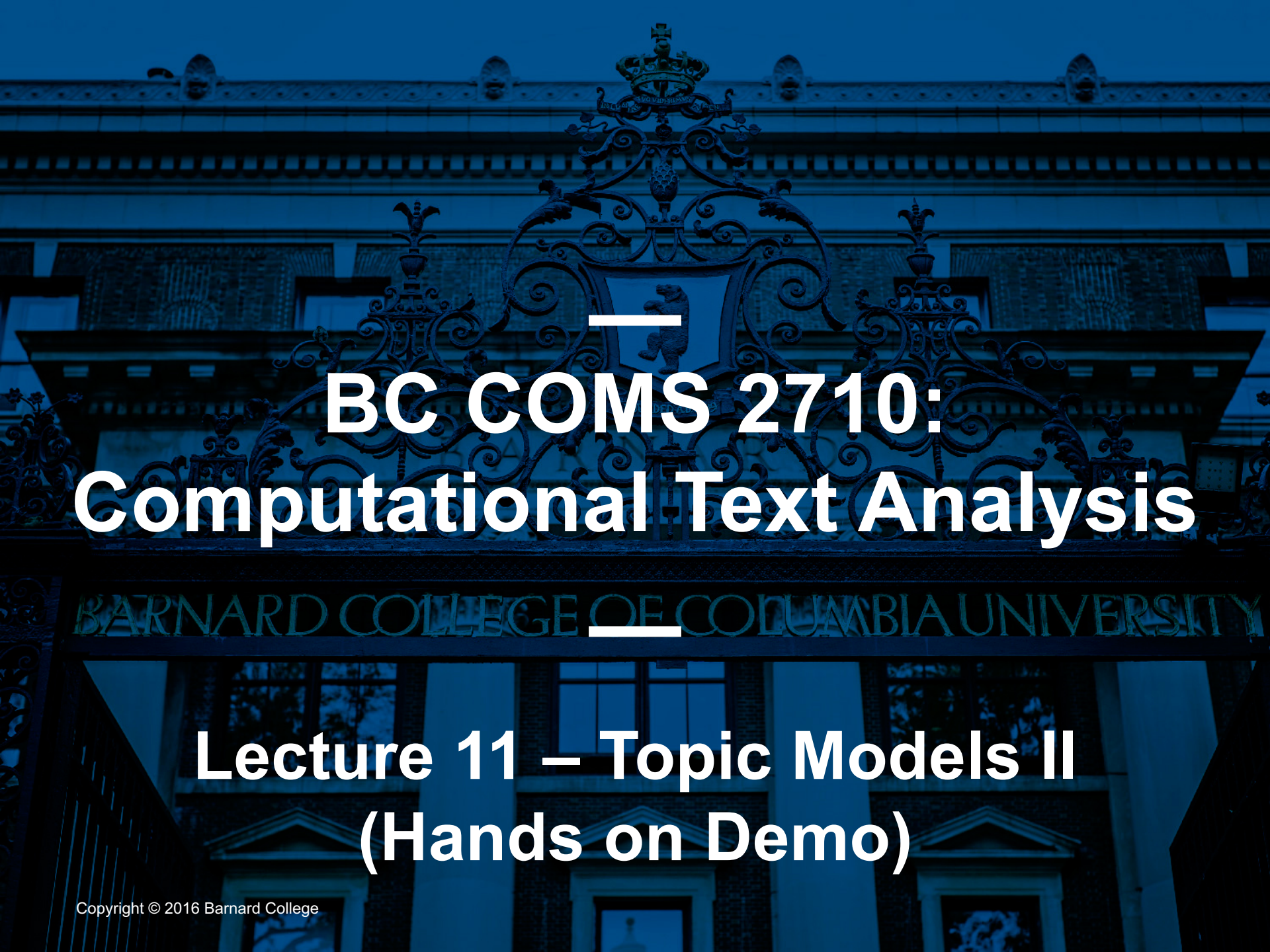


BC COMS 2710: Computational Text Analysis

Lecture 11 – Topic Models II (Hands on Demo)

Announcements – Assignments

- Tutorial 2.2
  - Due tonight (Thursday, 04/20)
  - Long - Broken into lots of small steps
- Readings:
  - Reading 03 – link course site, due Sunday
- HW 02:
  - Released later today
  - Shorter & Open ended assignment
- Office hours
  - Today 5-6pm
- Final Project
  - We will discuss on Monday

~~4~~ 3 Homeworks

- Readability of Inaugural Addresses
  - Due Monday 05/10 – available online
- Exploring NYTimes Obituaries
- Scraping and finding biases in CULPA reviews
- ~~Machine Learning~~

Course Feedback - Optional

- Mid-semester anonymous brief survey
- What have you learned so far and how comfortable do you feel with the material?
- What has been going well in the course so far? What are things you are enjoying about the course?
- What has not been going well in the course so far? What are things you are not enjoying about the course?
- What can we (the course staff) be doing better?

