AI INSTITUTE #AIISC
UNIVERSITY OF SOUTH CAROLINA

UNIVERSITY OF SOUTH CAROLINA

CSCE 771: Computer Processing of Natural Language
Lecture 20: Topic Analysis

PROF. BIPLAV SRIVASTAVA, AI INSTITUTE

29TH OCTOBER, 2024

Carolinian Creed: "I will practice personal and academic integrity."

Organization of Lecture 20

- Opening Segment
  - Announcements
- Main Lecture
- Concluding Segment
  - About Next Lecture – Lecture 21

Main Section

- Topic Analysis
- LSA
- LDA
- Topic Classification

Recap of Lecture 19

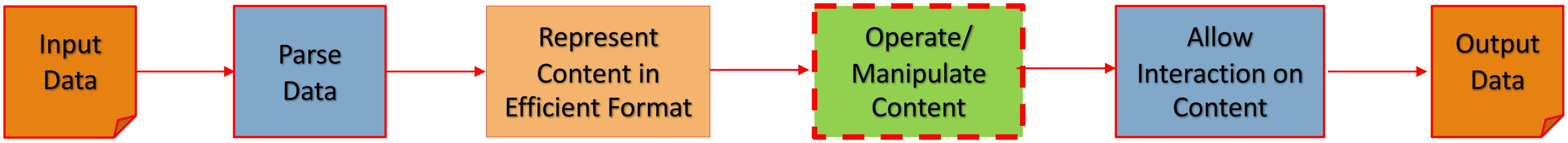
We looked at

- What is an event?
- Extraction and linking
- Spatio-temporal reasoning
- Applications

Main Lecture

Topic Detection and Analysis

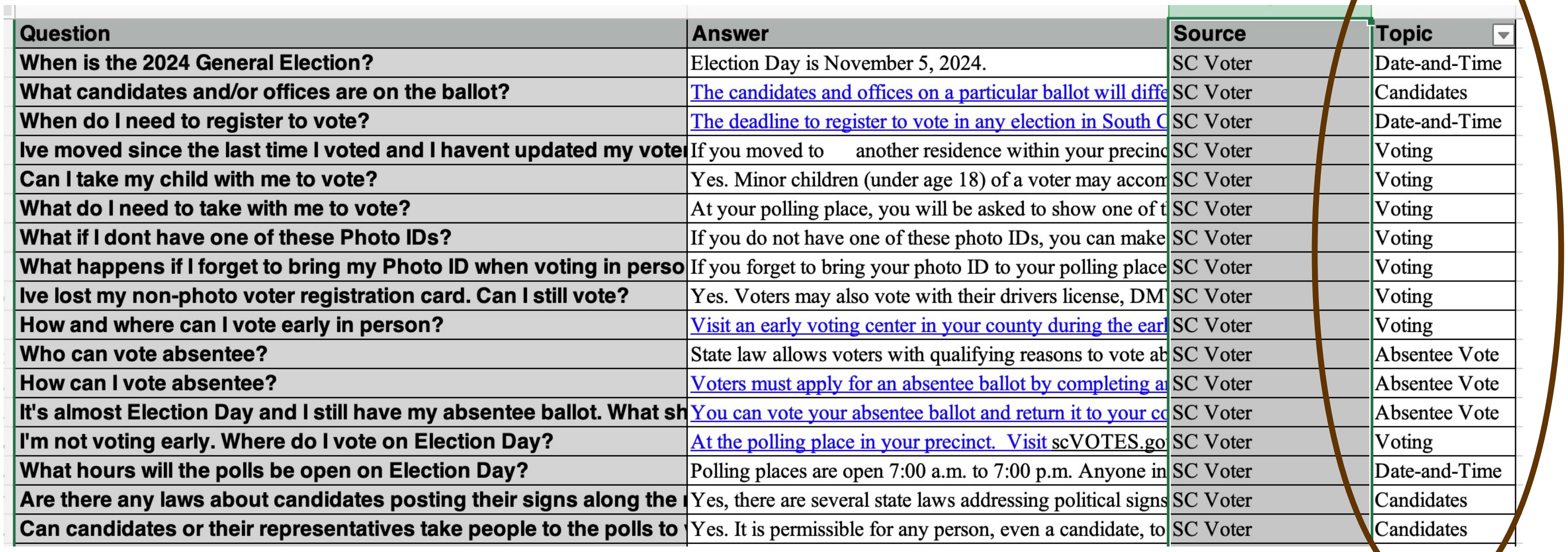
Statistical patterns identified from textual data


