</td>
<td>is, say, told, given, playing, would</td>
</tr>
<tr>
<td>.</td>
<td>punctuation marks</td>
<td>. , ; !</td>
</tr>
<tr>
<td>X</td>
<td>other</td>
<td>ersatz, esprit, dunno, gr8, univeristy</td>
</tr>
</tbody>
</table>
Closed class words

• Relatively fixed membership
• Usually **function** words: short, frequent words with grammatical function
  • determiners: *a, an, the*
  • pronouns: *she, he, I*
  • prepositions: *on, under, over, near, by, …*

Open class words

• Usually **content** words: Nouns, Verbs, Adjectives, Adverbs
  • Plus interjections: oh, ouch, uh-huh, yes, hello
• New nouns and verbs like iPhone or to fax
Word Classes Graphic

Open class ("content") words

Nouns
- Proper: Janet, Italy
- Common: cat, cats, mango

Verbs
- Main: eat, went

Adjectives
- old, green, tasty

Adverbs
- slowly, yesterday

Interjections
- Ow, hello

Numbers
- 122,312, one

Closed class ("function")

Determiners
- the, some

Conjunctions
- and, or

Auxiliary
- can, had

Prepositions
- to, with

Particles
- off, up

Pronouns
- they, its

Copyright © 2016 Barnard College

Dependency Parsing
The idea in dependency grammar is that the sentence “hangs” off the main verb like a mobile. The links between words describe how the words are connected.
<table>
<thead>
<tr>
<th>Universal DP Tags</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Clausal Argument Relations</strong></td>
<td><strong>Description</strong></td>
</tr>
<tr>
<td>NSUBJ</td>
<td>Nominal subject</td>
</tr>
<tr>
<td>DOBJ</td>
<td>Direct object</td>
</tr>
<tr>
<td>IOBJ</td>
<td>Indirect object</td>
</tr>
<tr>
<td>CCOMP</td>
<td>Clausal complement</td>
</tr>
<tr>
<td>XCOMP</td>
<td>Open clausal complement</td>
</tr>
<tr>
<td><strong>Nominal Modifier Relations</strong></td>
<td><strong>Description</strong></td>
</tr>
<tr>
<td>NMOD</td>
<td>Nominal modifier</td>
</tr>
<tr>
<td>AMOD</td>
<td>Adjectival modifier</td>
</tr>
<tr>
<td>NUMMOD</td>
<td>Numeric modifier</td>
</tr>
<tr>
<td>APPOS</td>
<td>Appositional modifier</td>
</tr>
<tr>
<td>DET</td>
<td>Determiner</td>
</tr>
<tr>
<td>CASE</td>
<td>Prepositions, postpositions and other case markers</td>
</tr>
<tr>
<td><strong>Other Notable Relations</strong></td>
<td><strong>Description</strong></td>
</tr>
<tr>
<td>CONJ</td>
<td>Conjunct</td>
</tr>
<tr>
<td>CC</td>
<td>Coordinating conjunction</td>
</tr>
</tbody>
</table>
Examples of tags

<table>
<thead>
<tr>
<th>Relation</th>
<th>Examples with head and dependent</th>
</tr>
</thead>
<tbody>
<tr>
<td>NSUBJ</td>
<td>United canceled the flight.</td>
</tr>
<tr>
<td>DOBJ</td>
<td>United diverted the flight to Reno.</td>
</tr>
<tr>
<td></td>
<td>We booked her the first flight to Miami.</td>
</tr>
<tr>
<td>IOBJ</td>
<td>We booked her the flight to Miami.</td>
</tr>
<tr>
<td>NMOD</td>
<td>We took the morning flight.</td>
</tr>
<tr>
<td>AMOD</td>
<td>Book the cheapest flight.</td>
</tr>
<tr>
<td>NUMMOD</td>
<td>Before the storm JetBlue canceled 1000 flights.</td>
</tr>
<tr>
<td>APPOS</td>
<td>United, a unit of UAL, matched the fares.</td>
</tr>
<tr>
<td>DET</td>
<td>The flight was canceled.</td>
</tr>
<tr>
<td>CONJ</td>
<td>Which flight was delayed?</td>
</tr>
<tr>
<td>CC</td>
<td>We flew to Denver and drove to Steamboat.</td>
</tr>
<tr>
<td>CASE</td>
<td>We flew to Denver and drove to Steamboat.</td>
</tr>
<tr>
<td></td>
<td>Book the flight through Houston.</td>
</tr>
</tbody>
</table>
Dependency Parsing - Example

I prefer the morning flight through Denver
Classify words into predefined categories:

- persons
- organizations
- locations
- expressions of times
- quantities
- monetary values
- percentages
Named Entity Recognition

- Classify words into predefined categories:
  - persons
  - organizations
  - locations
  - expressions of times
  - quantities
  - monetary values
  - percentages

Monday, October 30, Hillary Clinton will present her book in Chicago at the University of Chicago.
Named Entity Recognition

- Classify words into predefined categories:
  - persons
  - organizations
  - locations
  - expressions of times
  - quantities
  - monetary values
  - percentages

Slide from Federico Nanni
Approaches for NER

- regular expression to extract:
- Gazetteers
- Patters
- Machine Learning
Approaches for NER – Regular Expressions

- Extract:
  - telephone numbers
  - E-mails
  - Dates
  - Prices
  - Locations (e.g., word + “river” indicates a river -> Hudson river)
Approaches for NER - Gazetteers

- Dictionaries or list of proper names of:
  - Person
  - Location
  - Organization
Approaches for NER – Context Patterns

- context patterns, such as:
  - [Person] earns [Money]
  - [PERSON] joined [ORGANIZATION]
  - [PERSON] fly to [LOCATION]