THE UNIVERSITY of EDINBURGH

Text Technologies for Data Science

INFR11145

Comparing Text Corpora (2)

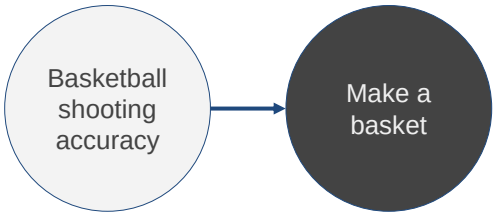
Instructor:
Steve Wilson

11-Nov-2020

1

LDA Overview

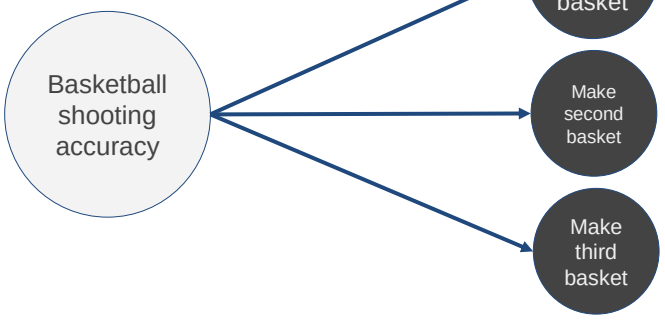
Background: Plate Notation



3

3

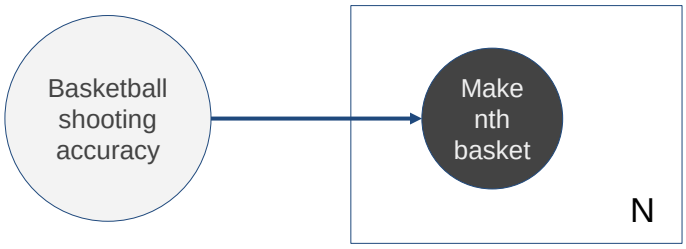
Background: Plate Notation



4

4

Background: Plate Notation

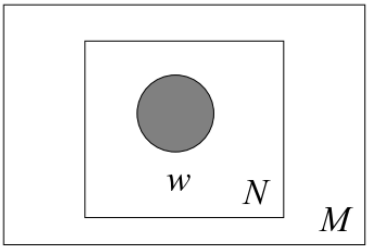


Latent Dirichlet Allocation

- Let's start with a very simple model
- We will work our way up to the full LDA model

Unigram Model

w is a word
N words in a document
M documents in a corpus
w is a vector of words (i.e. doc)



$$p(\mathbf{w}) = \prod_{n=1}^{N} p(w_n).$$

Figure from
Blei et al 2003

Probability with a Unigram Model

$$p(\mathbf{w}) = \prod_{n=1}^{N} p(w_n).$$

Compute the probability of the example sentence.

"My dog barked at another dog."

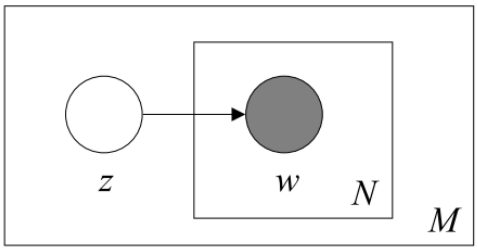
word	my	at	dog	another	barked
probability	.1	.1	.05	.04	.03

Unigram Model…

- What is the point of making these models more complex?
- Why not just use the basic unigram model for everything?
- Remember:
  - Higher text probability ***doesn't imply a better model***
  - We want to ***accurately describe*** the data
    - → higher probability for *real* documents, lower probability for noise

Mixture of Unigrams Model

z is the topic of a document



$$p(\mathbf{w}) = \sum_{z} p(z) \prod_{n=1}^{N} p(w_n \mid z).$$

Figure from Blei et al 2003

10

Probability with Mixture of Unigrams

$$p(\mathbf{w}) = \sum_z p(z) \prod_{n=1}^{N} p(w_n \mid z).$$

p(z=pets) = .6, p(z=vehicles) =.4

- Compute the probability of the sentence.
- Ignore stopwords: “my”, “after”, “the”