LDA Review

Training LDA Model

1. Randomly assign words to topics
2. Repeat many times:
    1. For each document:
        1. For each token, re-assign the topic based on:
            1. Topic assignment for every other token in the document
            2. Topic assignment for every other instance of the type in the the corpus
3. Return: Topics assignments for all tokens

LDA Columbia Connection

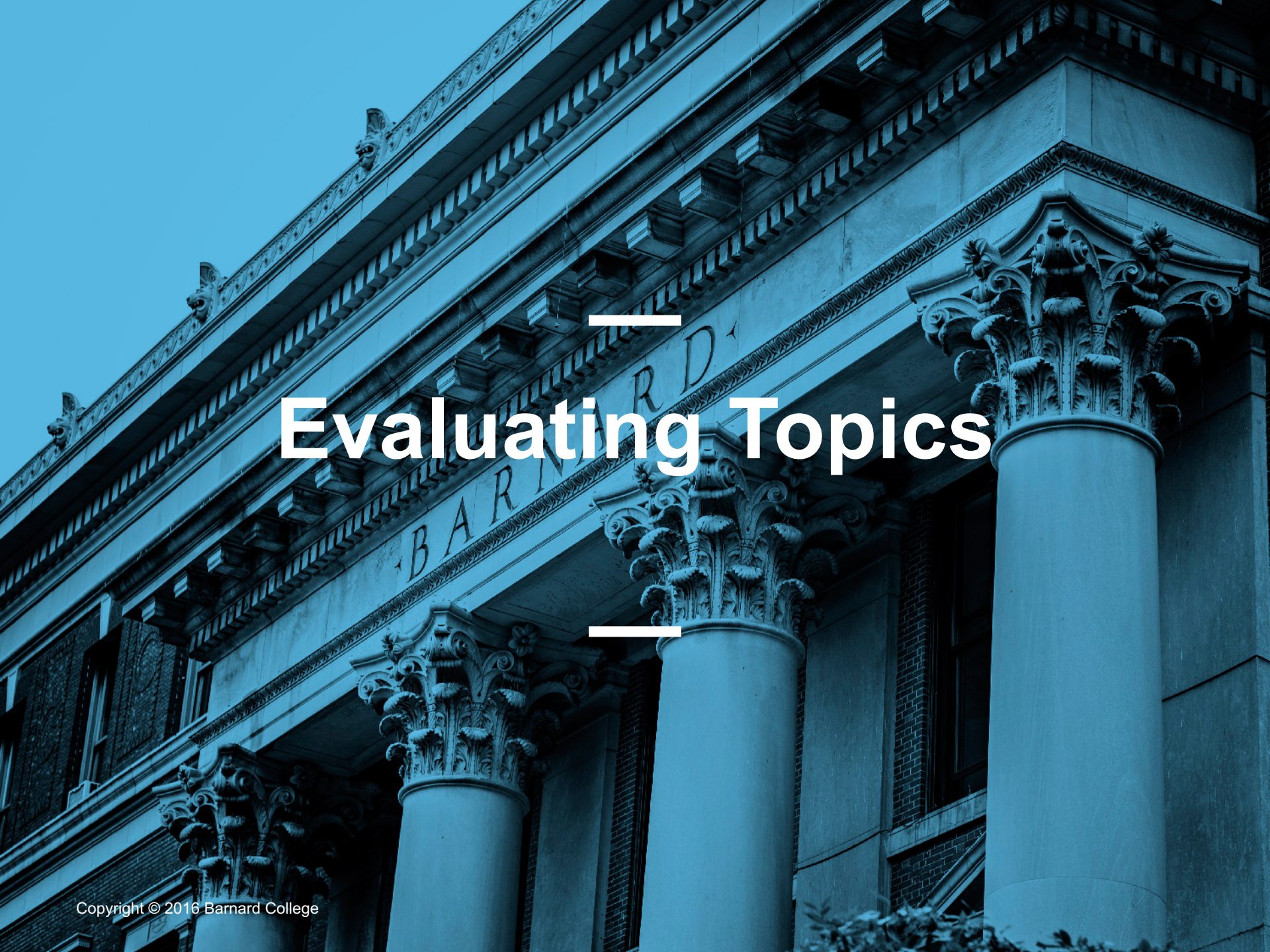
Journal of Machine Learning Research 3 (2003) 993-1022 Submitted 2/02; Published 1/03

Latent Dirichlet Allocation

David M. Blei BLEI@CS.BERKELEY.EDU
Computer Science Division
University of California
Berkeley, CA 94720, USA

