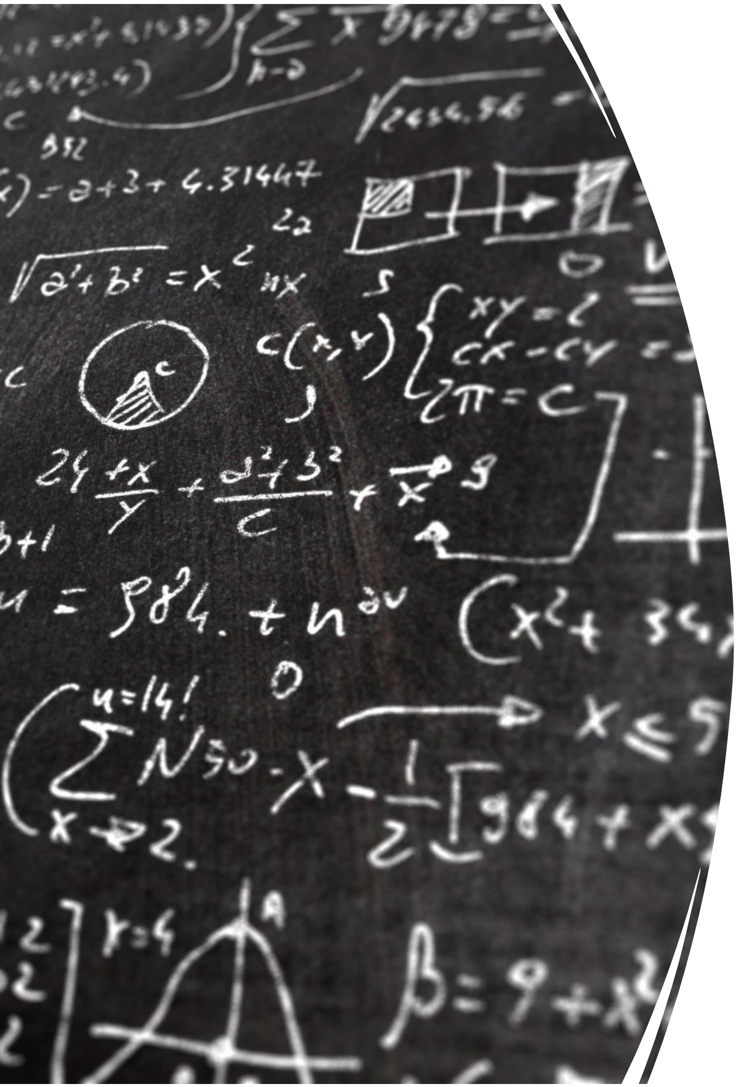


CS 383 – Computational Text Analysis

Lecture 7
LDA, Unsupervised misc

Adam Poliak

02/08/2023

Slides adapted David Mimno, Jordan Boyd-Graber

Announcements

- Office Hours:
  - This week: Thursday 3:30-4:30pm
    - Possibly 3-4:30, will confirm later today
- HW02 due tonight Wednesday 02/08
- Reading 03 released Monday
  - Due Monday 02/13
- HW03 due Wednesday 02/15
  - Released Monday
  - About 1.5 weeks