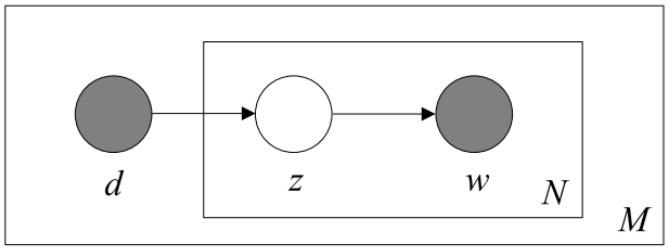
“My dog chased after the bus.”

word	cat	dog	chased	car	bus
P(w | z=pets)	.2	.3	.1	.01	.01
P(w | z=vehicles)	.01	.01	.1	.3	.2

11

Probabilistic Latent Semantic Indexing

d is a document ID



$$P(d,w) = \sum_{z \in Z} P(z)P(w \mid z)P(d \mid z)$$

Figure from Blei et al 2003

12

Probability with pLSI

$$P(d,w) = \sum_{z \in Z} P(z)P(w \mid z)P(d \mid z)$$

p(z=t1)	.5
p(z=t2)	.5
p(d | z=t1)	.6
p(d | z=t2)	.4

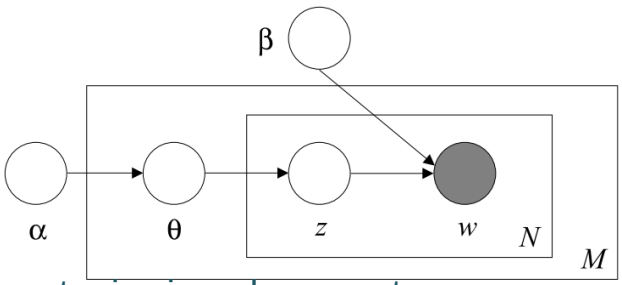
"The cat sat down."

Stopword = "The"

word	cat	sat	down	car	broke
p(w | z=t1)	.2	.1	.05	.01	.1
p(w | z=t2)	.01	.05	.1	.3	.1

13

Latent Dirichlet Allocation



θ is the distribution over topics in a document

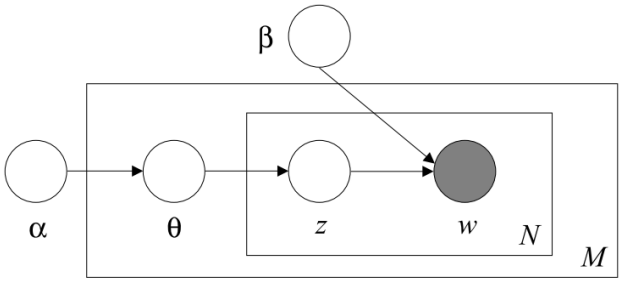
α is a prior over possible topic distributions within documents

β is a prior over word distributions within topics

Figure from Blei et al 2003

14

Latent Dirichlet Allocation



$$p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta),$$

Figure from Blei et al 2003

Probability with LDA

$$p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta),$$

topic	t1	t2
p(z=topic | θ)	.6	.4

"Fish swam by a submerged submarine."

Stopwords = ["a", "by"] **z** = [t1, t1, t2, t2] p(θ|α) = .7

word	fish	swam	submerged	submarine
p(w | z=t1, β)	.2	.1	.001	.05
p(w | z=t2, β)	.01	.02	.1	.3

Latent Dirichlet Allocation

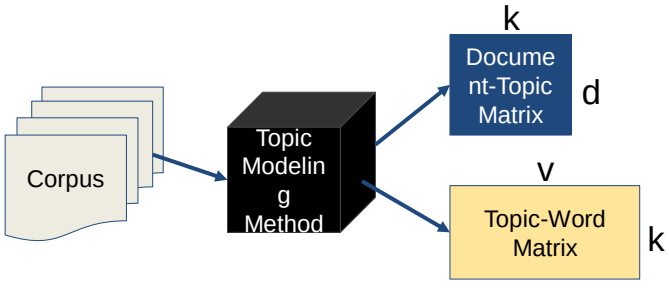
$$p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta),$$

$$p(\mathbf{w} \mid \alpha, \beta) = \int p(\theta \mid \alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n \mid \theta) p(w_n \mid z_n, \beta) \right) d\theta.$$

$$p(D \mid \alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d \mid \alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn} \mid \theta_d) p(w_{dn} \mid z_{dn}, \beta) \right) d\theta_d.$$