Input Data
Parse Data
Represent Content in Efficient Format
Operate/ Manipulate Content
Allow Interaction on Content
Output Data

Motivation for Topic Analysis

- Quickly find patterns in textual data (documents)
- Other related concepts -- examples
  - ***Word tag cloud*** – frequency based
  - **Topics** – statistical property
  - ***Summary*** – content based
- Usage
  - Manage documents
  - Classify text into groups

Example in SC Election FAQs



Question	Answer	Source	Topic
When is the 2024 General Election?	Election Day is November 5, 2024.	SC Voter	Date-and-Time
What candidates and/or offices are on the ballot?	The candidates and offices on a particular ballot will diffe	SC Voter	Candidates
When do I need to register to vote?	The deadline to register to vote in any election in South C	SC Voter	Date-and-Time
Ive moved since the last time I voted and I havent updated my voter	If you moved to another residence within your precinc	SC Voter	Voting
Can I take my child with me to vote?	Yes. Minor children (under age 18) of a voter may accom	SC Voter	Voting
What do I need to take with me to vote?	At your polling place, you will be asked to show one of t	SC Voter	Voting
What if I dont have one of these Photo IDs?	If you do not have one of these photo IDs, you can make	SC Voter	Voting
What happens if I forget to bring my Photo ID when voting in perso	If you forget to bring your photo ID to your polling place	SC Voter	Voting
Ive lost my non-photo voter registration card. Can I still vote?	Yes. Voters may also vote with their drivers license, DM	SC Voter	Voting
How and where can I vote early in person?	Visit an early voting center in your county during the earl	SC Voter	Voting
Who can vote absentee?	State law allows voters with qualifying reasons to vote ab	SC Voter	Absentee Vote
How can I vote absentee?	Voters must apply for an absentee ballot by completing a	SC Voter	Absentee Vote
It's almost Election Day and I still have my absentee ballot. What sh	You can vote your absentee ballot and return it to your co	SC Voter	Absentee Vote
I'm not voting early. Where do I vote on Election Day?	At the polling place in your precinct. Visit scVOTES.go	SC Voter	Voting
What hours will the polls be open on Election Day?	Polling places are open 7:00 a.m. to 7:00 p.m. Anyone in	SC Voter	Date-and-Time
Are there any laws about candidates posting their signs along the r	Yes, there are several state laws addressing political signs	SC Voter	Candidates
Can candidates or their representatives take people to the polls to v	Yes. It is permissible for any person, even a candidate, to	SC Voter	Candidates

What information do they convey? Are they accurate? Should there be more than one? When are they enough?

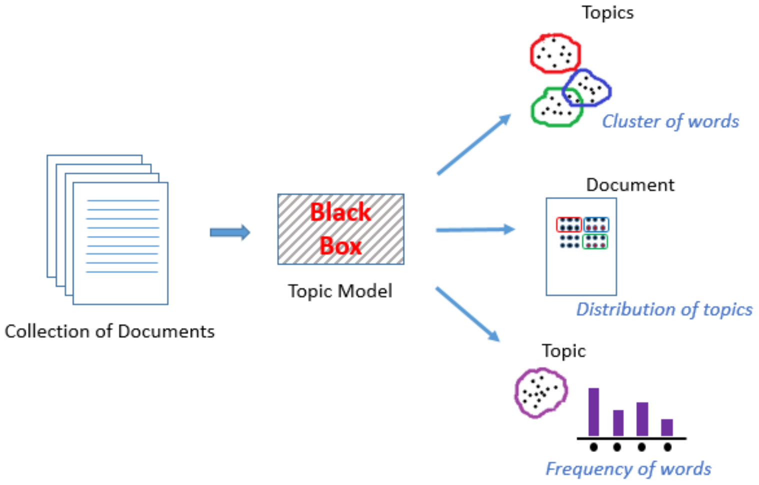
What is a Topic?

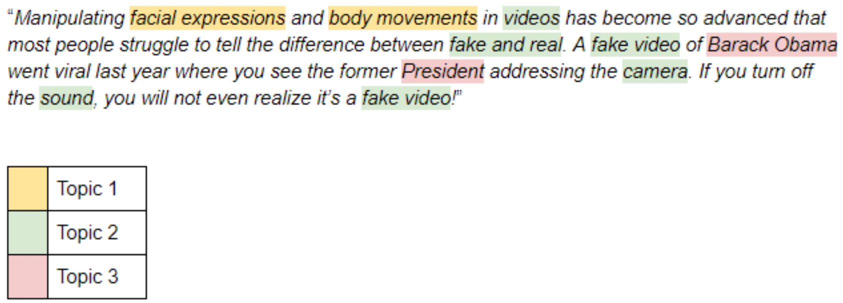
- Words: building block on language writing; separated by white-spaces
  - Other building blocks: sentences, paragraphs
- Documents: logical / physical organization of content
- Topics are:
  - Set of words/ phrases that are indicative of document/ corpus content