Outline

Topic Modeling

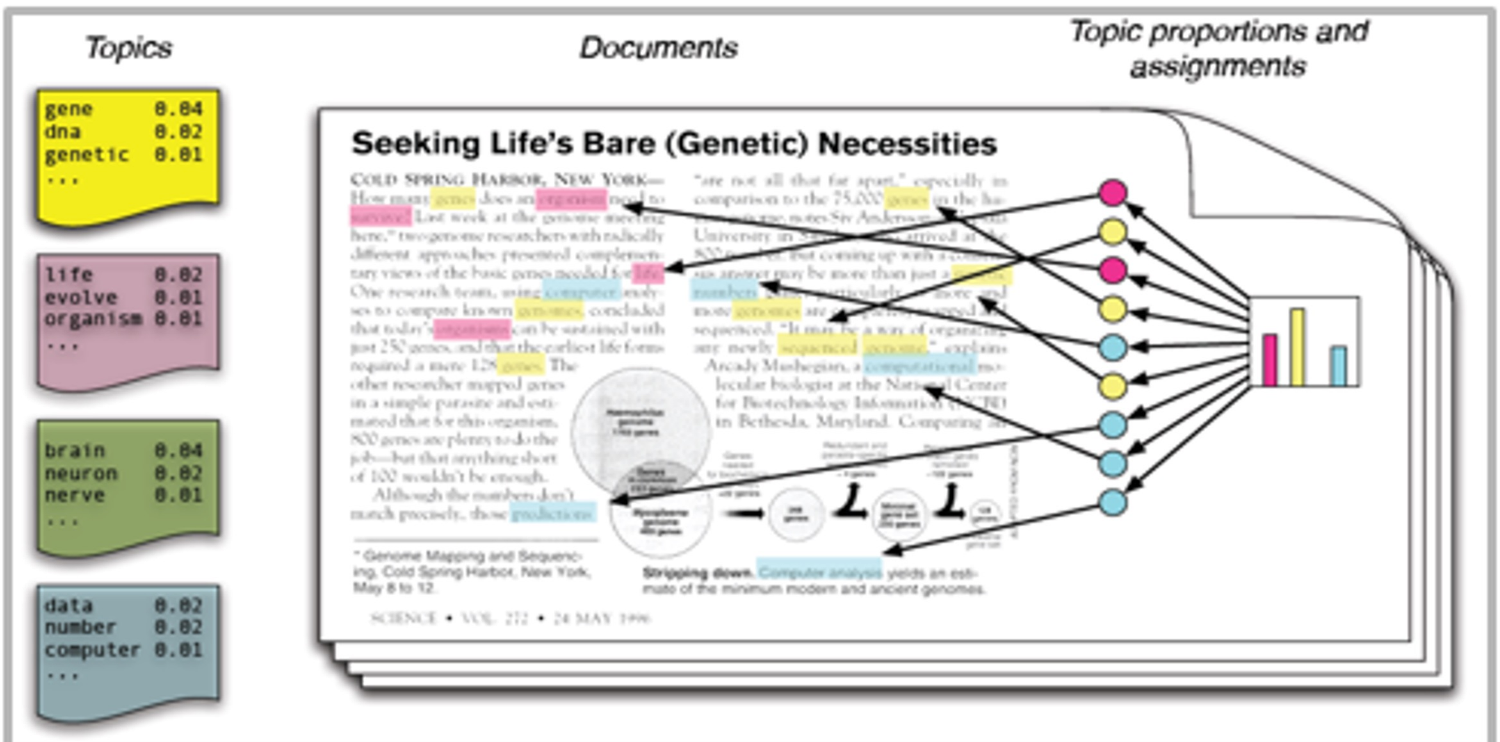
- Topic Model Review
- Evaluating Topic Models
- Implementation

Dictionary based methods

PMI

Topic Modeling

- Goal: Identify underlying topics across documents



 Slide from Federico Nanni

What are topics?

Observation

Tokens that are likely to appear in the same context

Hidden structure that determines how **tokens** appear in a corpus

Want to uncover

 Slide from Federico Nanni

Topic Modeling: Corpora -> Topics

Input:

Millions of Books

Output: topics

(distributions over words)

killed wounded sword slain arms military rifle wounds loss
human Plato Socrates universe philosophical minds ethics
inflammation affected abdomen ulcer circulation heart
ships fleet sea shore Admiral vessels land boats admiral
sister child tears pleasure daughters loves wont sigh warm
sentence clause syllable singular examples clauses syllables
provinces princes nations imperial possessions invasion
women Quebec Women Iroquois husbands thirty whom
steam engines power piston boilers plant supplied chimney
lines points direction planes Lines scale sections extending

Each row is a topic

 Slide from David Mimno

How do we discover topics?

- Latent Semantic Analysis
- Probabilistic Latent Semantic Analysis
- **Latent Dirichlet Allocation**

