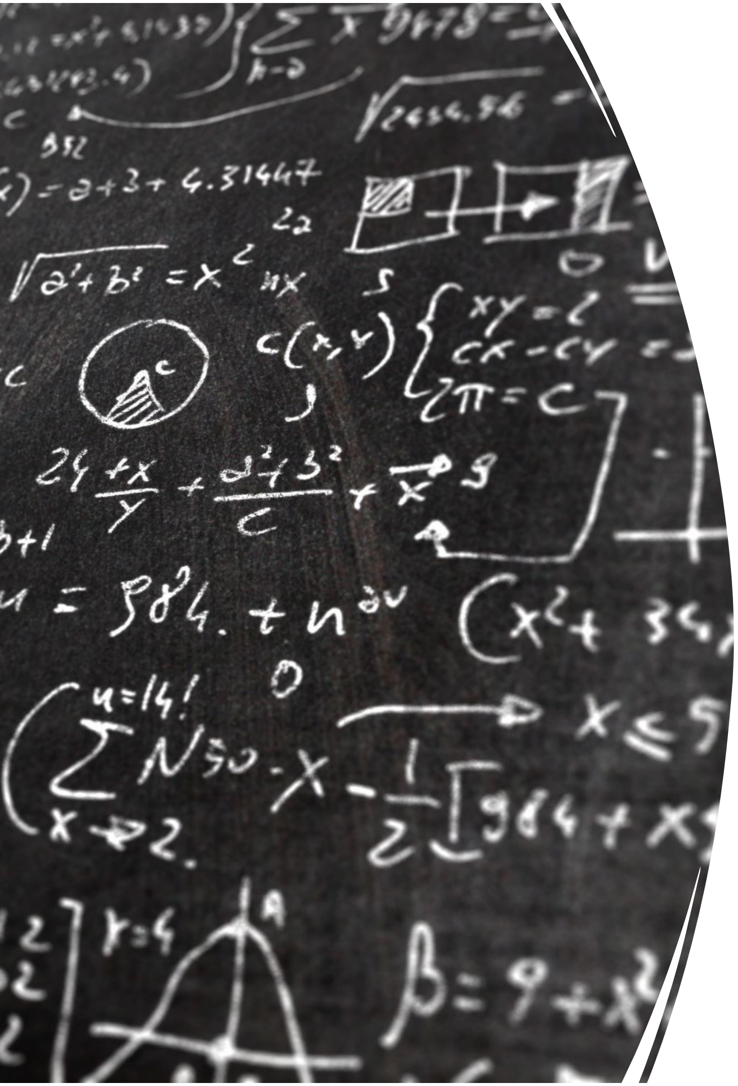


CS 383 – Computational Text Analysis

Lecture 1

Adam Poliak

01/18/2023

What is Computational Text

Computational Text Analysis for Social Science: Model Assumptions and Complexity

Brendan O'Connor* **David Bamman†** **Noah A. Smith†***
*Machine Learning Department

BIG DATA & SOCIET

ECR Forum

Big Data & So
July–Decembe
© The Autho
Reprints and
sagepub.com/
DOI: 10.1177
bds.sagepub.c
SAGE

Computational text analysis: Thoughts on the contingencies of an evolving method

Daniel Marciniak

Abstract
Mapping a public discourse with the tools of computational text analysis comes with many contingenci
corpus curation, data processing and analysis, and visualisation. However, the complexity of algorithm

B
D
& S

Commentary

Big D
July–
© T
Repr
sage
DOI:
bds.s
S

Adapting computational text analysis to social science (and vice versa)

Paul DiMaggio

Abstract
Social scientists and computer scientist are divided by small differences in perspective and disciplinary divide. In the field of text analysis, several such differences are noted: social scientists models to explore corpora, whereas many computer scientists employ supervised models to tra hold to more conventional causal notions than do most computer scientists, and often favor existing algorithms, whereas computer scientists focus more on developing new models; and com trust human judgment more than social scientists do. These differences have implications that pot practice of social science.

Keywords
Topic models, text analysis, unsupervised models, interpretation, sentiment analysis, supervised

Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts

Justin Grimmer
Department of Political Science, Stanford University, Encina Hall West 616 Serra Street, Stanford, CA 94305
e-mail: jgrimmer@stanford.edu (corresponding author)

Brandon M. Stewart
Department of Government and Institute for Quantitative Social Science, Harvard University, 1737 Cambridge Street, Cambridge, MA 02138
e-mail: bstewart@fas.harvard.edu

Edited by R. Michael Alvarez

Politics and political conflict often occur in the written and spoken word. Scholars have long recognized this, but the massive costs of analyzing even moderately sized collections of texts have hindered their use in political science research. Here lies the promise of automated text analysis: it substantially reduces the costs of analyzing large collections of text. We provide a guide to this exciting new area of research and show how, in many instances, the methods have already obtained part of their promise. But there are pitfalls to using automated methods—they are no substitute for careful thought and close reading and require extensive and problem-specific validation. We survey a wide range of new methods, provide guidance on how to validate the output of the models, and clarify misconceptions and errors in the literature. To conclude, we argue that for automated text methods to become a standard tool for political scientists, methodologists must contribute new methods and new methods of validation.

What is Data Science?

- *“Data science is the study of extracting value from data” –*

  Jeannette Wing

What is Data Science?

