CS 293/EDUC 473

Discovery & Exploration in Educational Text Data

Topic Modeling, Clustering

Stanford
GRADUATE SCHOOL OF
EDUCATION

Questions about HW1 & syllabus?

Reminders

- HW1 due next Tuesday at midnight
- Final project
  - Dora has extra office hours tomorrow between 3-4pm in CERAS
  - Project rationale due the following Tuesday (Oct 16)
  - See Ed Forum announcement for finding project partners
- Monday: guest lecture by Michael Madaio from Google Research on fairness and bias in AI for education
  - Reading commentary due Sunday at 5pm

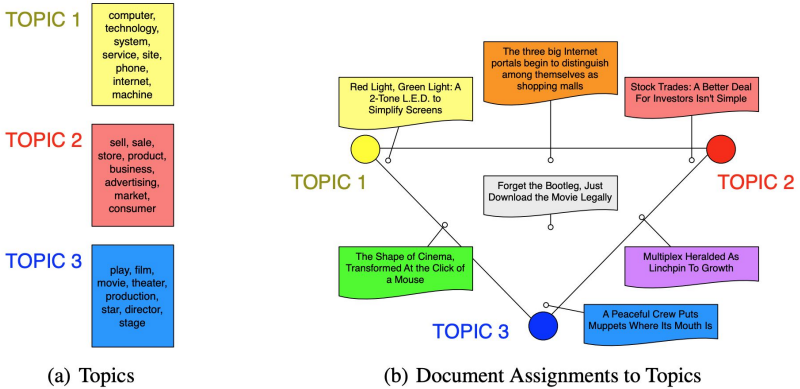
Today’s class

- Q&A with Dr Sarah Johnson from the Teaching Lab
- Brief lecture on topic modeling & clustering
- Reading discussion for Liu & Cohen (2021) led by Jurgen and Chenglei

Q&A with Sarah Johnson

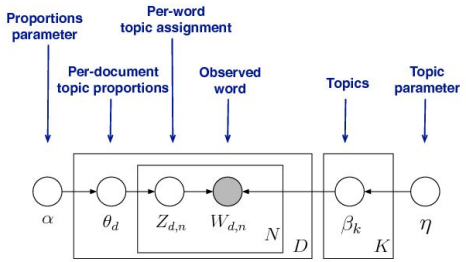
Topic modeling

- Going beyond looking at unique words to looking at **themes**
- There are several topic modeling approaches; in my experience, LDA MALLET (David Blei) works best for simple explorations
  - Original command line package
  - Python wrapper
- LDA outputs per-document topic proportions and per-topic word proportions
- Take a look at these slides here to learn more



(a) Topics (b) Document Assignments to Topics

LDA as a graphical model



- Encodes **assumptions**
- Defines a **factorization** of the joint distribution
- Connects to **algorithms** for computing with data

Credit: Jason Yosinski

Pre-processing decisions for topic modeling **matter**

- Highest probability words within topics are often stop words (e.g. “I”, “not”, “will”) → **remove stopwords (?)**
  - Sometimes stopwords lead to meaningful distinctions (negations, use of 1st vs 3rd person)
- Forms of the same word may appear together as highest probability words (e.g. “like”, “liked”, “liking”) → **stem / lemmatize words (?)**
  - but: they may decrease stability (Schoefield & Mimno, 2016) and can lead to combining words with different meanings
- **Lowercasing?**
  - may collapse named entities with non-named entities; removes emphasis (e.g. all caps)
- (Denny & Spirling, 2017) introduce an R package **preText** to measure the impact of preprocessing on topics

Evaluating topic models: Reading tea leaves (Chang et al., 2019)

- Manual methods are the most reliable!
- Two evaluation tasks
  - **Word Intrusion**
  - **Topic Intrusion**
    - More challenging with longer texts
- Relatively quick once it's set up
  - For resource-efficiency, first select ~3 most promising parameter settings (num topics + preprocessing decisions) by eyeballing + automated metrics and compare those

Please select the word which is out of place or does not belong with the others.