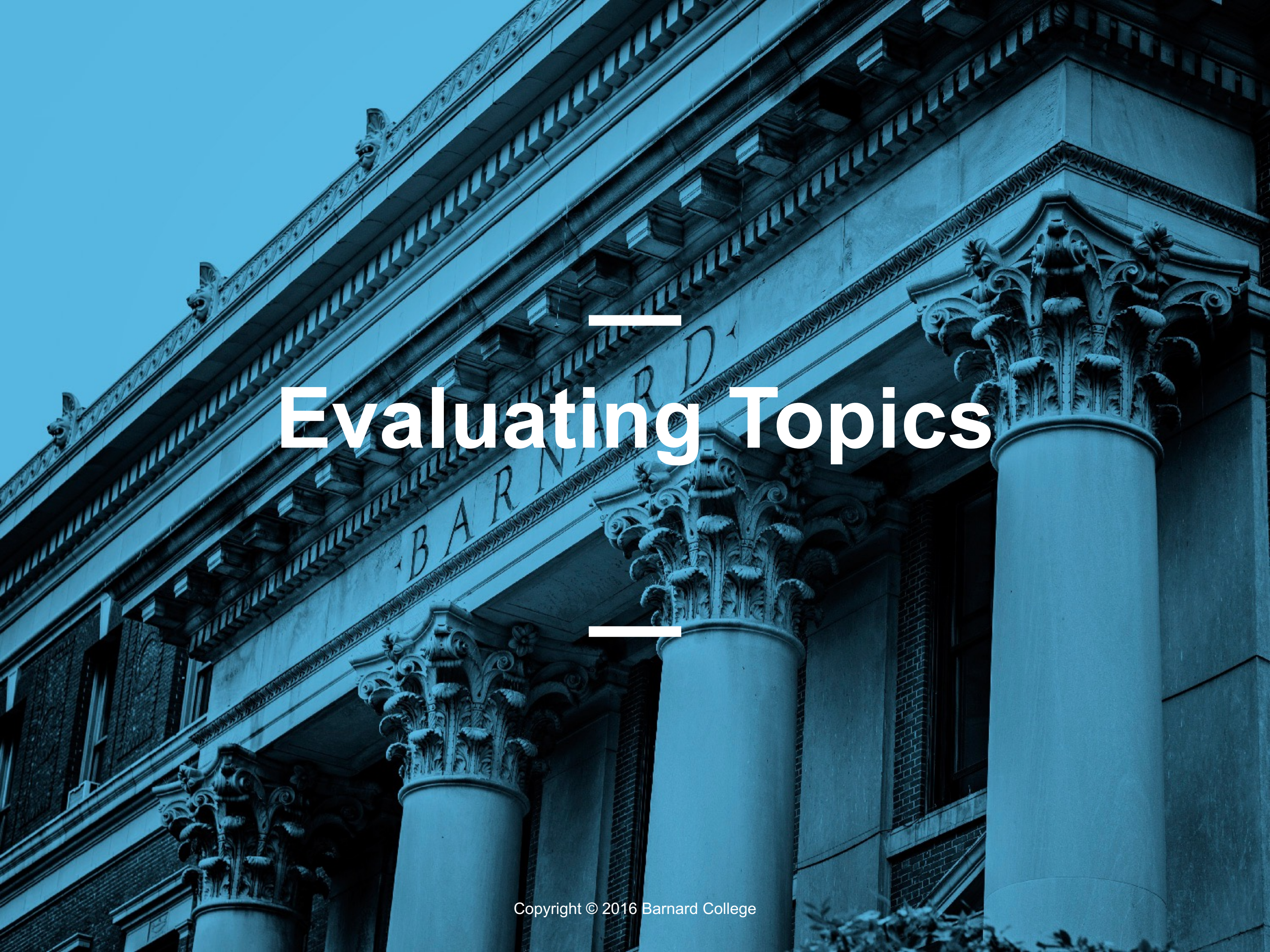
LDA

- Probabilistic model
- Generative model

LDA Generative Story

- Each word appears independent of each other

- Each word depends on the topic
  - Topics have a distribution of words
  - Topics have a distribution of documents
  - Both are multinomial distributions!



Evaluating Topics

Output of topic models

Top 10 topic terms
face, problem, depress, econom, suffer, economi, caus, great depress, crisi, prosper
bank, money, tax, pay, debt, loan, rais, fund, paid, govern
worker, labor, work, union, job, employ, strike, factori, industri, wage
govern, power, feder, nation, peopl, author, constitut, state, system, unit
roosevelt, wilson, peac, presid, treati, negoti, theodor roosevelt, taft, leagu, agreement
men, women, famili, children, young, work, woman, home, mother, husband
citi, york, urban, hous, live, town, center, communiti, move, chicago
railroad, build, line, technolog, transport, road, develop, travel, invent, canal
good, trade, product, manufactur, market, import, produc, economi, consum, tariff
farmer, farm, planter, small, land, cotton, plantat, crop, famili, larg

What makes topics bad?