- *"Data science is the study of extracting **<u>value</u>** from data" – Jeannette Wing*

- Value
  - Requires domain expertise to determine what value is
  - *Value from data* is different based on the domain and the needs

What is Data Science?

- *"Data science is the study of **<u>extracting</u>** value from data" –*

  Jeannette Wing

- Extracting
  - emphasizes action on data
  - mining information

What is Computational Text Analysis?

Computational Text Analysis

practice

- *"~~Data science is~~ the stud~~y of~~ extracting value from data" –*

large scale textual

Jeannette Wing

Adam Poliak

Computational Text Analysis

- *Computational text analysis is not a replacement for but rather an addition to the approaches one can take to analyze social and cultural phenomena using textual data. By moving back and forth between large-scale computational analyses and small-scale qualitative analyses, we can combine their strengths so that we can identify large-scale and long-term trends, but also tell individual stories*

How we do things with words …

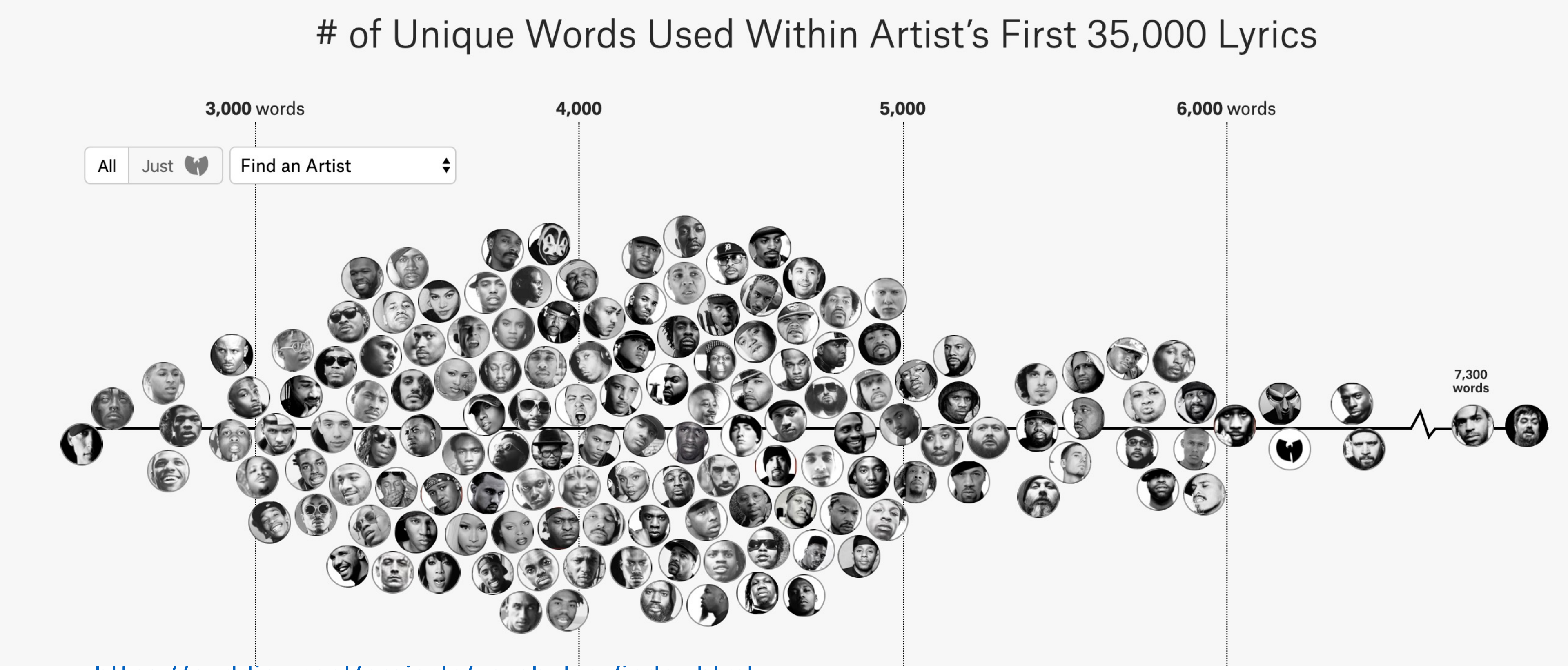
Computational Text Analysis

- *Computational text analysis is not a replacement for but rather an addition to the approaches one can take to analyze social and cultural phenomena using textual data. By moving back and forth between* ***large-scale computational analyses*** *and small-scale qualitative analyses, we can combine their strengths so that we can identify large-scale and long-term trends, but also tell individual stories*

What can we do with computational text analysis?

What can we do with large scale textual analysis?