Andrew Y. Ng ANG@CS.STANFORD.EDU
Computer Science Department
Stanford University
Stanford, CA 94305, USA

Michael I. Jordan JORDAN@CS.BERKELEY.EDU
Computer Science Division and Department of Statistics
University of California
Berkeley, CA 94720, USA



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 band songs rock album jazz pop song singer night	book life novel story books man stories love children family	art museum show exhibition artist artists paintings painting century works	game knicks nets points team season play games night coach	show film television movie series says life man character know
theater play production show stage street broadway director musical directed	clinton bush campaign gore political republican dole presidential senator house	stock market percent fund investors funds companies stocks investment trading	restaurant sauce menu food dishes street dining dinner chicken served	budget tax governor county mayor billion taxes plan legislature fiscal

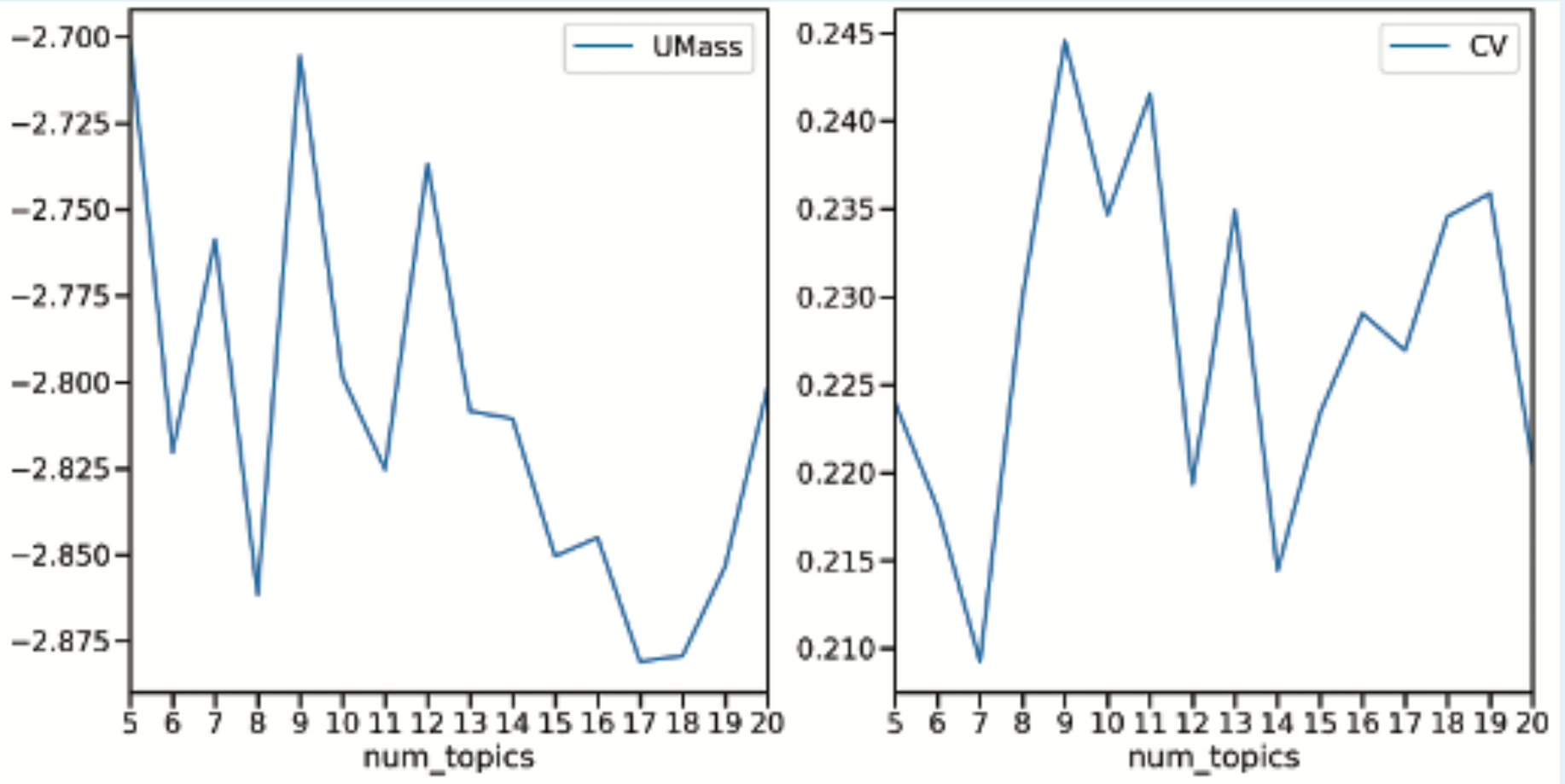
Automatic Metrics – Topic Coherence

- 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 Python

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