- **Random**, unrelated words
- ***Intruder*** words
- Boring, **overly general** words
- **Chimaeras:**
  - Multiple topics combined

Evaluation – Word Intrusion Task

- Take top k words in a topic
  - Usually 5 or 10
- Substitute 1 word with a top word from another topic
- Shuffle the works
- Ask someone to pick the intruder
  - If they can pick the intruder – it's a good topic

Automatic Metrics – Topic Coherence

- The average or median of pairwise word similarities formed by top words of a given topic.

music	book	art	game	show
band	life	museum	knicks	film
songs	novel	show	nets	television
rock	story	exhibition	points	movie
album	books	artist	team	series
jazz	man	artists	season	says
pop	stories	paintings	play	life
song	love	painting	games	man
singer	children	century	night	character
night	family	works	coach	know

theater	clinton	stock	restaurant	budget
play	bush	market	sauce	tax
production	campaign	percent	menu	governor
show	gore	fund	food	county
stage	political	investors	dishes	mayor
street	republican	funds	street	billion
broadway	dole	companies	dining	taxes
director	presidential	stocks	dinner	plan
musical	senator	investment	chicken	legislature
directed	house	trading	served	fiscal

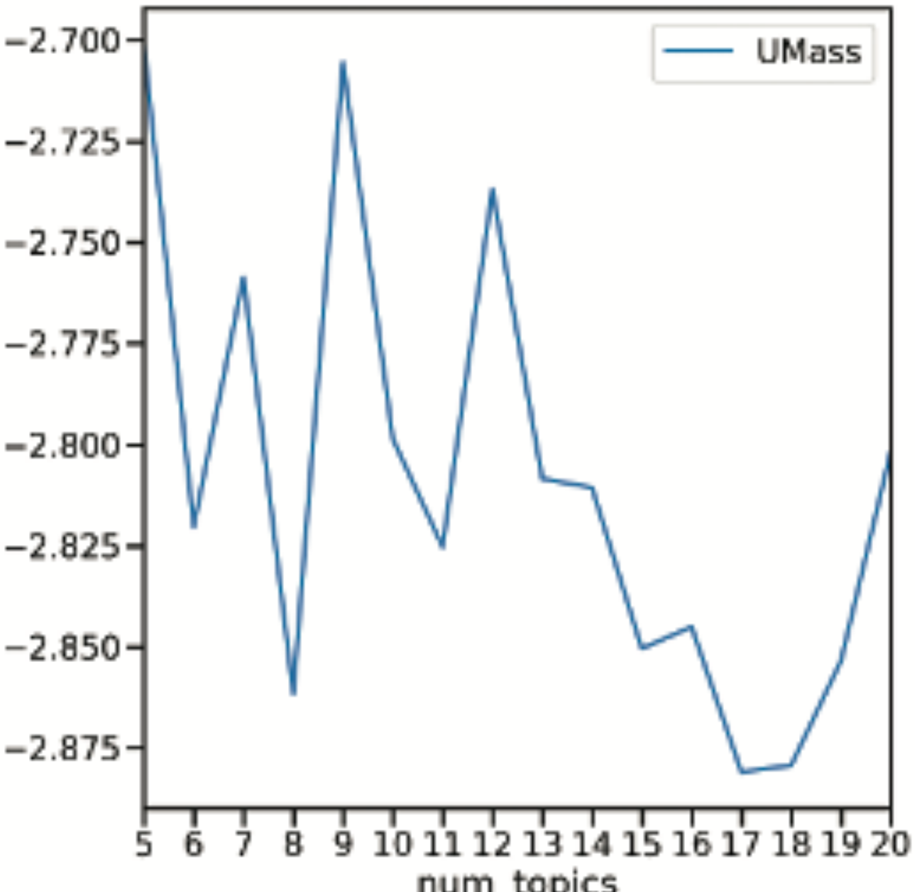
Automatic Metrics – Topic Coherence

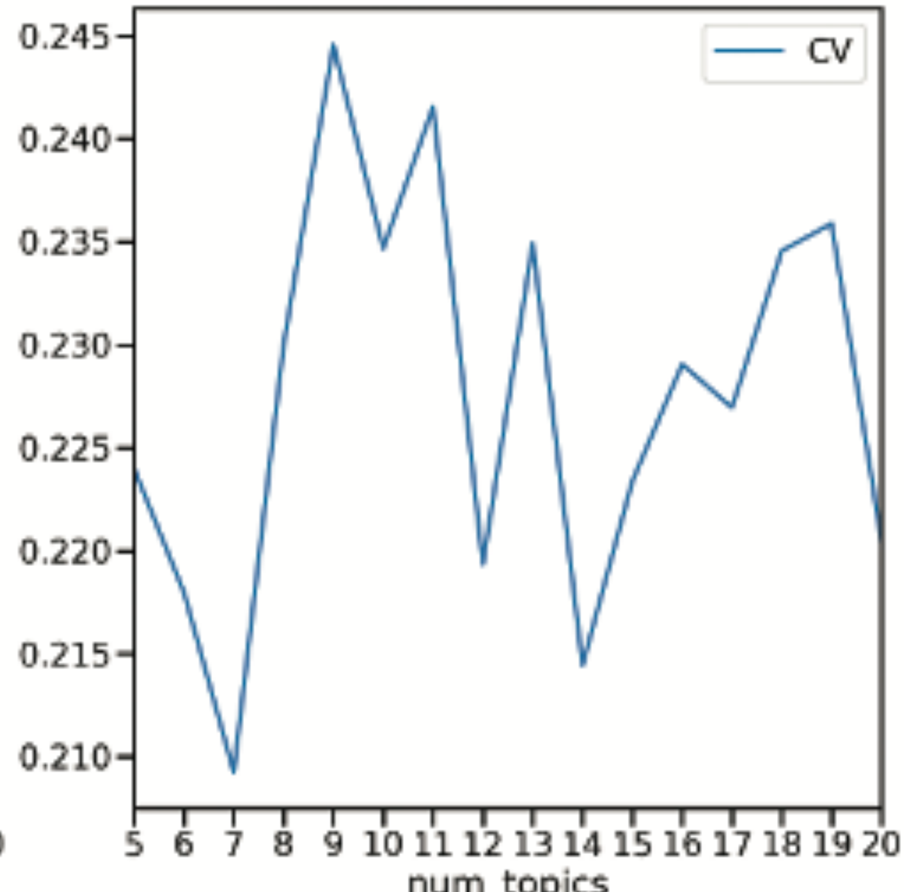
- The average or median of pairwise word similarities formed by top words of a given topic.