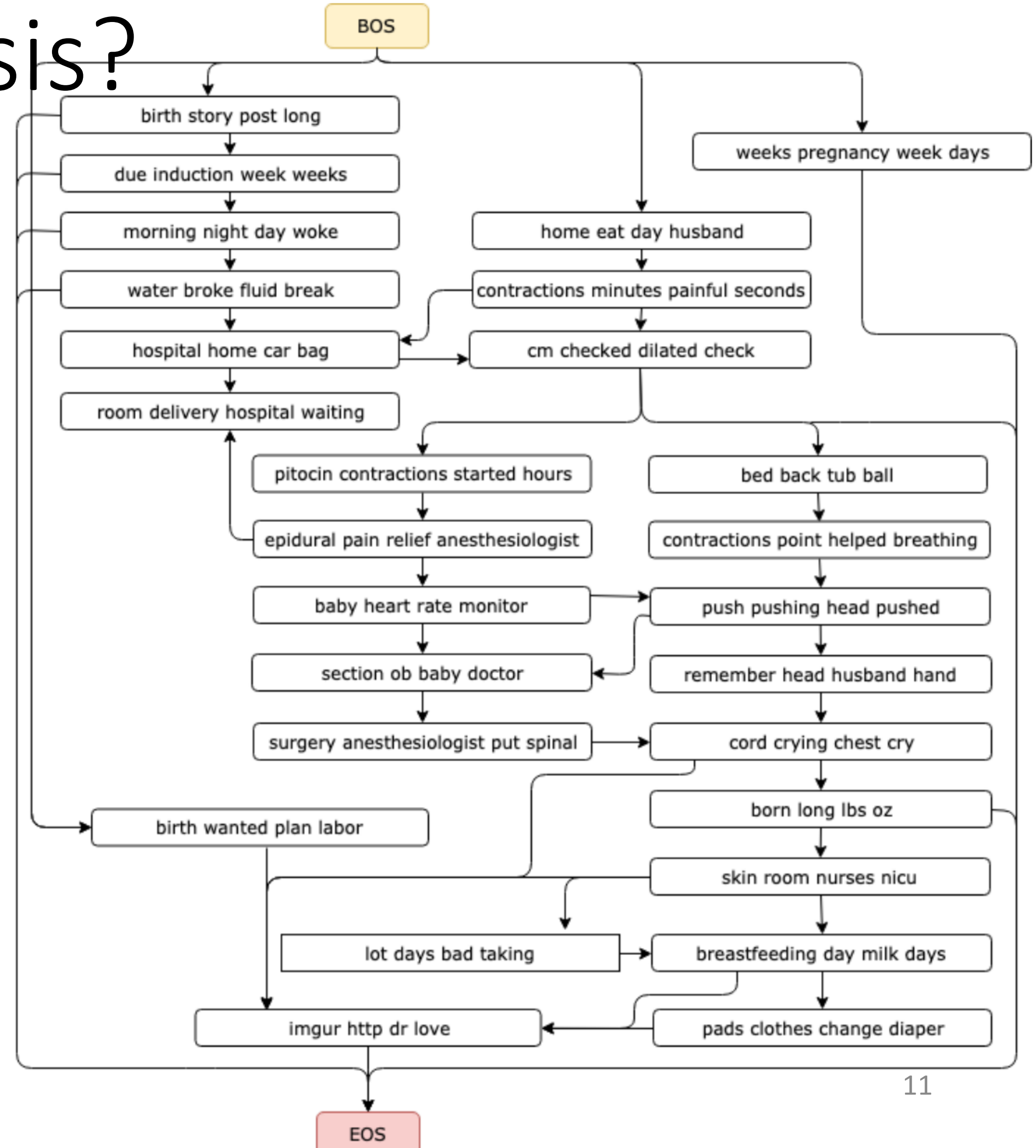
- Sort artists by their vocabulary



https://pudding.cool/projects/vocabulary/index.html

What can we do with large scale textual analysis?

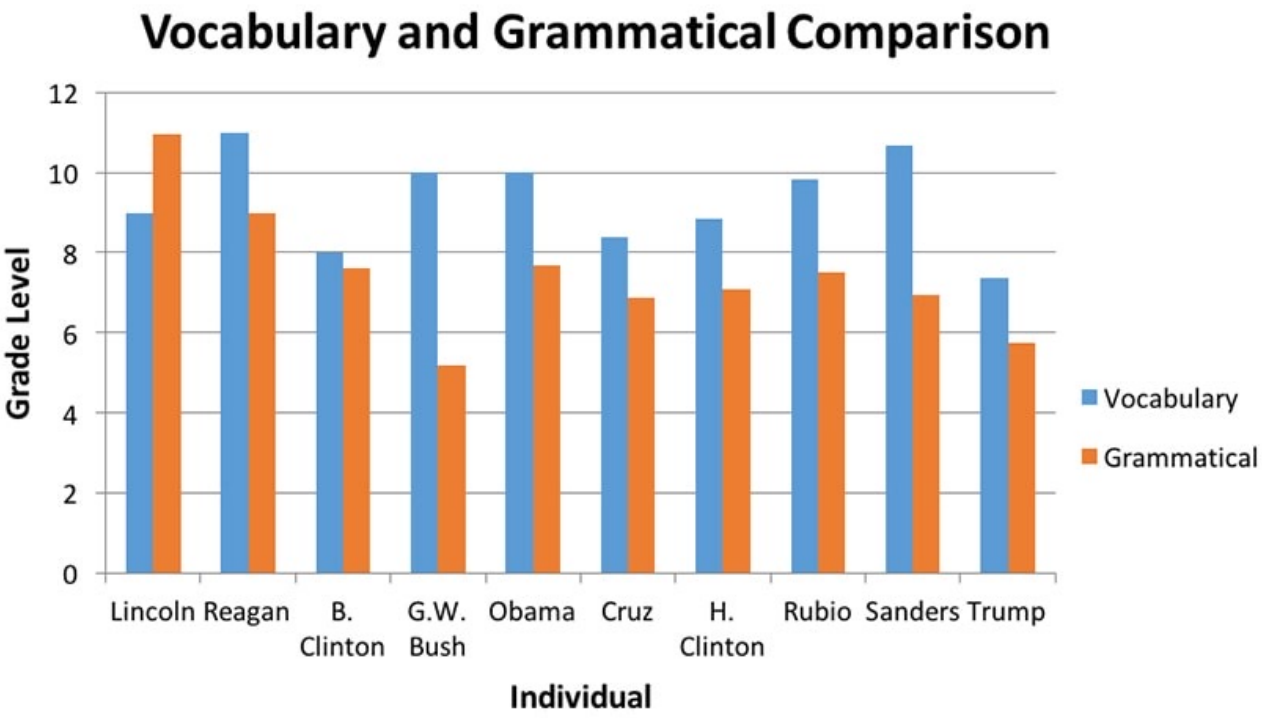
- Identify flow of topics in birthing narratives



