THE UNIVERSITY of EDINBURGH

Text Technologies for Data Science

INFR11145

Comparing Text Corpora (2)

Instructor:
Björn Ross

08-Nov-2023

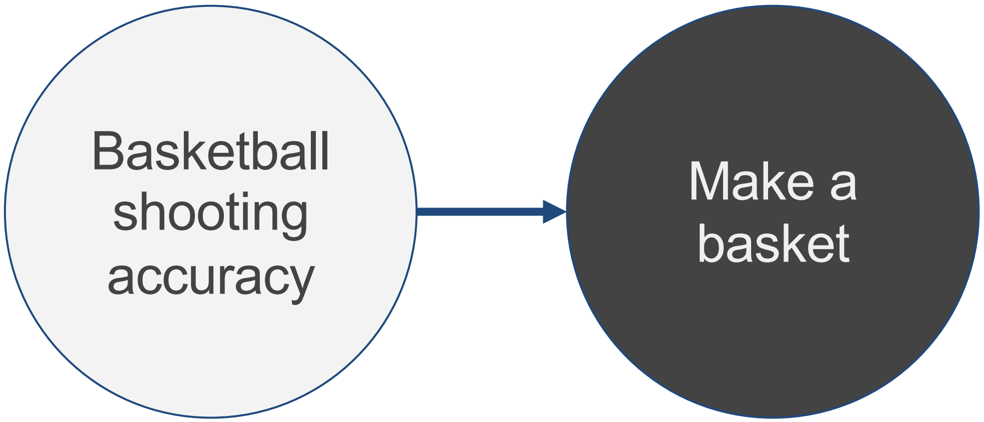
LDA Overview

Background: Plate Notation

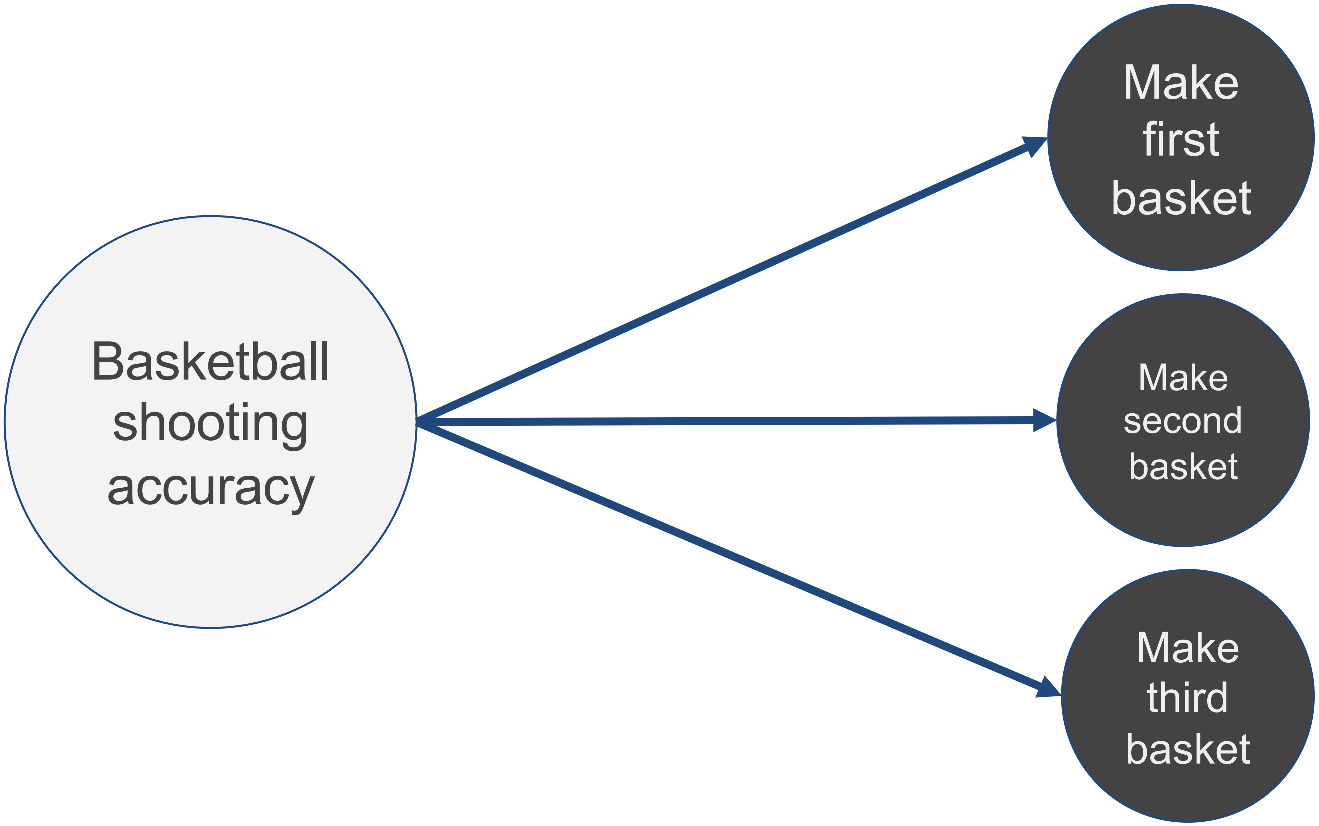
Background: Plate Notation

Make a basket

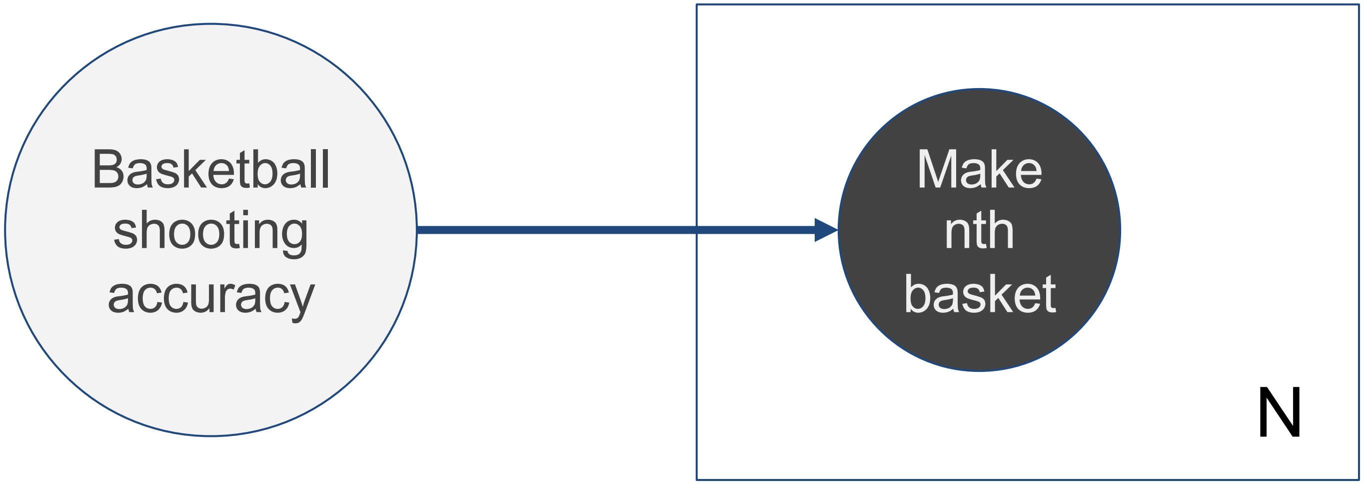
Background: Plate Notation



Background: Plate Notation



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

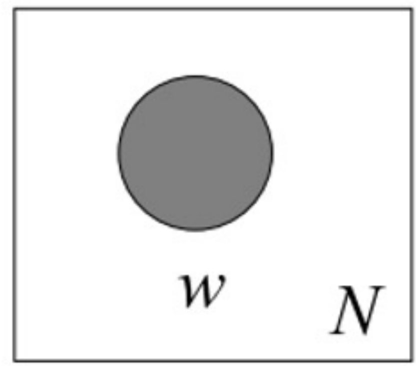