- Pairwise word similarities:
  - Umass Coherence:
    - log probability of word co-occurrences of topic words
  - UCI Coherence:
    - normalized pointwise mutual information of topics words

- Further reading:
  - Evaluating topic coherence measures - https://arxiv.org/pdf/1403.6397.pdf

Using Topic Coherence to choose *k*





LDA Popularity

- Straight-forward modeling approach
- Lots of easy-to-use implementations

Mallet

Mallet

- MAchine Learning for LanguagE Toolkit

- Java-based library for Natural Language Processing
  - Started at Umass by Andrew McCallum and his students
    - http://mallet.cs.umass.edu/
  - Currently maintained by David Mimno (Cornell) and his students
    - Public code: https://github.com/mimno/Mallet

Little Mallet Wrapper – Mallet in

Maria Antoniak
@maria_antoniak

If you want to call MALLET from Python, here's my little-mallet-wrapper!

It's pretty simple but also includes some plotting functions. Should be useful if you have students who are afraid of the command line or if you just don't feel like leaving the comfort of Jupyter.

> **Melanie Walsh** @mellymeldubs · Dec 15, 2020
> Replying to @pvierth @maria_antoniak and @heatherfro
> Maria also developed a Python wrapper for MALLET! github.com/maria-antoniak... I taught it in my undergrad class last semester, and I thought it was really successful