Two Categories of Techniques

- Topic Learning – *unsupervised*
  - Topic as implicit concept
- Topic Classification – *supervised*
  - Topic as label

Conceptual Framework and Example




"Manipulating facial expressions and body movements in videos has become so advanced that most people struggle to tell the difference between fake and real. A fake video of Barack Obama went viral last year where you see the former President addressing the camera. If you turn off the sound, you will not even realize it's a fake video!"

🟨	Topic 1
🟩	Topic 2
🟥	Topic 3

Credit: https://www.analyticsvidhya.com/blog/2018/10/stepwise-guide-topic-modeling-latent-semantic-analysis/

Topic Learning

- Words: building block on language writing; separated by white-spaces
  - Other building blocks: sentences, paragraphs
- Documents: logical / physical organization of content
- Topics:
  - Implicit concept - **Latent**
  - Set of words/ phrases that are indicative of document/ corpus content

Many techniques:

- Singular Value Decomposition (SVD)
- Latent Semantic Indexing (LSI) (Deerwester et al., 1988), Latent Semantic Analysis (LSA) (Deerwester et al., 1990)
- Latent Dirichlet Allocation (LDA) (Blei et al., 2003)
- Non-negative Matrix Factorization (NMF) (Lee and Seung, 1999)
- LDA2VEC, …

Singular-Value Decomposition

(Compact) SVD Idea:

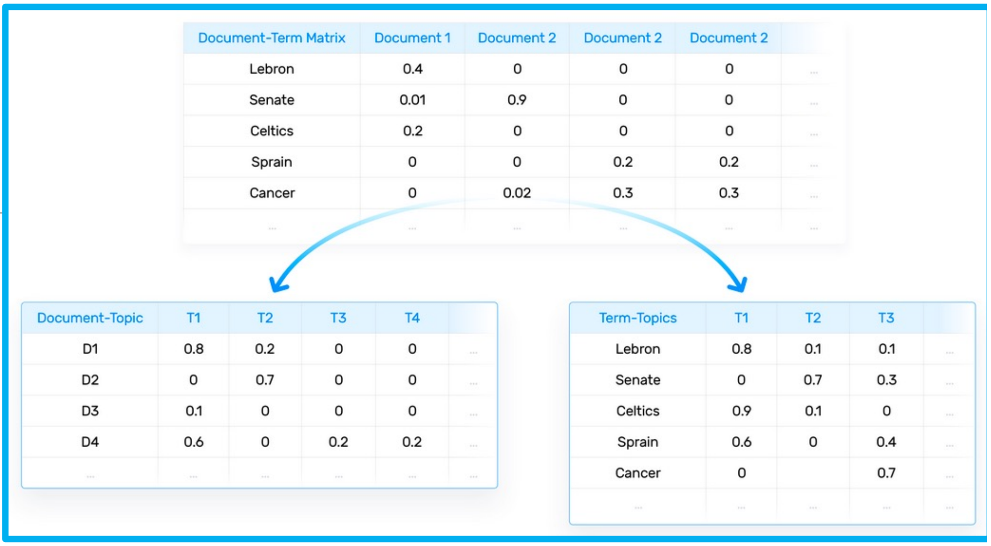
A(m*n) = U(m*r) x S(r*r) x V(r*n)

A(m*n) = U(m*m) x S(m*n) x V(n*n)

Matrix S is a diagonal matrix of the singular values of the original matrix.

Credits: https://monkeylearn.com/topic-analysis/, Mausam lecture slides

Document – Term Matrix



Document-Term Matrix	Document 1	Document 2	Document 2	Document 2	
Lebron	0.4	0	0	0	...
Senate	0.01	0.9	0	0	...
Celtics	0.2	0	0	0	...
Sprain	0	0	0.2	0.2	...
Cancer	0	0.02	0.3	0.3	...
...	...	...	...	...	...

Document-Topic	T1	T2	T3	T4	
D1	0.8	0.2	0	0	...
D2	0	0.7	0	0	...
D3	0.1	0	0	0	...
D4	0.6	0	0.2	0.2	...
...	...	...	...	...	...