Figure from
Blei et al 2003

Unigram Model

w is a word
N words in a document
M documents in a corpus

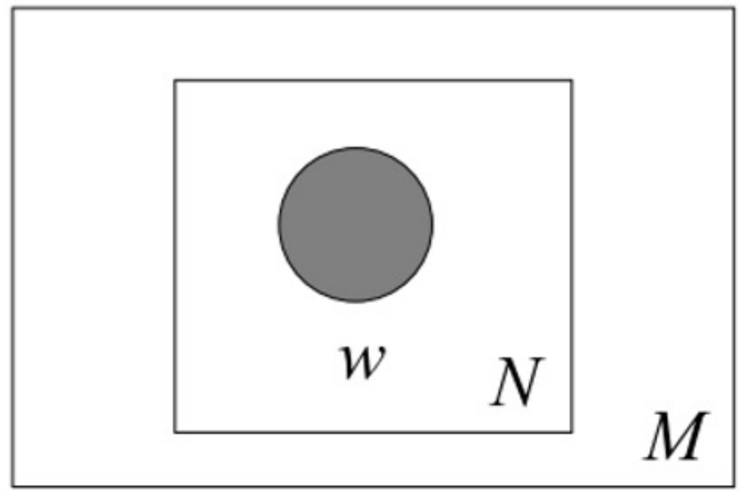


Figure from
Blei et al 2003

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)

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

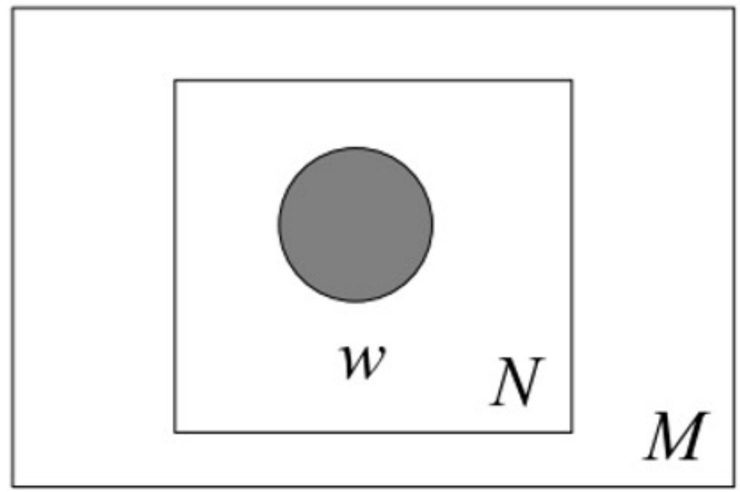


Figure from
Blei et al 2003

Probability with a Unigram Model

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

What is the probability of the example sentence?