10:50 AM · Dec 15, 2020 · Twitter Web App

Outline

Topic Modeling

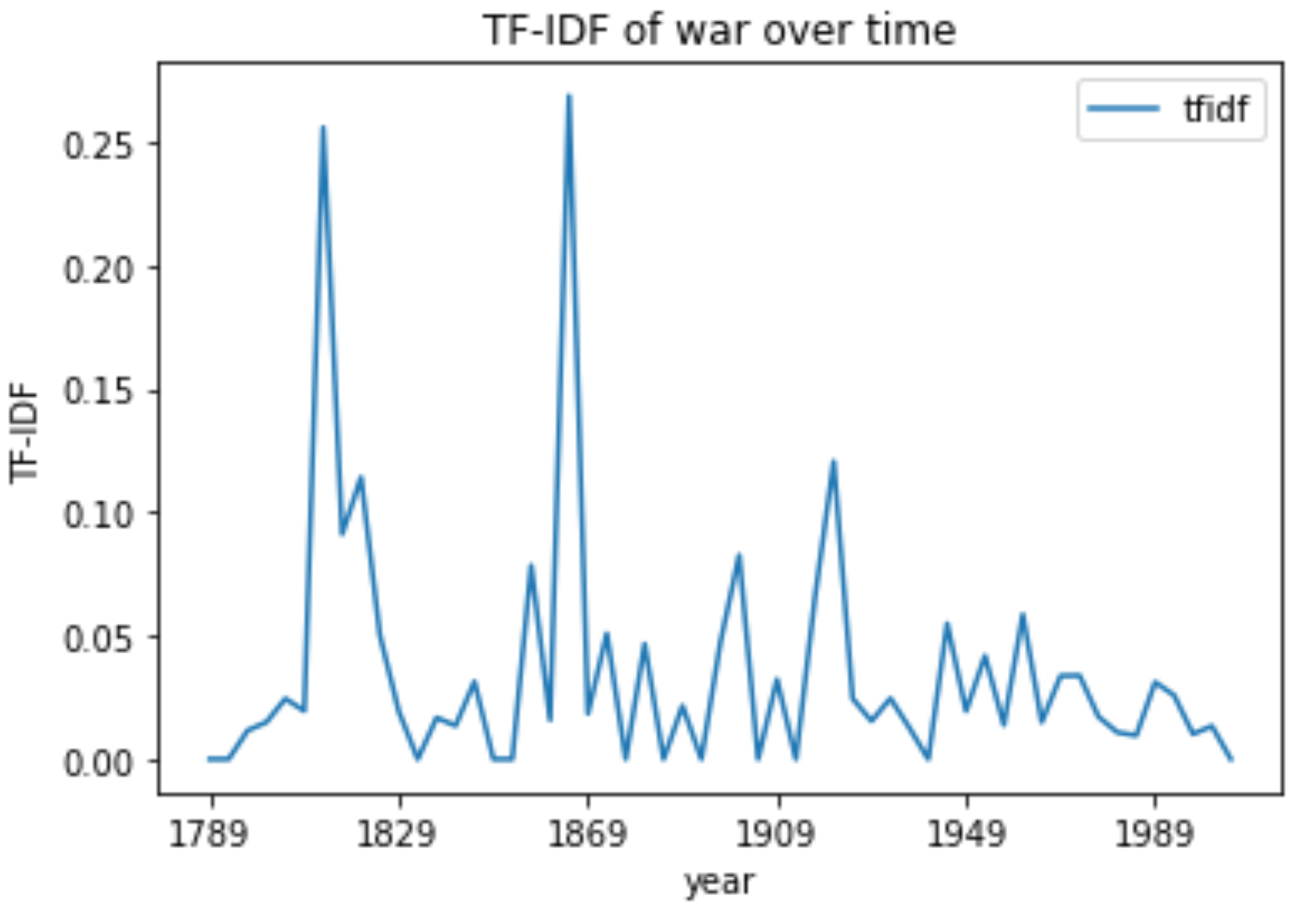
- Topic Model Review
- Evaluating Topic Models
- Implementation

Dictionary based methods

PMI

What did we count so far?

Words, words, words



Dictionary based methods

- Goal: Connect counts c_i to attributes v_i

- Dictionary-based methods:
  - Specify $\hat{v}_i = f(c_i)$ for some known function $f(\cdot)$
  - Define $f(\cdot)$ based on a prespecified dictionary of terms capturing particular categories of text
  - Common method in the social science literature using text
    - Appropriate in cases where prior information is strong

Text as Data, Gentzkow, Kelly, and Taddy
Journal of Economic Literature 2019

Connotations in the Vocabulary

- Words have connotations

- Goal of Dictionaries:
  - Build lexical recourses the represent word connotations

- Dictionary-based methods:
  - Deploy connotation-dictionaries to detect and categorize text

Sentiment Lexicons

Affective Meaning

- Affective: relating to moods, feelings, and attitudes

- Drawing on literatures in
  - affective computing – Rosalind Picard
    - ***if we want computers to be genuinely intelligent and to interact naturally with us, we must give computers the ability to recognize, understand, even to have and express emotions.***