https://maria-antoniak.github.io/resources/2019_cscw_birth_stories.pdf

What can we do with large scale textual analysis?

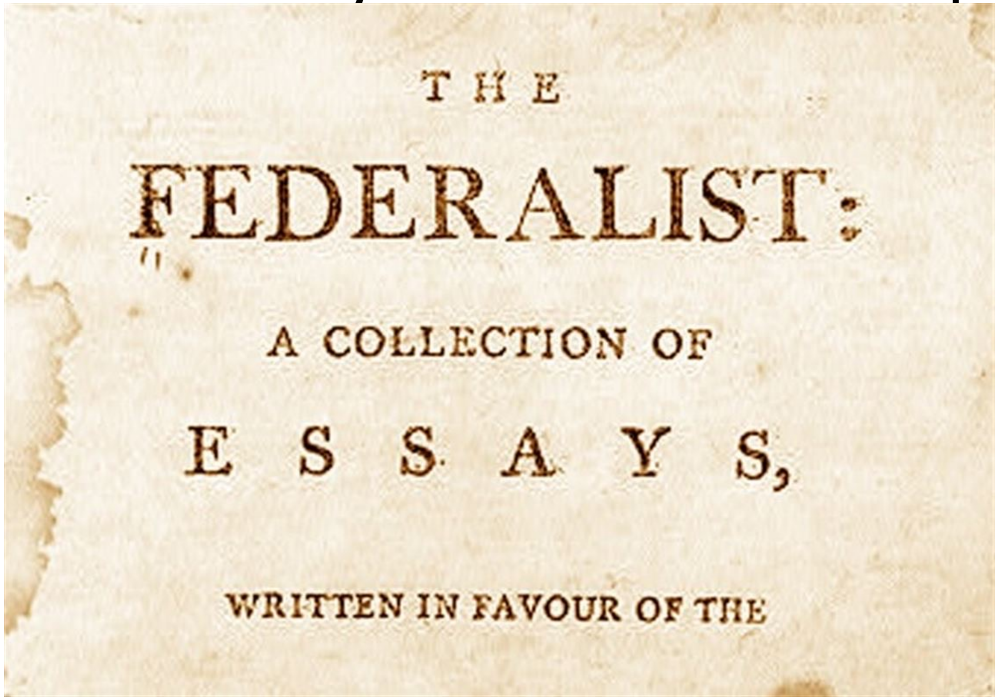
- Categorize the level of presidential candidates' speeches



https://arxiv.org/pdf/1603.05739.pdf

What can we do with large scale textual analysis?

- Who wrote the anonymous Federalist Papers?