◯ password ◯ help ◯ log ◉ woke ◯ account

Word Intrusion Task

Please select the group of words which is out of place or does not belong with the tweet.

Tweet: @SpotifyCares Keep getting Gateway error messages. Don't have time for this.

◉ lyrics; back; notifications; app; feature

◯ error; get; tried; message; says

Topic Intrusion Task

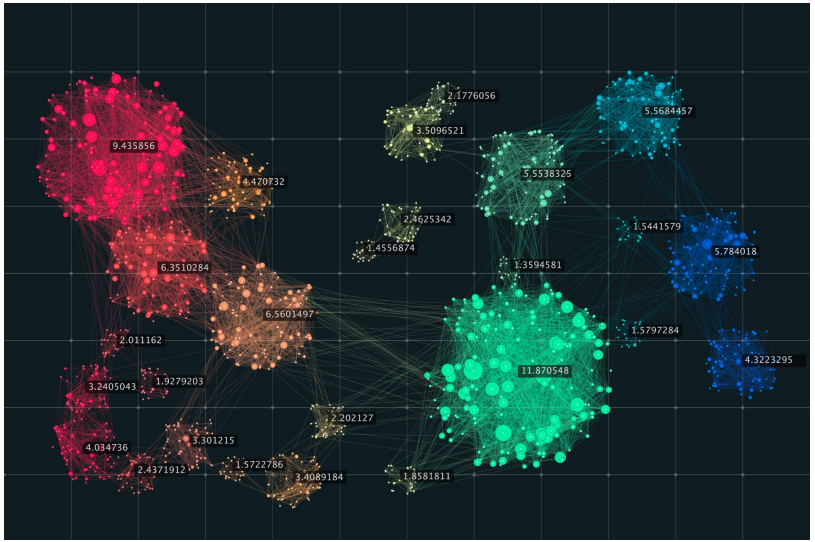
image credit: (Guzman et al., 2017)

Clustering

Group a set of data points into a number of clusters, so that

- Data points in the same cluster are similar to each other
- Data points in different clusters are dissimilar

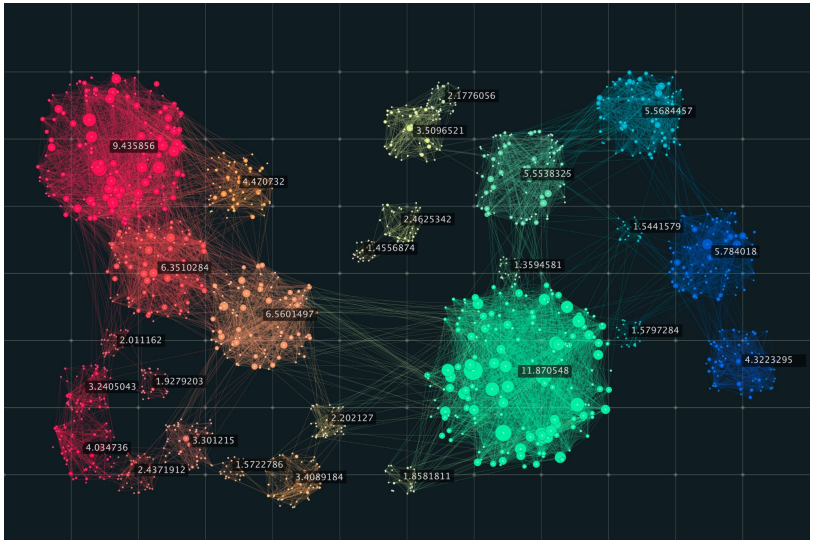
Use only X, not Y → unsupervised method



Slide credit: Diyi Yang

Cluster structures

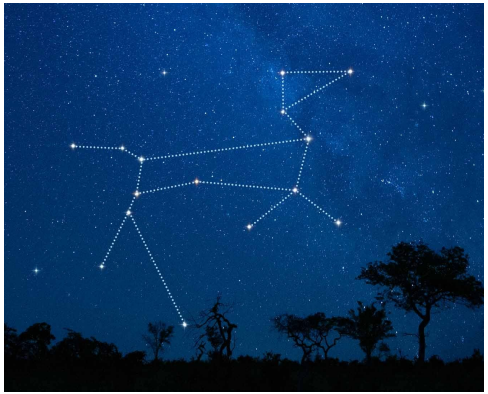
- Partitioning a group of data point into K disjoint sets (K-means clustering)
- Assigning X to hierarchical structures (Hierarchical clustering)
- Assigning X to partial membership in K different sets (Graphic models, GMM)
- Learning a representation that puts similar data points closer to each other (Word embeddings w/ deep learning)