Affective Meaning

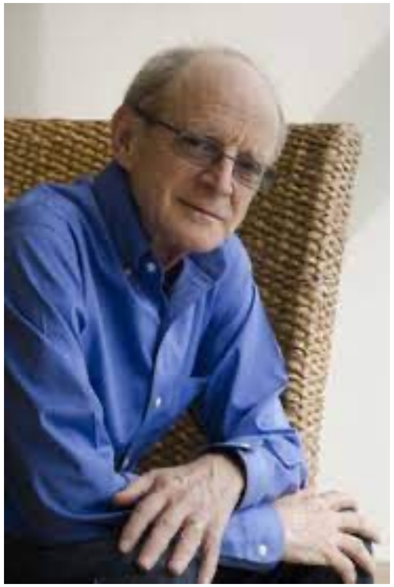
- Affective: relating to moods, feelings, and attitudes

- Drawing on literatures in
  - Linguistic subjectivity – Janyce Wiebe
    - ***Subjectivity in natural language refers to aspects of language used to express opinions, evaluations, and speculatins.***

Affective Meaning

- Affective: relating to moods, feelings, and attitudes

- Drawing on literatures in
  - Social Psychology – James Pennebaker
  - Developed LIWIC:
    - **a program that simply looked for and counted words in psychology-relevant categories across multiple text files.**



Affective Meaning

- Can we identify:
  - sentiment
  - emotion
  - personality
  - mood
  - attitudes

Slide take from Dan Jurafsky

Why Compute Affective Meaning?

- Detection and Categorization
  - sentiment towards politicians, products, countries, ideas
  - frustration of callers to a help line
  - stress in drivers or pilots
  - depression and other medical conditions
  - confusion in students talking to e-tutors
  - emotions in novels (e.g., for studying groups that are feared over time)

Slide take from Dan Jurafsky

Dictionaries of Attitudes

The General Inquirer

Philip J. Stone, Dexter C Dunphy, Marshall S. Smith, Daniel M. Ogilvie. 1966. The General Inquirer: A Computer Approach to Content Analysis. MIT Press

- Home page: http://www.wjh.harvard.edu/~inquirer
- List of Categories: http://www.wjh.harvard.edu/~inquirer/homecat.htm
- Spreadsheet: http://www.wjh.harvard.edu/~inquirer/inquirerbasic.xls
- Categories:
  - Positive (1915 words) and Negative (2291 words)
  - *Strong vs Weak, Active vs Passive, Overstated versus Understated*
  - *Pleasure, Pain, Virtue, Vice, Motivation, Cognitive Orientation,*
- Free for Research Use

Slide taken from Dan Jurafsky

MPQA Subjectivity Cues Lexicon

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann (2005). Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. Proc. of HLT-EMNLP-2005.

Riloff and Wiebe (2003). Learning extraction patterns for subjective expressions. EMNLP-2003.

- Home page: http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/
- 6885 words from 8221 lemmas
  - 2718 positive
  - 4912 negative
- Each word annotated for intensity (strong, weak)
- GNU GPL

LIWC (Linguistic Inquiry and Word Count)

Pennebaker, J.W., Booth, R.J., & Francis, M.E. (2007). Linguistic Inquiry and Word Count: LIWC 2007. Austin, TX

- Home page: http://www.liwc.net/
- 2300 words, >70 classes
- **Affective Processes**
  - negative emotion (*bad, weird, hate, problem, tough*)
  - positive emotion (*love, nice, sweet*)
- **Cognitive Processes**
  - Tentative (*maybe, perhaps, guess*), Inhibition (*block, constraint*)
- **Pronouns, Negation** (*no, never*), **Quantifiers** (*few, many*)
- $30 or $90 fee