https://www.jstor.org/stable/2283270

investigations | joshua freedman and dan jurafsky

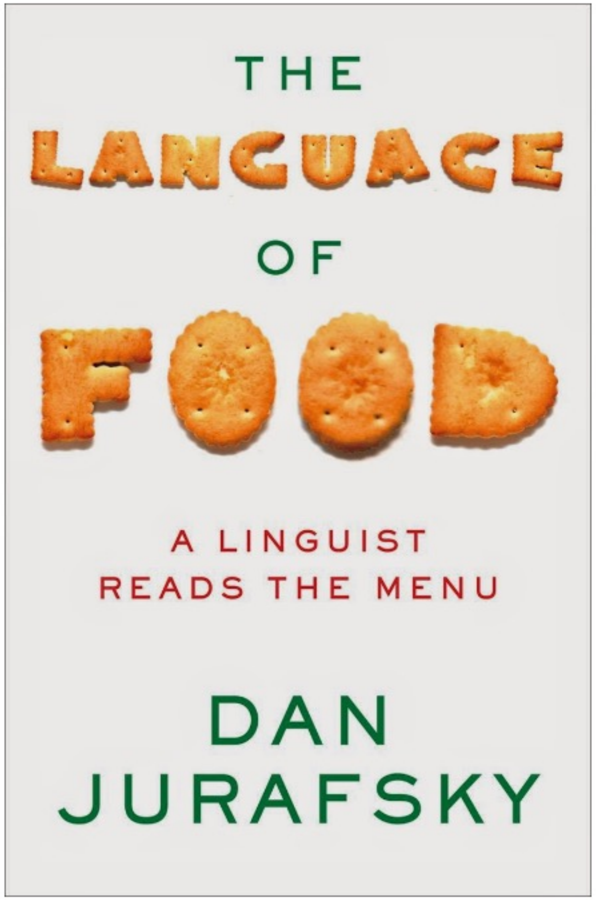
Authenticity in America

Class Distinctions in Potato Chip Advertising

Naturalness/Ingredients in Expensive Chips		Historicity/Locality in Inexpensive Chips	
Naturalness	all natural	**Historicity**	using an old family recipe
Naturalness	great taste…naturally	**Historicity**	time-tested standard
Naturalness	nothing fake or phony	**Historicity**	almost 85-year-old recipe
Naturalness	still made with all natural oil	**Historicity**	a time-honored tradition
Naturalness	totally natural	**Historicity**	since 1986
Naturalness	absolutely nothing artificial	**Historicity**	since 1921
Naturalness	only real food ingredients	**Historicity**	the chips that built our company
Ingredients	Yukon Gold potatoes	**Historicity**	Jim Herr, Founder
Ingredients	Sea salt	**Historicity**	Bill and Sally Utz believed
Ingredients	only the finest potatoes	**Location**	in the shadow of the Cascade Mountains
Process	hand-rake every batch	**Location**	made in the great Pacific Northwest
Process	kettle cooked	**Location**	classic American snacks
Process	special cooking techniques	**Location**	freshness and authenticity of the islands

https://online.ucpress.edu/gastronomica/article-abstract/11/4/46/44534/Authenticity-in-America-Class-Distinctions-in

What can we do with large scale textual analysis?



What can we do with large scale textual analysis?

A Lot!

Computational Text Analysis in this course

- Aggregate large scale textual data
- Text Processing
- Discovering patterns in data
  - Applying NLP/ML tools to text

Course Objectives

Learn and master the methods behind:

1. Natural Language Processing & Text-based Machine Learning

2. Aggregate large scale textual data

3. Discovering patterns in data

4. Complete an independent research project

Course Outline

- Text Processing, Unsupervised Learning

  4 weeks

- Supervised Machine Learning

  3 weeks

- Hypothesis Testing

  2 weeks

- Data Collection

  2 weeks

- Advanced Topics

  2 weeks

Logistics

Communication

- Course webpage:
  - https://cs.brynmawr.edu/cs383-cta
- Piazza:
  - Online discussion board
- Gradescope:
  - Submitting assignments

Lectures

- Live classes
  - Primarily lectures
  - Q/A
  - Recorded
  - Discussions

- Readings:
  - Readings associated with the lecture’s material
    - Make sure to read before lecture
  - Distributed on course schedule

Assesement

- Midterms
    - March 2nd
    - April 13th
    - flexible grading policy

- Final Exam

Assignments & Assessment

- Weekly long homeworks
- Reading reflections
- Midterm – Wednesday 04/12
- Final Project (pairs)

Reading reflections

- Usually due Friday midnight
- For each reading:
  - 3-4 sentence summary
  - 1 sentence about something in particular that you like
  - 1 sentence about something you didn't like or something you found confusing and you'd like me to explain
  - 1 question for future work
- Goal: Examples of computational text analysis
  - Preparation for final projects
- Complete individually

Homeworks

- A mix of programming and written analysis
  - Usually given starter code
- Implement methods covered in class
- Must completely individually

Final Project

- Develop Research Question
- Collect Textual Data to Answer Question
- Data Exploration & Analysis
- Machine Learning
- Can use toolkits/APIs

Final Project – Deliverables

- Project ideation – TBD
- Project proposal – TBD
- Project presentations – Wed 04/26 (last day of classes)
- Project submissions – end of finals

Assignment Logistics

- Distribution:
  - Course website
  - Can work on your own machines or CS lab machines
- Gradescope (for submission)
- Final project:
  - Likely use CS lab machines

Participation Grade

- During class meetings:
  - Topic discussion
  - Asking questions
- Asynchronous
  - Active on Piazza

Course staff

Adam Poliak (apoliak@brynmawr.edu)

- PhD in Computer Science from Johns Hopkins University
- Taught 2 years at Barnard
  - Data Science and this course (for non-majors)
- 2nd semester at BMC
- Research:
  - Natural Language Processing
  - Data Science applied to text data

Our job is to help you succeed!

Course Policies

Collaboration

- Encouraged to discuss problems
- Do not share solutions

Late Days

- Late Days – 10 late days
- Can use at most 2 late days on an assignment
- Can be used only on homeworks and reading responses

Announcements – Assignments

- Homework 00
  - Due Monday night

- Readings:
  - Reading 01 – due Monday night (available already)
  - Reading 02 – due next Friday night (posted later this week)

Today’s focus:
Words, words, words

Why focus on words?

- Words suggest meaning
- If we can identify words, we can count them
- If we 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?

How many words?

I am planning to play a new show in New York before going to watch a new play

Outline

- Tokenization
- Lemmatization
- Stemming
- Stopwords
- Part of Speech
- Dependency Parsing
- Named Entities

Tokenization

Tokenization

"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

Tokenization - Example

<u>“Mr. Smith doesn’t like apples.”</u>

How many tokens are in the sentence?