Model Inference

- Want to learn the model parameters
- Exact inference becomes intractable



Model Inference

- Instead, use an approximate method such as:
  - Gibbs sampling
  - Variational Inference

19

Gibbs Sampling for LDA

Want to learn Φ, θ given a set of documents D
Φ = topic-word probabilities
θ = document-topic probabilities

20

Gibbs Sampling for LDA

Want to learn Φ, θ given a set of documents D

1. Randomly initialize Φ, θ
2. Repeat until convergence:
   a. Sample a new topic assignment for every word in every document
   b. Use newly sampled topics to update Φ and θ

Gibbs Sampling for LDA

Green eggs and ham.
Ham and green peppers.
Ham and cheese.

	green	eggs	and	ham	peppers	cheese
t1	.1	.4	.05	.1	.05	.3
t2	.05	.15	.1	.2	.4	.1

	s1	s2	s3
t1	.5	.2	.4
t2	.5	.8	.6

Random initialization.

Gibbs Sampling for LDA

Green eggs and ham.
Ham and green peppers.
Ham and cheese.

	green	eggs	and	ham	peppers	cheese
t1	.1	.4	.05	.1	.05	.3
t2	.05	.15	.1	.2	.4	.1

	s1	s2	s3
t1	.5	.2	.4
t2	.5	.8	.6

Green eggs and ham.
Ham and green peppers.
Ham and cheese.

23

Gibbs Sampling for LDA

Green eggs and ham.
Ham and green peppers.
Ham and cheese.

	green	eggs	and	ham	peppers	cheese
t1	1/6	1/6	1/6	1/6	1/6	1/6
t2	1/5	0/5	2/5	2/5	0/5	0/5

Green eggs and ham.
Ham and green peppers.
Ham and cheese.

	s1	s2	s3
t1	2/4	2/4	2/3
t2	2/4	2/4	1/3

24

Gibbs Sampling for LDA

[Repeat until convergence or max iterations]

25

Topic Modeling Examples

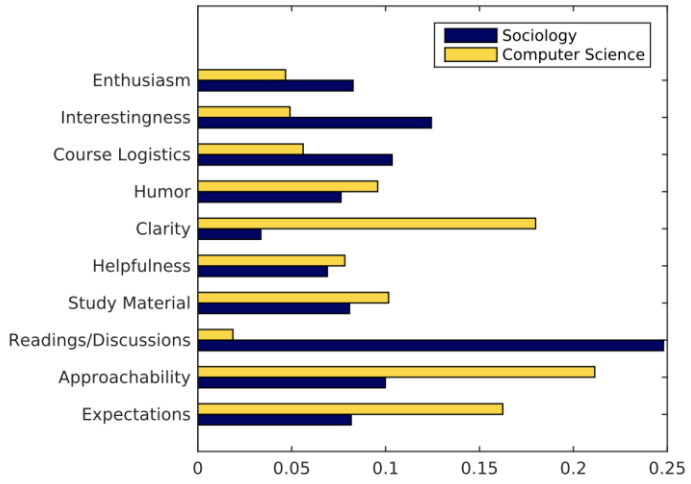
26

What do students look for in a professor?

Topic	Sample words
Approachability	prof, fair, clear, helpful, teaching, approachable, nice, organized, extremely, friendly, super, amazing
Clarity	understand, hard, homework, office, material, clear, helpful, problems, explains, accent, questions, extremely
Course Logistics	book, study, boring, extra, nice, credit, lot, hard, attendance, make, fine, attention, pay, mandatory
Enthusiasm	teaching, passionate, awesome, enthusiastic, professors, loves, cares, wonderful, fantastic, passion
Expectations	hard, work, time, lot, comments, tough, expects, worst, stuff, avoid, horrible, classes
Helpfulness	helpful, nice, recommend, cares, super, understanding, kind, extremely, effort, sweet, friendly, approachable
Humor	guy, funny, fun, awesome, cool, entertaining, humor, hilarious, jokes, stories, love, hot, enjoyable
Interestingness	interesting, material, recommend, lecturer, engaging, classes, knowledgeable, enjoyed, loved, topics
Readings/ Discussions	readings, papers, writing, ta, interesting, discussions, grader, essays, boring, books, participation
Study Material	exams, notes, questions, material, textbook, hard, slides, study, answer, clear, tricky, attend, bring, understand