Slide credit: David Bamman & Diyi Yang

When to use topic modeling & clustering?

- Discovering **interesting or unexpected structures** can be useful for hypothesis generation
- Unsupervised learning generates **alternative representations** as features for some subsequent (unsupervised or supervised) models
  - E.g. topics or clusters can be used as **inputs in a log odds comparison** (previous lecture)
  - E.g. topics or clusters can be used in a **classification / regression model**



Slide partial credit: David Bamman & Diyi Yang

Word embedding based representations

- Simple but tough to beat baseline sentence embeddings
  - tf-idf weighted average of word embeddings
- Sentence transformers (see HW1)
  - base model is sBERT

```
Cluster  11
['This is our first question.', 'Here is the next one.', 'Here is your first question.', 'This is our next question.',
"Let's get directed to the next question.", "Let's move on to the next question.", 'Here is our first practice question.
', "Let's get directed to the next question.", 'Let us get directed to the next question.', "Let's get directed to the
next question.", 'This is your next question.', 'Here is our first question.', 'This is your first independent question.
', 'Here you go for the next one.', 'Here is your next question.', 'This is our first question.', 'Here comes the next
```

```
Cluster  7
['That was a good try.', 'That is a very good start.', "That's good.", "That's good.", 'Very good.', 'Great!', "That's
fantastic.", 'Well done.', 'All the best.', 'Excellent.', 'Great going.', 'That was a nice try.', "Let's get started.",
'That was a great start.', 'Now try this one.', 'Awesome!', "Let's get started.", 'This is a very good start.', "That's
great.", "That's great!", 'That was a good start.', 'Great try.', 'That was a great try.', 'That was a good try.', 'Good
work!', 'That was a good try!', 'That was a good try.', "Let's get started.", "That's good.", 'Nice One!', "Let's get
started.", "That's good.", 'Good try.', "That's great.", 'Good job!', 'Lets get started!', 'That was a good start.',
```

What to think about when doing clustering?

- How to **represent** each data point?
- How to **calculate the similarity** between data points?
- What is the **number of clusters** to use?
- How can we **evaluate** the resulting clusters?

Slide credit: Diyi Yang

Large Language Models for data exploration

E.g. Thematic Analysis (De Paoli, 2023; Gamieldien et al., 2023), Topic Modeling (Wang et al., 2023)

Pros:

- Can do very well at producing highly interpretable topics with good coverage

Downside:

- Hallucinations
- Black box process; the outputs require extensive evaluation to ensure they're accurately representing the data

Nr.	Theme	Description
1	Teaching Data Analysis and Interpretation	This group includes topics related to teaching students how to analyze and interpret data, including identifying good and bad graphs, understanding statistical knowledge, and teaching critical thinking about data.
2	Mentoring and Diversifying the Field	This group includes topics related to mentoring young students and making a difference in diversifying the field of data analysis.
3	Teaching GIS and Geospatial Data	This group includes topics related to teaching GIS software and geospatial data, including challenges in teaching and the practical use of the software.
4	Collaborative Learning and Interpersonal Interaction	This group includes topics related to the benefits of collaborative learning and interpersonal interaction in acquiring quantitative skills.
5	Teaching with Big Data	This group includes topics related to incorporating big data into teaching and the challenges of introducing it to students.
6	Training and Support for Teaching with Data	This group includes topics related to the lack of training and support for instructors in teaching with data, and suggestions for workshops and resources.

Reading Discussion for Liu & Cohen (2021)