"My dog barked at another dog."

word	my	at	dog	another	barked
probability	.10	.10	.05	.04	.03

Probability with a Unigram Model

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

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

Solution:

My dog barked at another dog.
.1 * .05 * .03 * .1 * .04 * .05 = 3e-8

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

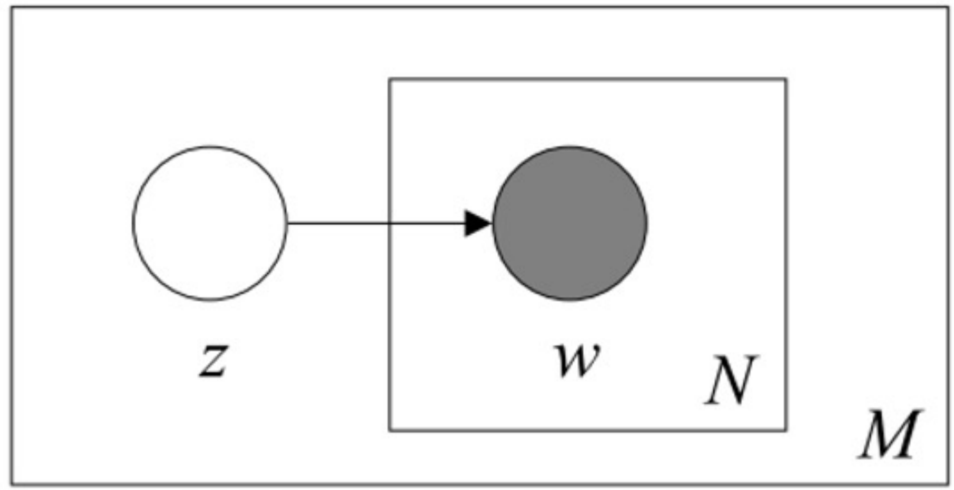
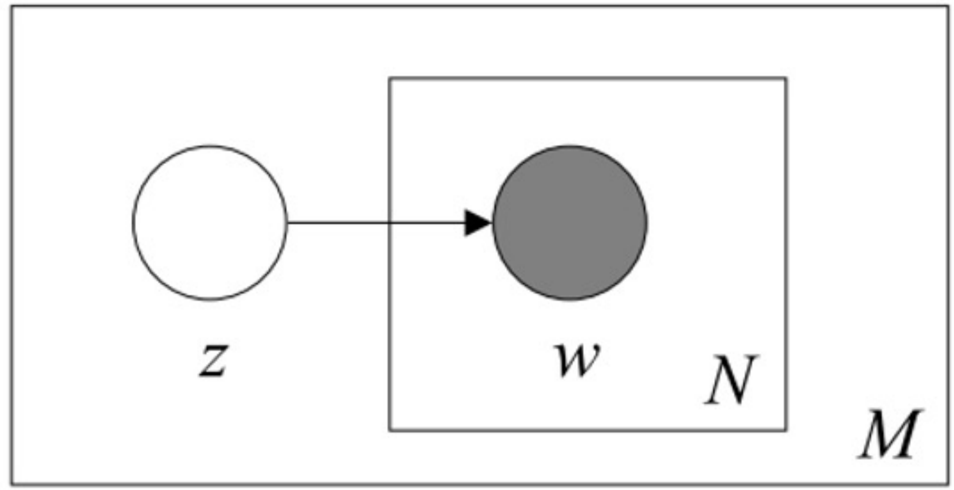


Figure from
Blei et al 2003

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

Probability with Mixture of Unigrams

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

What is the probability of the sentence?
Ignore stopwords: “my”, “after”, “the”

“My dog chased after the bus.”

w_i	cat	dog	chased	car	bus
$P(w_i \mid$ z = pets)	.20	.30	.10	.01	.01
$P(w_i \mid$ z = vehicles)	.01	.01	.10	.30	.20

$p(z = pets) = 0.6,$
$p(z = vehicles) = 0.4$

Probability with Mixture of Unigrams

word	cat	dog	chased	car	bus
$P(w_i \mid z = \text{pets})$	.2	.3	.1	.01	.01
$P(w_i \mid z = \text{vehicles})$	.01	.01	.1	.3	.2

Solution:

My dog chased after the bus.