Tokenization - Example

“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”

Tokenization - Example

“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”

Tokenization - Example

“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”***

Tokenization - Example

“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
- ***V*** = vocabulary, i.e. set of tokens
- ***|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"

Chapter 5

Basic Text Processing In: Text Mining:

A Guidebook for the Social Sciences

boys ->

children ->

am, are, is -> b

Lemmatization - example

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.*

Stemming

"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

Porter Algorithm for Stemming

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 [redacted] 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.

Porter Stemming Explained

"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

https://www.cs.toronto.edu/~frank/csc2501/Readings/R2_Porter/Porter-1980.pdf

Stemming Example

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.

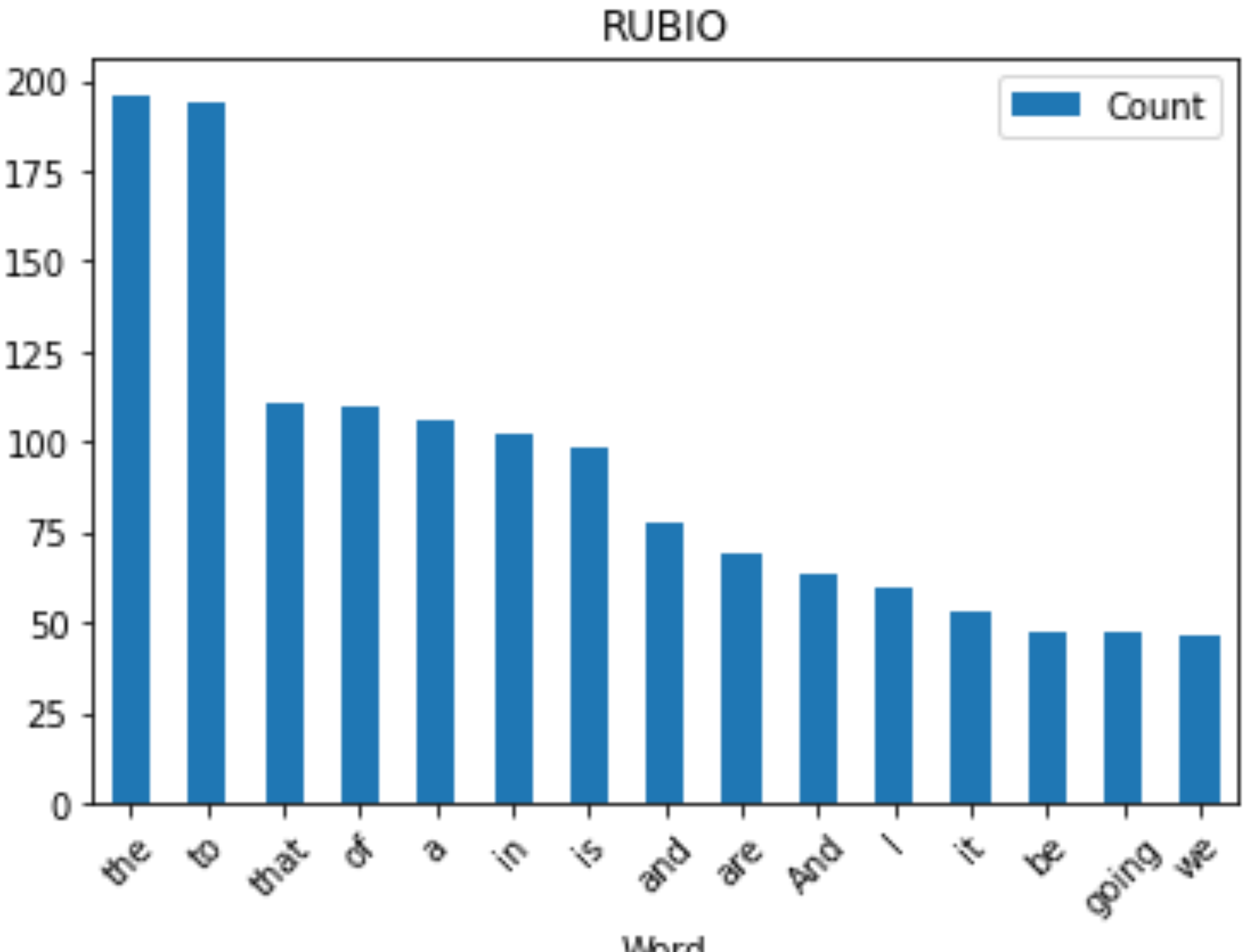
→

Thi wa not the map we found in Billi Bone s chest but an accur copi complet in all thing name and height and sound with the singl except of the red cross and the written note .