Slide taken from Dan Jurafsky

LIWC Categories

LIWC			LIWC Cont.		
Category	Example	T-statistics	Category	Example	T-statistics
Linguistics Processes			Negative emotion	hurt, ugly, nasty	6.49***
Words > 6 letters		-3.41**	Anxiety	fearful, nervous	2.37
Dictionary words		9.60****	Anger	hate, kill, annoy	5.30***
Total function words		8.98****	Sadness	cry, grief, sad	3.54***
Personal pron.	I, them, her	7.07****	Cognitive process	cause, ought	6.09***
1st pers singular	I, me, mine	9.83****	Insight	think, know	0.11
1st pers plural	we, us, our	-2.38	Causation	effect, hence	0.93
2nd person	you, your, thou	-0.91	Discrepancy	should, would	5.53***
3rd pers singular	she, her, him	3.63**	Tentative	maybe, perhaps	5.95***
3rd pers plural	their, they'd	2.47	Certainty	always, never	4.02***
Impersonal pron.	it, it's, those	7.07****	Inhibition	block, constrain	0.32
Articles	a, an, the	4.13***	Inclusive	with, include	4.74 ***
Common verbs	walk, went, see	6.27***	Exclusive	but, without	7.53 ****
Auxiliary verbs	am, will, have	5.76***	Perceptual process		1.93
Past tense	went, ran, had	8.70****	See	view, saw, seen	1.68
Present tense	is, does, hear	4.00***	Hear	listen, hearing	-0.88
Future tense	will, gonna	5.84***	Feel	feels, touch	1.94
Adverbs	very, really	7.92****	Biological process		4.22***
Prepositions	to, with, above	7.62****	Body	cheek, spit	5.02***
Conjunctions	and, whereas	4.59***	Health	clinic, flu, pill	1.51
Negations	no, not, never	1.71	Sexual	horny, incest	-0.61
Quantifiers	few, many, much	2.98*	Ingestion	dish, eat, pizza	4.37***
Numbers	second, thousand	-3.68**	Relativity	area, bend, exit	9.52 ****
Swear words	damn, piss, fuck	5.53***	Motion	arrive, car	3.07*
Spoken Categories			Space	down, in, thin	8.87****
Assent	agree, OK, yes	7.05****	Time	end, until	5.87***
Nonfluency	er, hm, umm	1.41	**Personal Concerns**		
Filters	blah, imean		Work	job, majors	0.05
Psychological			Leisure	chat, movie	2.97*
Social process	mate, talk, child	0.10	Achievement	earn, win	-1.22
Family	son, mom, aunt	2.24	Home	family, kitchen	3.37**
Friends	buddy, neighbor	2.10	Money	audit, cash	0.23
Humans	adult, baby, boy	0.89	Religion	church, altar	-0.77
Affective process	happy, cry	3.55**	Death	bury, coffin	0.49
Positive emotion	love, nice, sweet	0.08			

Table 1. Two-sample T-test statistics of linguistic variables between geo-locator and non-locators. Significant differences of each LIWC attribute are indicated in the third column. (*p <0.01, **p<0.001, ***p<0.0001, ****p<1e-10)

Plenty of more examples

- See Chris Pott's Tutorial on Sentiment Lexicons
  - http://sentiment.christopherpotts.net/lexicons.html
  - Compares different dictionaries of sentiment

Discovering Connotations

Point-wise Mutual Information

- How much more do events ***x*** and ***y*** co-occur than if they were independent?

$$PMI(X,Y) \; = \; \log_2 \frac{P(X,Y)}{P(X)P(Y)}$$

Point-wise Mutual Information

$$PMI(X,Y) \quad = \quad \log_2 \frac{P(X,Y)}{P(X)P(Y)}$$

- PMI between words and categories:

$$PMI\left(word_i, category_j\right) \\ = \qquad \log_2 \frac{P(word_i, category_j)}{P(word_i)P(category_j)}$$

Discovering Gender Bias via PMI

First Women, Second Sex: Gender Bias in Wikipedia

Eduardo Graells-Garrido
Universitat Pompeu Fabra
Barcelona, Spain
eduard.graells@upf.edu

Mounia Lalmas
Yahoo Labs
London, UK
mounia@acm.org

Filippo Menczer
Yahoo Labs
Sunnyvale, USA

We approach the analysis of gender bias by defining a methodology for comparing the characterizations of men and women in biographies. In particular we refer to three dimensions of biographies: meta-data, language usage, and structure of the network built from links between articles. Our results show that, indeed, there are differences in characterization and structure.

PMI to measure gender bias

Associativity. To explore which words are more strongly associated with the different genders, we measure *Pointwise Mutual Information* (Church and Hanks, 1990) over the set of vocabulary in both genders. PMI is defined as:

$$\mathrm{PMI}(c, w) = \log \frac{p(c, w)}{p(c)p(w)}$$

where c is a class (*men* or *women*), and w is a word. The probabilities can be estimated from the proportions of biographies about men and women, and the corresponding proportions of words and bigrams. Since PMI overweights words with very small frequencies, we consider only words that appear in at least 1% of men or women biographies.

https://arxiv.org/abs/1502.02341

Next week

Supervised learning

Math concepts:

- Partial derivatives
- Optimization

Readings will include examples of dictionary methods