.6(.3 * .1 * .01) = .00018

.4(.01 * .1 * .2) = .00008

Total = .00026

Probabilistic Latent Semantic Indexing

d is a document ID

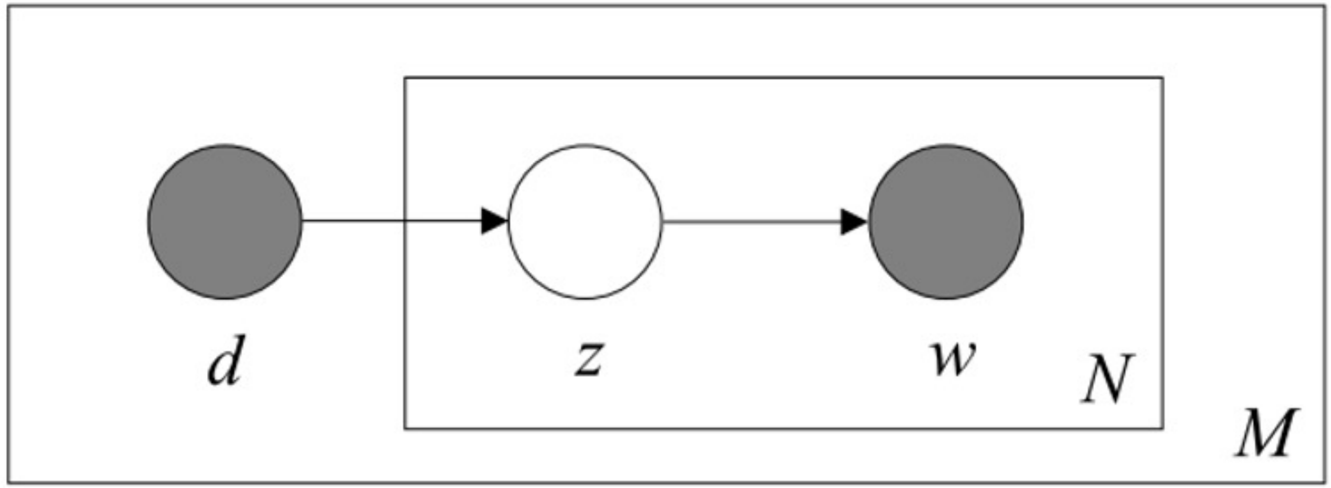
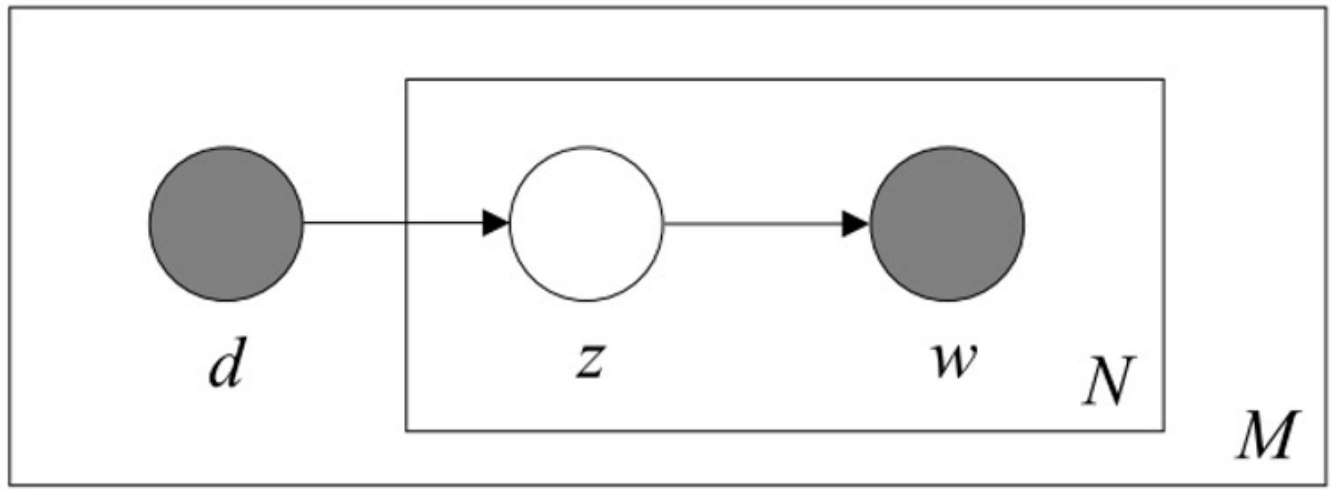


Figure from Blei et al 2003

Probabilistic Latent Semantic Indexing

d is a document ID



$$p(d, w_n) = p(d) \sum_{z} p(w_n \mid z) p(z \mid d)$$

Figure from Blei et al 2003

Probability with pLSI

w_i	cat	sat	down	car	broke
$p(w_i \vert z = t_1)$	.2	.1	.05	.01	.1
$p(w_i \vert z = t_2)$	.01	.05	.1	.3	.1

d_1 "The **cat** sat down."

$p(d = d_1)$	.01
$p(z = t_1 \vert d = d_1)$	.6
$p(z = t_2 \vert d = d_1)$	.4

w_i	cat	sat	down	car	broke
$p(w_i \vert z = t_1)$	.2	.1	.05	.01	.1
$p(w_i \vert z = t_2)$	.01	.05	.1	.3	.1

What is the joint probability of the document and the word "cat"?

Probability with pLSI

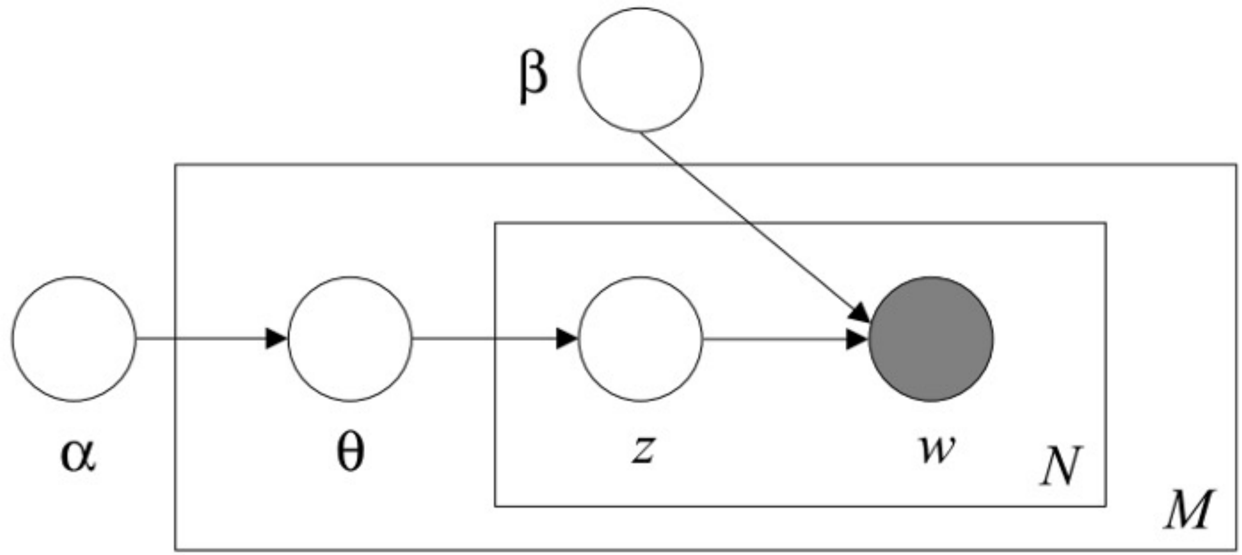
$$p(d, w_n) = p(d) \sum_z p(w_n | z) p(z | d)$$

Solution:

The **cat** sat down.