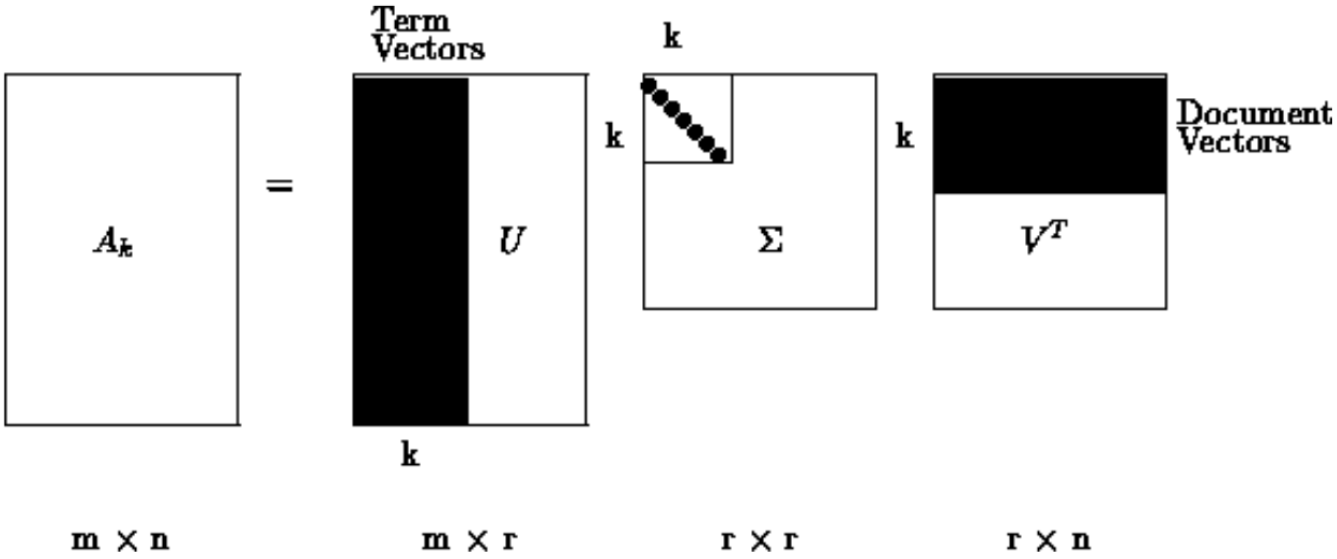
Term-Topics	T1	T2	T3	
Lebron	0.8	0.1	0.1	...
Senate	0	0.7	0.3	...
Celtics	0.9	0.1	0	...
Sprain	0.6	0	0.4	...
Cancer	0		0.7	...
...	...	...	...	...

Document – Topic Matrix

Term – Topic Matrix

Informally: consider documents in a corpus as a distribution over topics – a latent set words – which is distributed over terms in the documents

LSA - Latent Semantic Analysis



Document – Term Matrix

Document – Topic Matrix

Topic – Topic Matrix

Terms – Topic Matrix

Credits: Mausam lecture slides

Elements of S (i.e., Σ) are the topics

LDA - Latent Dirichlet Allocation

- Each topic is represented by an (unknown) set of words.
- Assumption: Every document is composed of a mixture of topics, and every word has a probability of belonging to a certain topic.
- Cover all the (known) documents in the corpus to the (unknown) topics in a way such that the words in each document are mostly captured by those topics.
- **Objective**: “a generative probabilistic model of a corpus that not only assigns high probability to members of the corpus, but also assigns high probability to other “similar” documents.”

- Video lecture by Prof. Blei: https://www.youtube.com/watch?v=FkckgwMHP2s

LDA paper: https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf
Blog: https://monkeylearn.com/topic-analysis/,

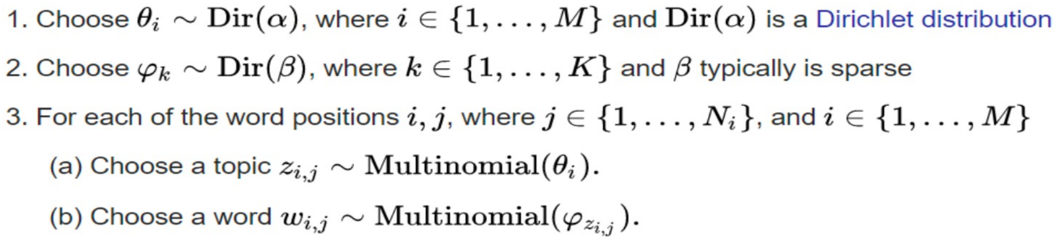
LDA - Latent Dirichlet Allocation

- Intuition
  - any corpus (collection of documents) can be represented as a **Document-Term** matrix. The value of i,j cell gives the frequency count of word Wj in document Di.
  - LDA converts this **Document-Term** Matrix into two lower dimensional matrices – M1 and M2.
  - M1 is a **Document-Topics** matrix and M2 is a **Topic – Terms** matrix with dimensions (N, K) and (K, M) respectively, where N is the number of documents, K is the number of topics and M is the vocabulary size.