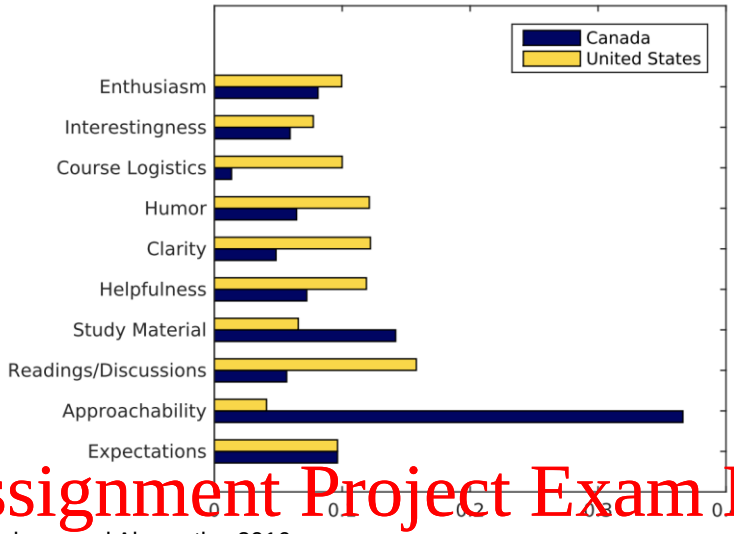
Azab, Mihalcea, and Abernathy, 2016

What do students look for in a professor?



Azab, Mihalcea, and Abernathy, 2016

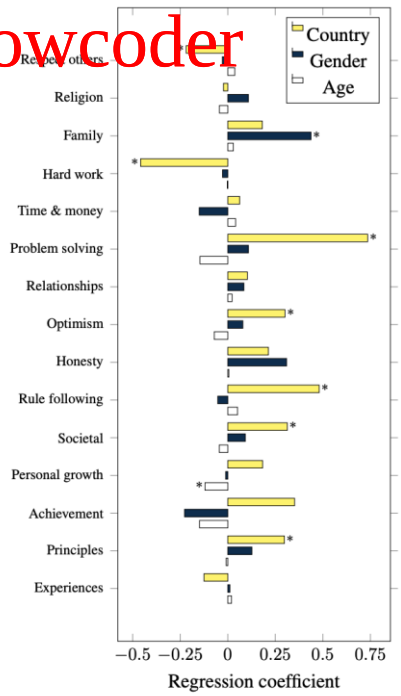
What do students look for in a professor?



Azab, Mihalcea, and Abernathy, 2016

How do personal attributes relate to values?

Theme	Example Words
Respect others	people, respect, care, human, treat
Religion	god, heart, belief, religion, right
Family	family, parent, child, husband, mother
Hard Work	hard, work, better, honest, best
Time & Money	money, work, time, day, year
Problem solving	consider, decision, situation, problem
Relationships	family, friend, relationship, love
Optimism	enjoy, happy, positive, future, grow
Honesty	honest, truth, lie, trust, true
Rule following	moral, rule, principle, follow
Societal	society, person, feel, thought, quality
Personal Growth	personal, grow, best, decision, mind
Achievement	heart, achieve, complete, goal
Principles	important, guide, principle, central
Experiences	look, see, experience, choose, feel



Wilson, Mihalcea, Boyd, and Pennebaker 2016

Annotation + Classification

Annotation + Classification

- Method 1: Traditional Supervised Learning
  - Annotate representative samples
  - Train a classifier
  - Apply to rest of data
- Method 2: Transfer Learning
  - Find another large, but similar dataset
  - Train a classifier on that dataset
  - *Optionally: fine-tune classifier to your smaller dataset*
  - Apply to rest of your data

After Classification

- Which features are most relevant for each class?
- What are common words/topics for each class?
- How do predicted classes relate to other variables?

- *More about text classification coming up next week!*

Wrap-up

- Content analysis background
- Word-level differences
- Dictionaries and Lexica
- Topic modeling
- Annotation + classification

Readings

- Manning: IR book section 13.5
- "Probabilistic Topic Models" by David Blei