0.01 * (0.2 * 0.6 + 0.01 * 0.4) = 0.00124

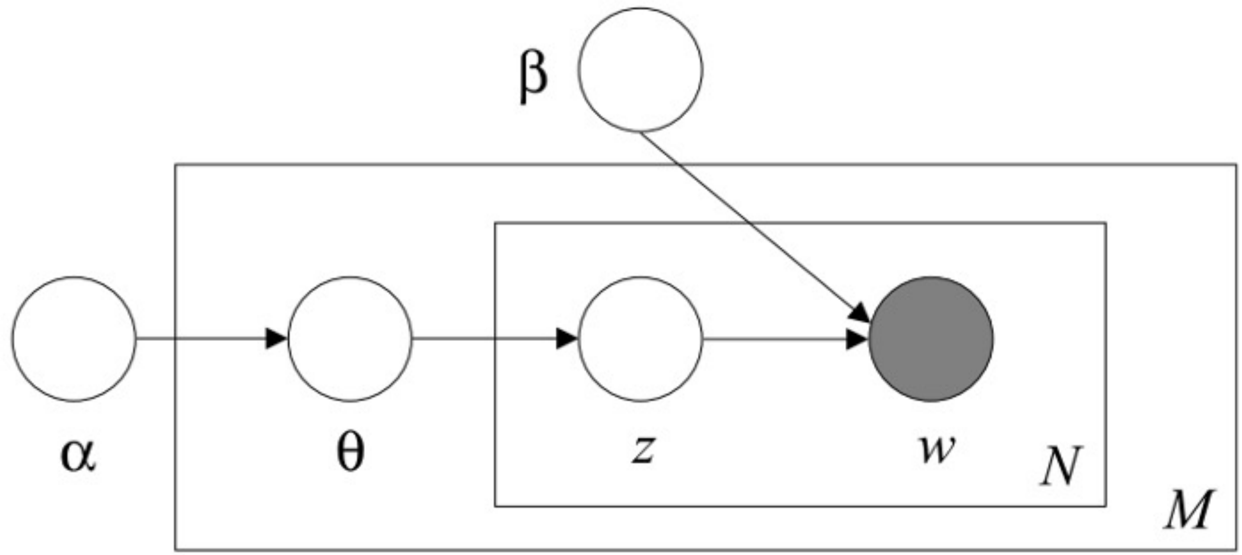
Latent Dirichlet Allocation



θ is the distribution over topics in a document
α is the parameter of a Dirichlet distribution giving possible topic distributions within documents
β gives word distributions within topics

Figure from
Blei et al 2003

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

Latent Dirichlet Allocation

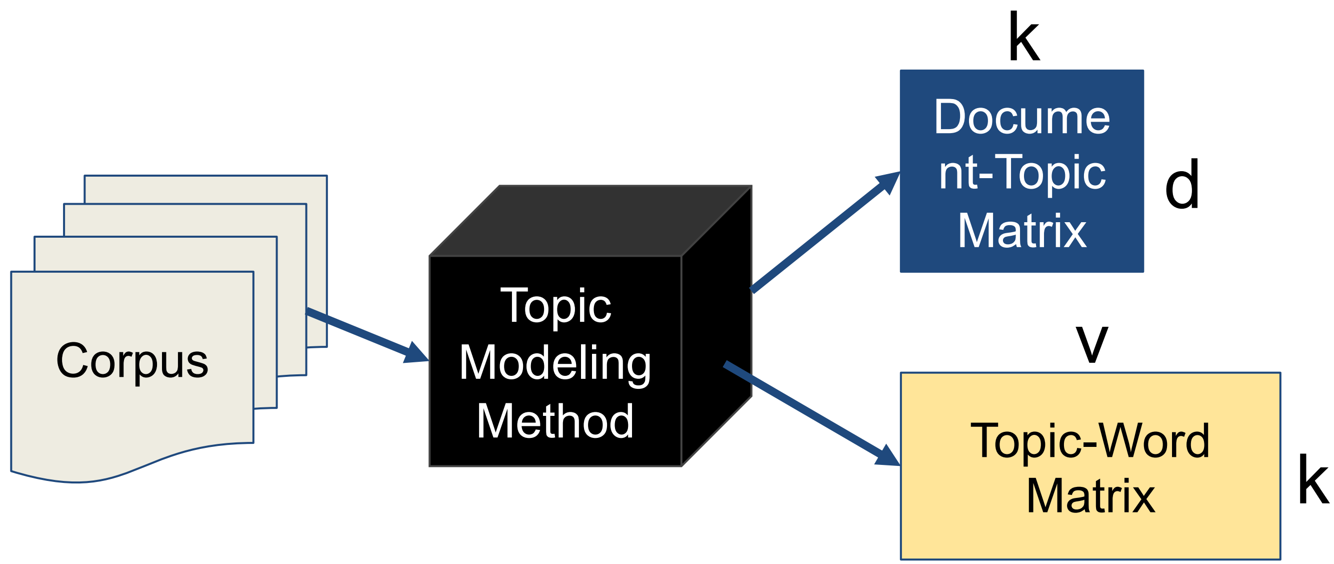
$$p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta)$$