Credit: https://www.analyticsvidhya.com/blog/2016/08/beginners-guide-to-topic-modeling-in-python/

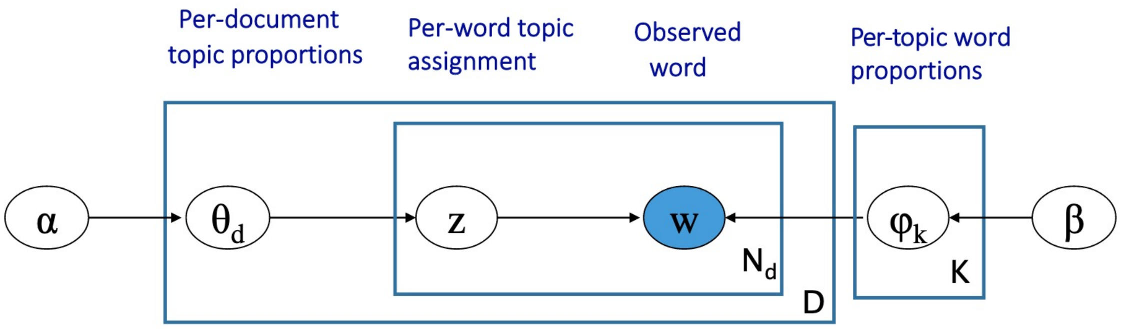
LDA - Latent Dirichlet Allocation

- Generative Model



1. Choose $\theta_i \sim \mathrm{Dir}(\alpha)$, where $i \in \{1, \ldots, M\}$ and $\mathrm{Dir}(\alpha)$ is a Dirichlet distribution
2. Choose $\varphi_k \sim \mathrm{Dir}(\beta)$, where $k \in \{1, \ldots, K\}$ and β typically is sparse
3. For each of the word positions i, j, where $j \in \{1, \ldots, N_i\}$, and $i \in \{1, \ldots, M\}$
   - (a) Choose a topic $z_{i,j} \sim \mathrm{Multinomial}(\theta_i)$.
   - (b) Choose a word $w_{i,j} \sim \mathrm{Multinomial}(\varphi_{z_{i,j}})$.

Credit: Mausam slides;
LDA paper:
https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf



Per-document topic proportions

Per-word topic assignment

Observed word

Per-topic word proportions

From LDA paper - The boxes are "plates" representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document.

LDA Parameters

Alpha and Beta Hyperparameters –

- Alpha represents **document-topic densit**y
  - Higher the value of alpha, documents are composed of more topics and lower the value of alpha, documents contain fewer topics.

- Beta represents **topic-word density**
  - Higher the beta, topics are composed of more number of words in the corpus, and with the lower value of beta, they are composed of fewer words.

Credit: https://www.analyticsvidhya.com/blog/2016/08/beginners-guide-to-topic-modeling-in-python/

Code Exercises

- Working code: https://github.com/biplav-s/course-nl-f22/blob/main/sample-code/l19-topic/ExploreTopics.ipynb
- Exercise #1
  - Data: Copy file-1 (Example-TDBank-PersonalAcctAgree) data into local directory.
  - Activity: Run notebook on it. Compare output of url fetch v/s local file
- Exercise #2
  - Data: Take your favorite piece of text. Example resume
  - Activity: Run notebook on it. Explore output of LDA visualizer

Code Example

https://github.com/biplav-s/course-nl/blob/master/l17-topicanalysis/ExploreTopics.ipynb