[Example from: https://web.stanford.edu/~jurafsky/slp3/slides/2_TextProc_Mar_25_2021.pdf](https://web.stanford.edu/~jurafsky/slp3/slides/2_TextProc_Mar_25_2021.pdf)

Stop Words

Frequency of Rubio’s terms in 2016 Miami debate



Stopwords

"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

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

Universal Tag Set

Tag	Description	Example
ADJ	Adjective: noun modifiers describing properties	*red*, *young*, *awesome*
ADV	Adverb: verb modifiers of time, place, manner	*very*, *slowly*, *home*, *yesterday*
NOUN	words for persons, places, things, etc.	*algorithm*, *cat*, *mango*, *beauty*
VERB	words for actions and processes	*draw*, *provide*, *go*
PROPN	Proper noun: name of a person, organization, place, etc..	*Regina*, *IBM*, *Colorado*
INTJ	Interjection: exclamation, greeting, yes/no response, etc.	*oh*, *um*, *yes*, *hello*
ADP	Adposition (Preposition/Postposition): marks a noun's spacial, temporal, or other relation	*in, on, by under*
AUX	Auxiliary: helping verb marking tense, aspect, mood, etc.,	*can, may, should, are*
CCONJ	Coordinating Conjunction: joins two phrases/clauses	*and*, *or*, *but*
DET	Determiner: marks noun phrase properties	*a, an, the, this*
NUM	Numeral	*one, two, first, second*
PART	Particle: a preposition-like form used together with a verb	*up, down, on, off, in, out, at, by*
PRON	Pronoun: a shorthand for referring to an entity or event	*she, who, I, others*
SCONJ	Subordinating Conjunction: joins a main clause with a subordinate clause such as a sentential complement	*that*, *which*
PUNCT	Punctuation	; , ()
SYM	Symbols like $ or emoji	$, %
X	Other	asdf, qwfg

Simplified Tag set

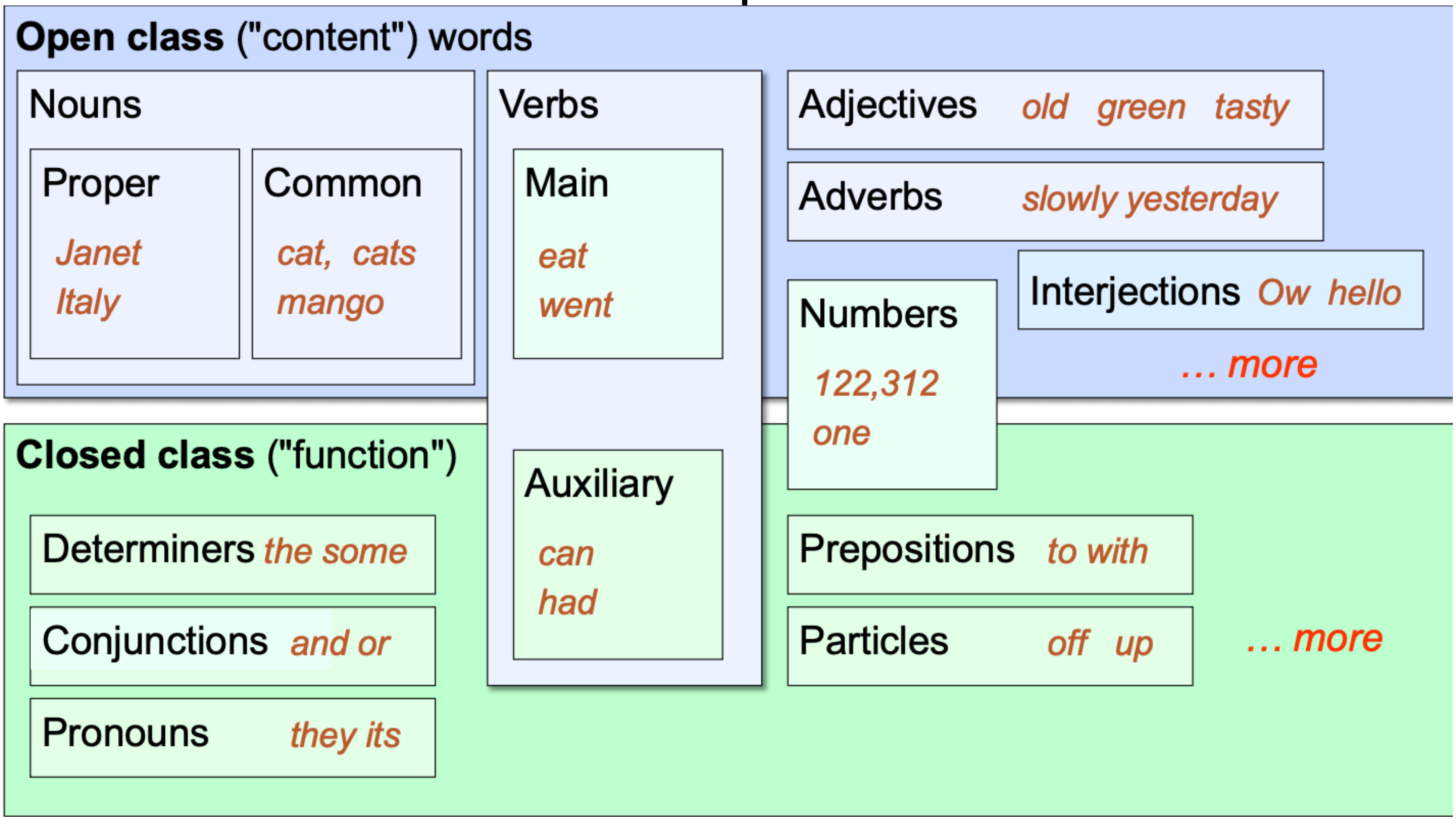
Tag	Meaning	English Examples
ADJ	adjective	*new, good, high, special, big, local*
ADP	adposition	*on, of, at, with, by, into, under*
ADV	adverb	*really, already, still, early, now*
CONJ	conjunction	*and, or, but, if, while, although*
DET	determiner, article	*the, a, some, most, every, no, which*
NOUN	noun	*year, home, costs, time, Africa*
NUM	numeral	*twenty-four, fourth, 1991, 14:24*
PRT	particle	*at, on, out, over per, that, up, with*
PRON	pronoun	*he, their, her, its, my, I, us*
VERB	verb	*is, say, told, given, playing, would*
.	punctuation marks	*. , ; !*
X	other	*ersatz, esprit, dunno, gr8, univeristy*