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

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

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

Gibbs Sampling for LDA

Goal: Learn Φ, θ given a set of documents D

Φ = topic-word probabilities
θ = document-topic probabilities

Known:
corpus, α, β and the probability that a word is from a topic conditional on the assignments of all other words to topics

$$P\left(z_i = j \mid \mathbf{z}_{-i}, w_i, d_i, \cdot\right) \propto \frac{C^{WT}_{w_i j} + \beta}{\sum_{w=1}^{W} C^{WT}_{wj} + W\beta} \frac{C^{DT}_{d_i j} + \alpha}{\sum_{t=1}^{T} C^{DT}_{d_i t} + T\alpha}$$

Note: the ∝ symbol means "proportional to"

Gibbs Sampling for LDA

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

1. Assign each word a topic randomly
2. Calculate count matrices
3. Repeat until convergence:
   - For every document d
     - For every word i
       - Decrement count matrices C^{WT} and C^{DT} for current topic assignment
       - Sample a new topic assignment
       - Increment count matrices C^{WT} and C^{DT} for new topic assignment
4. Calculate Φ and θ

Gibbs Sampling for LDA

d1 Green eggs and ham.
d2 Ham and green peppers.
d3 Ham and cheese.

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

Random initialization.

Gibbs Sampling for LDA

C^{WT}	green	eggs	and	ham	peppers	cheese
t1	1	1	1	1	1	1
t2	1	0	2	2	0	0

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

C^{DT}	d1	d2	d3
t1	2	2	2
t2	2	2	1

Gibbs Sampling for LDA

Assume (for the moment) $\alpha = \beta = 0$

θ	green	eggs	and	ham	peppers	cheese
t1	0.17	0.17	0.17	0.17	0.17	0.17
t2	0.20	0.00	0.40	0.40	0.00	0.00

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

Φ	d1	d2	d3
t1	0.50	0.50	0.66
t2	0.50	0.50	0.33

Gibbs Sampling for LDA

C^{WT}	green	eggs	and	ham	peppers	cheese
t1	1	1	1	1	1	1
t2	1	0	2	2	0	0

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

C^{DT}	d1	d2	d3
t1	2	2	2
t2	2	2	1

Gibbs Sampling for LDA

C^{WT}	green	eggs	and	ham	peppers	cheese
t1	1	1	1	1	1	1
t2	1	0	2	2	0	0

$\boxed{\text{Green}}$ eggs and ham.
Ham and green peppers.
Ham and cheese.

C^{DT}	d1	d2	d3
t1	2	2	2
t2	2	2	1

Gibbs Sampling for LDA

$$\frac{C^{WT}_{w_i j} + \beta}{\sum_{w=1}^{W} C^{WT}_{wj} + W\beta} \frac{C^{DT}_{d_i j} + \alpha}{\sum_{t=1}^{T} C^{DT}_{d_i t} + T\alpha}$$

Assume (for the moment) $\alpha = \beta = 0$

C^{WT}	green	eggs	and	ham	peppers	cheese
t1	0	1	1	1	1	1
t2	1	0	2	2	0	0

$\boxed{\text{Green}}$ eggs and ham.
Ham and green peppers.
Ham and cheese.

C^{DT}	d1	d2	d3
t1	1	2	2
t2	2	2	1

Gibbs Sampling for LDA

C^{WT}	green	eggs	and	ham	peppers	cheese
t1	0	1	1	1	1	1
t2	2	0	2	2	0	0

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

C^{DT}	d1	d2	d3
t1	1	2	2
t2	3	2	1

Gibbs Sampling for LDA

C^{WT}	green	eggs	and	ham	peppers	cheese
t1	0	1	1	1	1	1
t2	2	0	2	2	0	0

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

C^{DT}	d1	d2	d3
t1	1	2	2
t2	3	2	1

Gibbs Sampling for LDA

$$\frac{C^{WT}_{w_i j} + \beta}{\sum_{w=1}^{W} C^{WT}_{wj} + W\beta} \frac{C^{DT}_{d_i j} + \alpha}{\sum_{t=1}^{T} C^{DT}_{d_i t} + T\alpha}$$

Assume (for the moment) $\alpha = \beta = 0$

C^{WT}	green	eggs	and	ham	peppers	cheese
t1	0	0	1	1	1	1
t2	2	0	2	2	0	0

Green $\boxed{\text{eggs}}$ and ham.
Ham and green peppers.
Ham and cheese.

C^{DT}	d1	d2	d3
t1	0	2	2
t2	3	2	1

Gibbs Sampling for LDA

$$\frac{C^{WT}_{w_i j} + \beta}{\sum_{w=1}^{W} C^{WT}_{wj} + W\beta} \frac{C^{DT}_{d_i j} + \alpha}{\sum_{t=1}^{T} C^{DT}_{d_i t} + T\alpha}$$

$C^{WT} + \alpha$	green	eggs	and	ham	peppers	cheese
t1	0.01	0.01	1.01	1.01	1.01	1.01
t2	2.01	0.01	2.01	2.01	0.01	0.01

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

$C^{DT} + \beta$	d1	d2	d3
t1	0.01	2.01	2.01
t2	3.01	2.01	1.01

Gibbs Sampling for LDA

- Repeat until convergence
- Probabilistic algorithm – results depend on random initialisation and random samples!