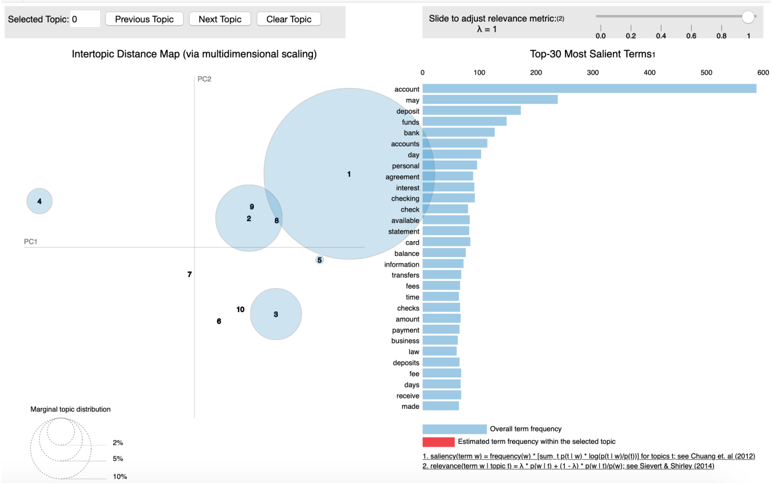
Libraries:

- Gensim: https://radimrehurek.com/gensim/models/ldamodel.html, https://radimrehurek.com/gensim/auto_examples/core/run_topics_and_transformations.html
- Scikit-learn: https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html
- LDA2VEC: https://github.com/cemoody/lda2vec?tab=readme-ov-file
  - Notebook: http://nbviewer.jupyter.org/github/cemoody/lda2vec/blob/master/examples/twenty_newsgroups/lda2vec/lda2vec.ipynb

Visualization of Topics

- LDA: PyLDAVis - https://github.com/bmabey/pyLDAvis
  - Talk on visualizing topics for 20 Newsgroups
    - https://www.youtube.com/watch?v=IksL96ls4o0

- Other measures (SVD)
  - Arrange documents by similarity of topics using bokeh – https://nlpforhackers.io/topic-modeling/



Selected Topic: 0

Slide to adjust relevance metric:(2)
λ = 1

Overall term frequency

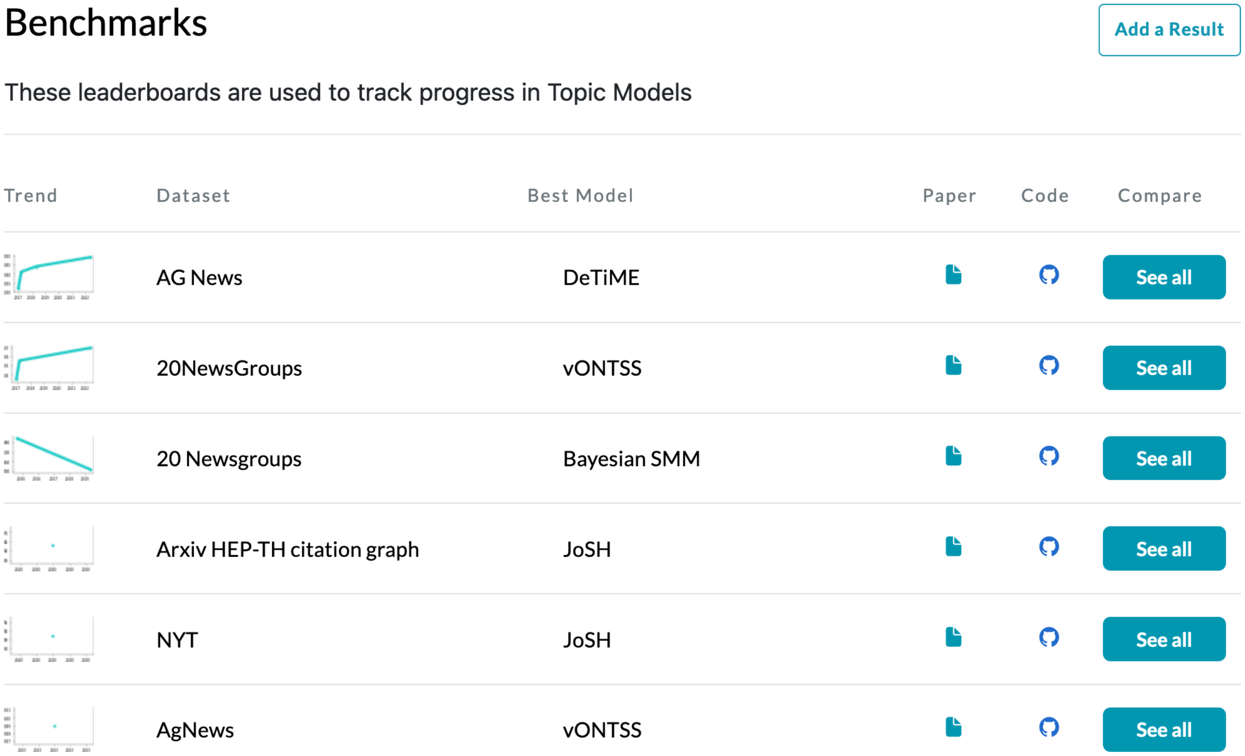
🟥 Estimated term frequency within the selected topic

