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
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?
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 corpus

3. Return: Topics assignments for all tokens
Latent Dirichlet Allocation

David M. Blei
Computer Science Division
University of California
Berkeley, CA 94720, USA

Andrew Y. Ng
Computer Science Department
Stanford University
Stanford, CA 94305, USA

Michael I. Jordan
Computer Science Division and Department of Statistics
University of California
Berkeley, CA 94720, USA
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
The average or median of pairwise word similarities formed by top words of a given topic.
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:
Using Topic Coherence to choose $k$
LDA Popularity

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

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