Word Classes: Open vs Closed

- 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

Slide from https://web.stanford.edu/~jurafsky/slp3/slides/8_POSNER_intro_May_6_2021.pdf

Word Classes Graphic



Dependency Parsing

Dependency Parsing - Idea

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.

Chapter 2

Dirk Hovy textbook

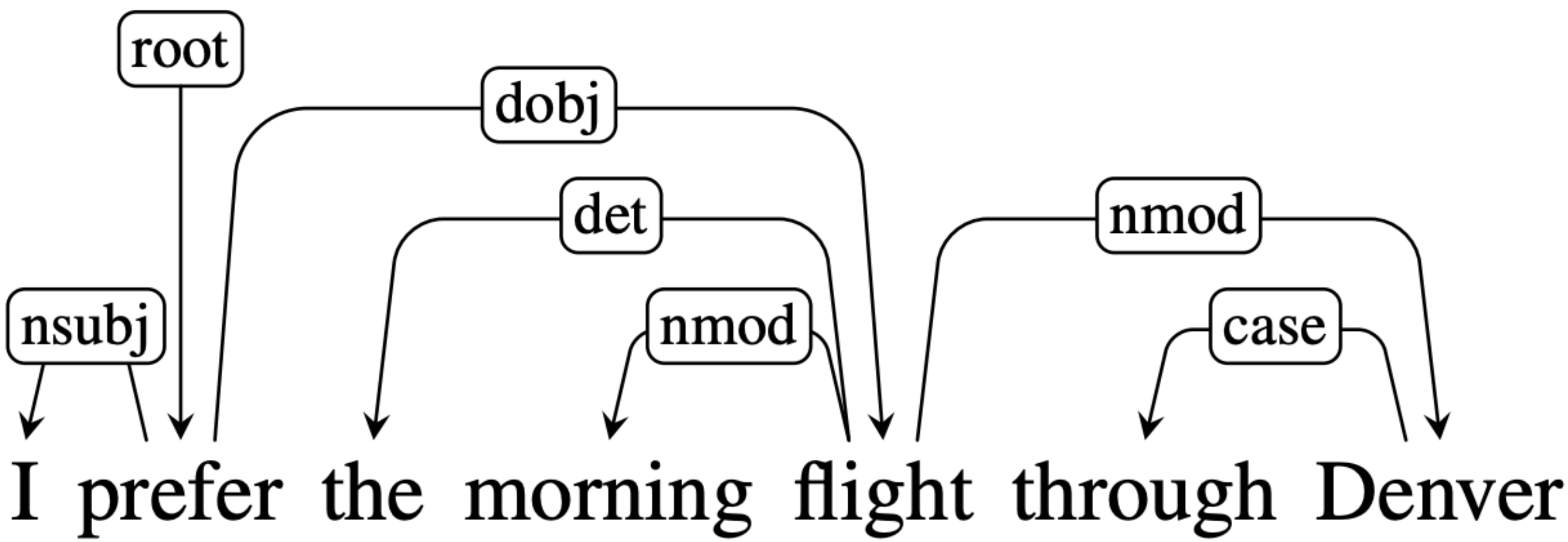
Universal DP Tags

Clausal Argument Relations	Description
NSUBJ	Nominal subject
DOBJ	Direct object
IOBJ	Indirect object
CCOMP	Clausal complement
XCOMP	Open clausal complement
Nominal Modifier Relations	**Description**
NMOD	Nominal modifier
AMOD	Adjectival modifier
NUMMOD	Numeric modifier
APPOS	Appositional modifier
DET	Determiner
CASE	Prepositions, postpositions and other case markers
Other Notable Relations	**Description**
CONJ	Conjunct
CC	Coordinating conjunction

Examples of tags

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

Dependency Parsing - Example



Named Entities

Named Entity Recognition

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

Slide from Federico Nanni

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.

Slide from Federico Nanni

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.

Slide from Federico Nanni

Approaches for NER

- regular expression to extract:
- Gazetteers
- Patterns
- Machine Learning

Slide from Federico Nanni

Approaches for NER – Regular Expressions

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

Slide from Federico Nanni

Approaches for NER - Gazetteers

- Dictionaries or list of proper names of:
  - Person
  - Location
  - Organization

Slide from Federico Nanni

Approaches for NER – Context Patterns

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

Slide from Federico Nanni

Summary

- Course overview & Logistics

- Simple Text Processing: words, words, words
  - Tokenization
  - Lemmatization
  - Stemming
  - Stopwords
  - Part of Speech
  - Dependency Parsing
  - Named Entities

TODOs

- Read the assigned reading for Monday's lecture:
  - Language Modeling
- HW00
- Reading Response 01