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)

2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Tracking Best Performance

https://paperswithcode.com/task/topic-models



Benchmarks

Add a Result

These leaderboards are used to track progress in Topic Models

Trend	Dataset	Best Model	Paper	Code	Compare
	AG News	DeTiME	📄	●	See all
	20NewsGroups	vONTSS	📄	●	See all
	20 Newsgroups	Bayesian SMM	📄	●	See all
	Arxiv HEP-TH citation graph	JoSH	📄	●	See all
	NYT	JoSH	📄	●	See all
	AgNews	vONTSS	📄	●	See all

Measuring Quality of Topics

- Coherence (c_v) of topics identified
- References:
  - Blog: https://towardsdatascience.com/c%E1%B5%A5-topic-coherence-explained-fc70e2a85227
  - Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. WSDM 2015 - Proceedings of the 8th ACM International Conference on Web Search and Data Mining, 399–408. https://doi.org/10.1145/2684822.2685324
  - Newman, J. H. Lau, K. Grieser, and T. Baldwin. Automatic evaluation of topic coherence. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pages 100–108. 2010.

References

- Blogs
  - LSA, https://www.analyticsvidhya.com/blog/2018/10/stepwise-guide-topic-modeling-latent-semantic-analysis/, 2024

- Cvitanić, Tonči, Bumsoo Lee, Hyeon Ik Song, Katherine K. Fu and David W. Rosen. “LDA v. LSA: A Comparison of Two Computational Text Analysis Tools for the Functional Categorization of Patents.” ICCBR Workshops (2016).
- Zengul, Ferhat Devrim, Ayşegül Bulut, Nurettin Oner, Abdulaziz Ahmed, Manju Yadav, Hope G. Gray and Bunyamin Ozaydin. “A Practical and Empirical Comparison of Three Topic Modeling Methods Using a COVID-19 Corpus: LSA, LDA, and Top2Vec.” Hawaii International Conference on System Sciences (2023).

Topic Classification

- Supervised task of assigning labels to a document
  - Assumption: topics for the population corpus are known
- For documents in corpus:
  - From the set of topics assigned to document, pick the topic with the highest probability
- For new documents:
  - Train a supervised classifier on known documents using topic labels from corpus
  - Assign topic to new documents from the learned classifier

Also see: https://www.kdnuggets.com/2019/11/topics-extraction-classification-online-chats.html

Review Paper

Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2021. Deep Learning--based Text Classification: A Comprehensive Review. ACM Comput. Surv. 54, 3, Article 62 (April 2022), 40 pages. https://doi.org/10.1145/3439726

Topic – Practical Considerations

- Can we assume topics are distributed across corpus ?
- How to be robust
  - Common words
  - Noisy text
- Drift of topics over time

Comments: Topic and Language Models

- Topic Modeling in Embedding Spaces, Adji B. Dieng, Francisco J. R. Ruiz, David M. Blei, TACL 2020
  - Embedded Topic Model (ETM) – "the etm models each word with a categorical distribution whose natural parameter is the inner product between the word's embedding and an embedding of its assigned topic"
  - Handles rare words and stop words

https://paperswithcode.com/paper/topic-modeling-in-embedding-spaces

Discussion

- Topic learning or topic classification
- Metric to measure goodness
- What to do for finding topics in election FAQs ?

Lecture 19: Concluding Comments

- We reviewed topic analysis
- Statistical property indicating key insights about a document
- Topic modeling/ detection
  - Identify topics
- Topic classification

Concluding Segment

Discussion: Course Project

Theme: Analyze quality of official information available for elections in 2024 [in a state]

- Take information available from
  - Official site: State Election Commissions
  - Respected non-profits: League of Women Voters
- Analyze information
  - State-level: Analyze quality of questions, answers, answers-to-questions
  - Comparatively: above along all states (being done by students)
- Benchmark and report
  - Compare analysis with LLM
  - Prepare report

- Process and analyze using NLP
  - Extract entities
  - Assess quality – metrics
    - Content – *Englishness*
    - Content – *Domain* -- election
  - … other NLP tasks
  - Analyze and communicate overall

Major dates for project check

- Sep 10: written – project outline
- Oct 8: in class
- **Oct 31: in class // LLM**
- Dec 5: in class // Comparative

Obtaining Election Data

Here are a few things to do:

A) **Official data** backed by laws: state election commission

a) Find the state's election commission

b) Find the Q/As they provide. They may be as FAQs or on different web pages.

c) Collect the Q/A programmatically

B) **Secondary data** sources: non-profit

a) Find Q/As from Vote 411 which is supported by the non-profit: LWV.