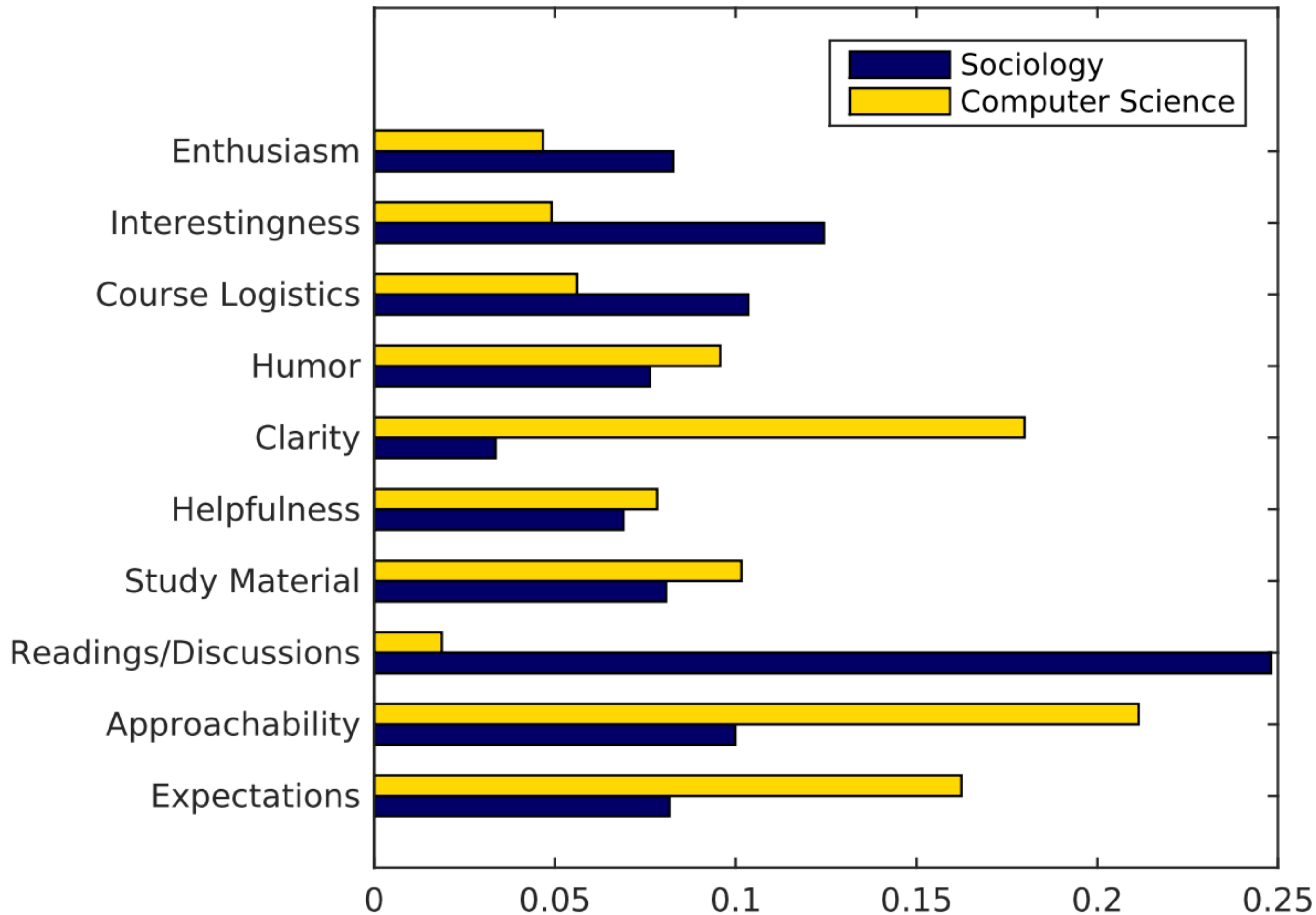
Topic Modeling Examples

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, long, understand

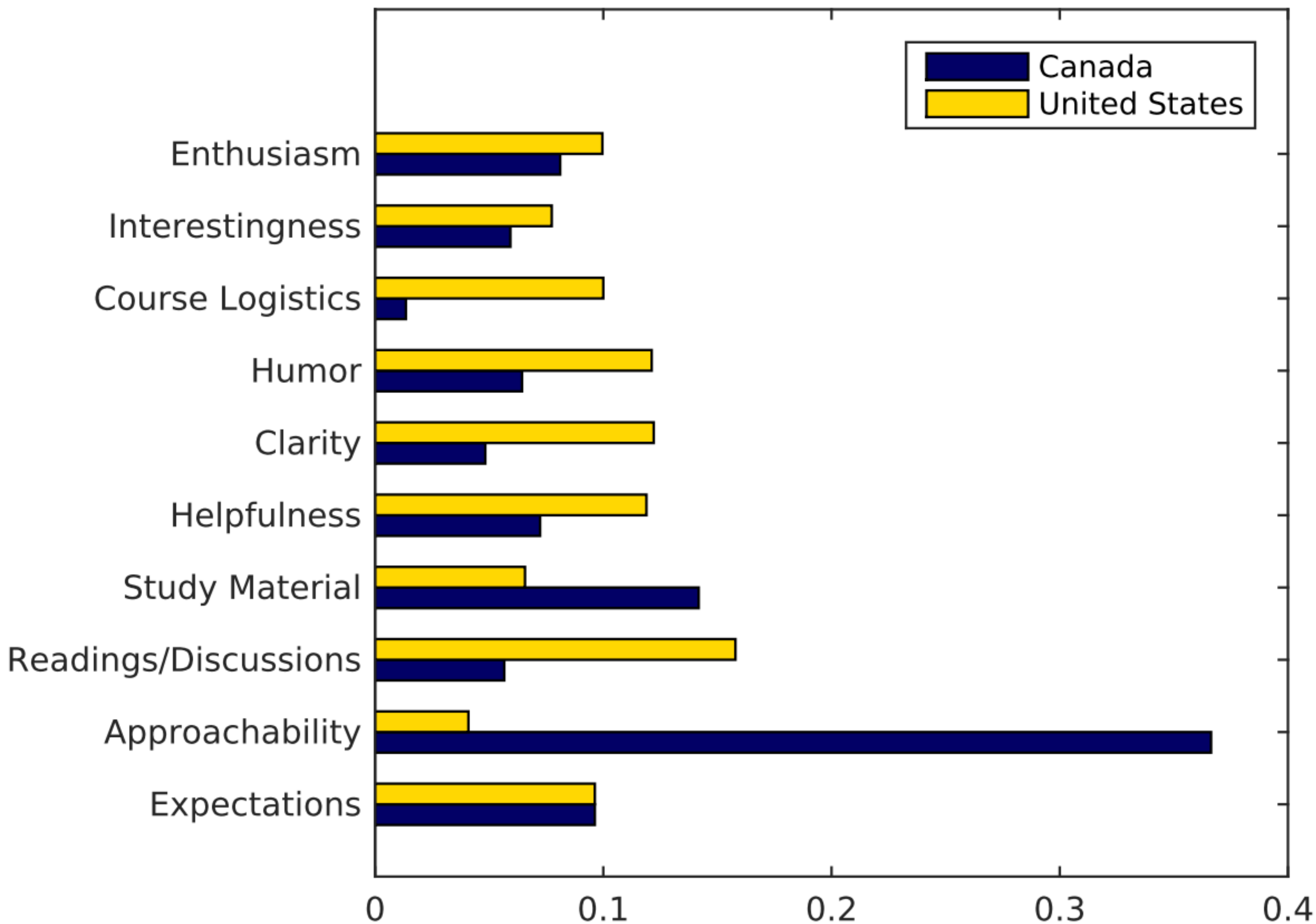
Azab, Mihalcea, and Abernathy, 2016

What do students look for in a professor?



Azab, Mihalcea, and Abernathy, 2016

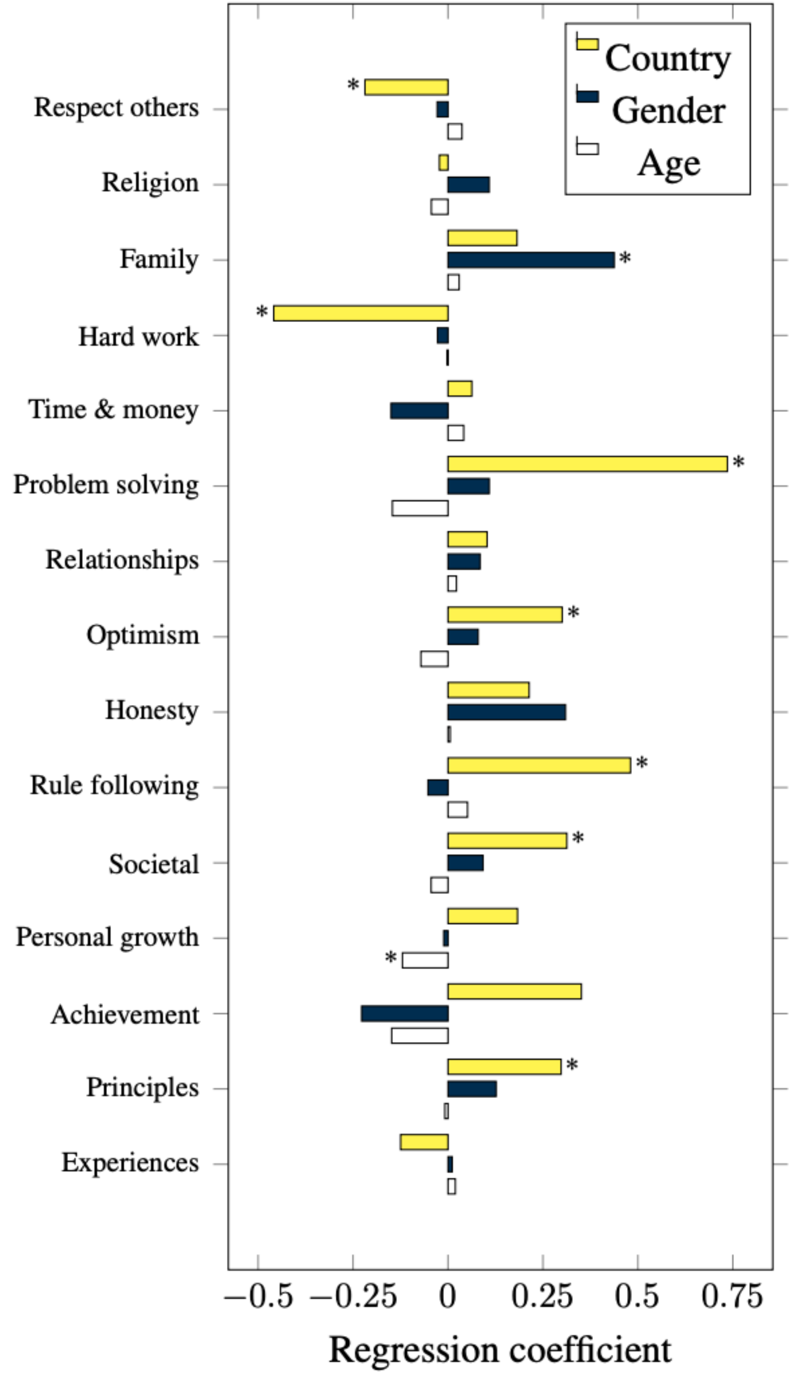
What do students look for in a professor?



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
- "Latent Dirichlet Allocation" by David Blei, Andrew Y. Ng, and Michael I. Jordan
- "Probabilistic Topic Models" by Mark Steyvers and Tom Griffiths

To watch:

- Guest lecture (2017) by David Blei at University of Edinburgh School of Informatics