For reference, for SC,

A) Official - https://scvotes.gov/voters/voter-faq/

B) Secondary - https://www.vote411.org/south-carolina

For extraction, one or more approaches:

- Manually annotating
- BeautifulSoup,
- Tika
- or other open source libraries.

<Student Name>

CSCE 771 Fall 2024: Milestone#1

Initial analysis of questions (Q)

*

1. State Selected:
2. Election data sites:
   - Official site (e.g.,State Election Commission) url
   - Secondary site (e.g., League of Women Voters) url
1. Report how data collected and Q/A statistics

1. Take on NLP methods you will use and why for Q/A analysis
   1. State-level (right)
   2. Comparatively: above along states being done by peers

Initial analysis of answers (A)

*

Initial analysis of an answer (a_i) for a question (q_i)

*

Election Q/A for Your State

- Format in .json; name file as "**xy**_qa.json", where xy is the two-character US state acronym
- Fixed attributes in .json
  - state: xy
  - num_questions: **a**, where a is the number of questions
  - num_answers: **b**, where b is the number of answers
  - contributor: student name
- questions: List of Q/As with attributes for each it:
  - q // question
  - a // answer
  - s // source url from where the information is taken
  - t // time when the information is taken – UTC format
- Store it in your github repo; put in sub-dir like "project/data"
- Instructor will keep it in common place inside course github repo and share.

Election Q/As for Multiple States

- Instructor will keep it common place inside course github repo and share.
- You will be able to access Q/As of all states from common location
  - To compare data across all states

<Student Name>

CSCE 771 Fall 2024: Milestone#2

1. State Selected:
   1. Optionally: do for SC
   2. Optionally: compare your state and SC
2. Election data sites:
   - Official site (e.g.,State Election Commission) url
   - Secondary site (e.g., League of Women Voters) url

1. LLM (s) used:
   1. Tasks used for:
   2. Comments: <highlight effort, if any>

Analysis of questions (Q)

*LLM-specific comments

Analysis of answers (A)

*LLM-specific comments

Analysis of an answer (a_i) for a question (q_i)

*LLM-specific comments

Discussion – a Paper Based on All Data?

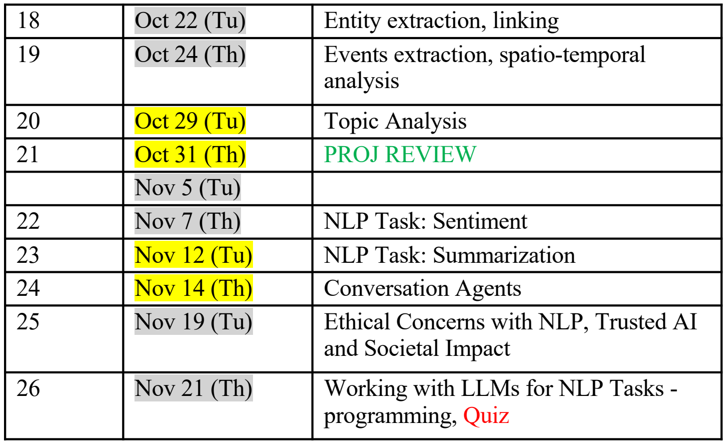
- Contributions
  - Analysis of current situation, perspective on gaps and opportunities with NLP
  - Dataset
- Logistics
  - Target venue
  - People
  - Timeline

About Next Lecture – Lecture 21

Lecture 21 Outline: Project Milestone #2 Update

- How has LLMs been used for election data analysis
  - for your state
  - Optional: for SC as well and comparison

- Template given in good-drive



18	Oct 22 (Tu)	Entity extraction, linking
19	Oct 24 (Th)	Events extraction, spatio-temporal analysis
20	Oct 29 (Tu)	Topic Analysis
21	Oct 31 (Th)	PROJ REVIEW
	Nov 5 (Tu)	
22	Nov 7 (Th)	NLP Task: Sentiment
23	Nov 12 (Tu)	NLP Task: Summarization
24	Nov 14 (Th)	Conversation Agents
25	Nov 19 (Tu)	Ethical Concerns with NLP, Trusted AI and Societal Impact
26	Nov 21 (Th)	Working with LLMs for NLP Tasks - programming, Quiz