CSCE 771: Computer Processing of Natural Language
Lecture 19: Topic Analysis

PROF. BIPLAV SRIVASTAVA, AI INSTITUTE

25TH OCTOBER, 2022

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

Organization of Lecture 19

- Opening Segment
  - Announcements
- Main Lecture
- Concluding Segment
  - About Next Lecture – Lecture 20

Main Section

- Topic Analysis
- LSA
- LDA
- Topic Classification

Recent Classes

Oct 11 (Tu)	Guest Lecture – Dr. Amitava Das: Using lang models to solve NLP tasks
Oct 13 (Th)	
Oct 18 (Tu)	Entity extraction, linking
Oct 20 (Th)	Events extraction, spatio-temporal analysis
Oct 25 (Tu)	Topic Analysis
Oct 27 (Th)	PROJ REVIEW
Nov 1 (Tu)	NLP Task: Sentiment
Nov 3 (Th)	NLP Task: Summarization

Review of Lecture 18

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

Announcements

- Quiz 2 evaluated
- All did well
  - Most marks lost by late submission

Project Assessment Discussion

Course Project – Deadlines and Penalty Rubric

- Project plan **not** ready by Sep 15, 2020 **[-20%]**
  * Project Title
  * Description: motivation and expected output
  * Illustrative Test cases: i.e., Example input / output
  * Data sources:
  * Technique and tools to use:
  * Metric for measuring output
  * How will you collect results
  * Format of report, presentation
  * Time schedule:
- Project report **not** ready by Nov 10, 2022 **[-20%]**
- Project presentations **not** ready by Nov 15, 2022 **[-10%]**

From Class 5

- W1 - Sep 26
- W2 – Oct 3
  - **Review presentation for class: 3 min each – Oct 4, 2022**
- W3 – Oct 10
- W4 – Oct 17
- W5 – Oct 24
  - **Review presentation for class: 3 min each – Oct 27, 2022**
- W6 – Oct 31
- W7 – Nov 7
- W8 – Nov 14
- W9 – Nov 21

Project Rubric

- **Project results** – 60%
  - Working system ? – 30%
  - Evaluation with results superior to baseline? – 20%
  - Considered related work? – 10%
- **Project efforts** – 40%
  - Project report – 20%
  - Project presentation (updates, final) – 20%

- **Bonus**
  - Challenge level of problem – 10%
  - Instructor discretion – 10%
- **Penalty**
  - Lack of timeliness as per announced policy (right) - up to 60%

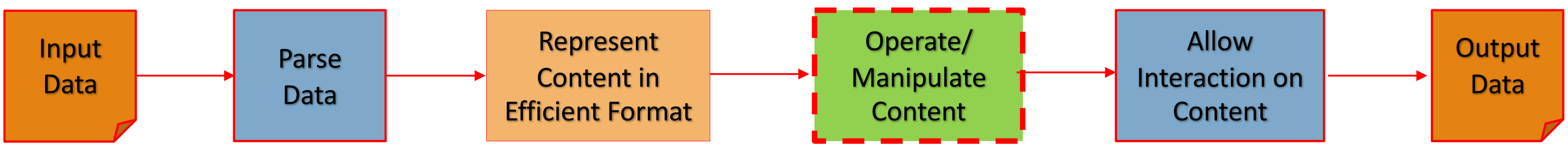
Milestones

- Penalty: **not** ready by Sep 15, 2022 **[-20%]**
- Project report **not** ready by Nov 10, 2022 **[-20%]**
- Project presentations **not** ready by Nov 15, 2022 **[-10%]**

Main Lecture

Topic Detection and Analysis

Statistical patterns identified from textual data



Motivation for Topic Analysis

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

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

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)

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.

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	...
...	...	...	...	...

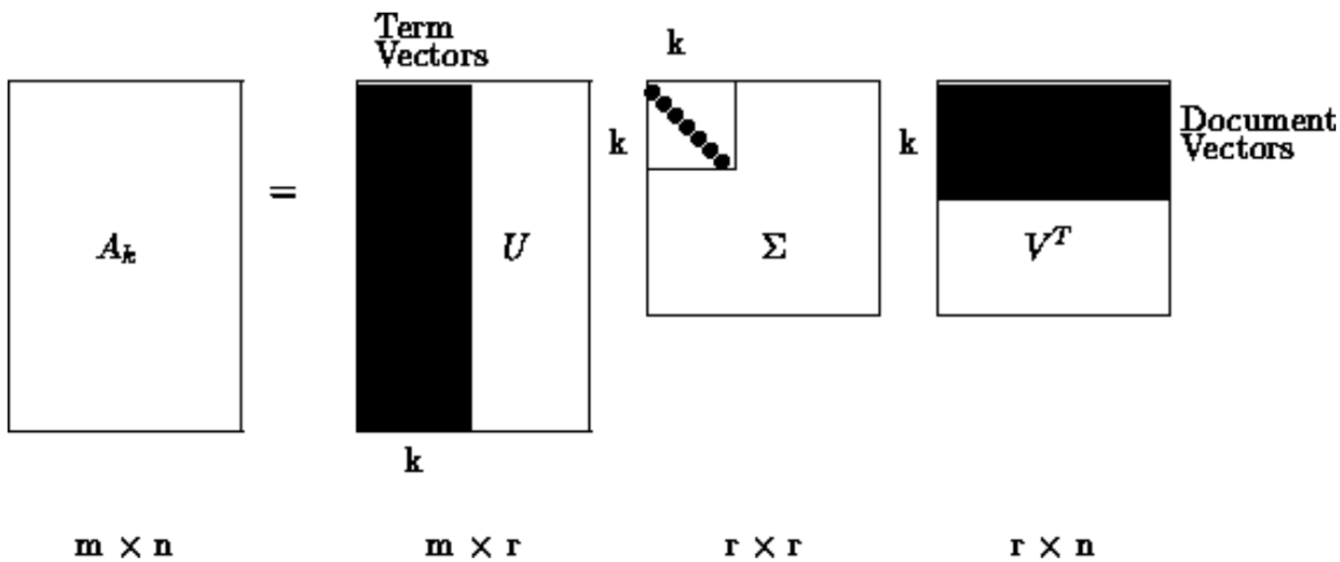
Document – Topic Matrix

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

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

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

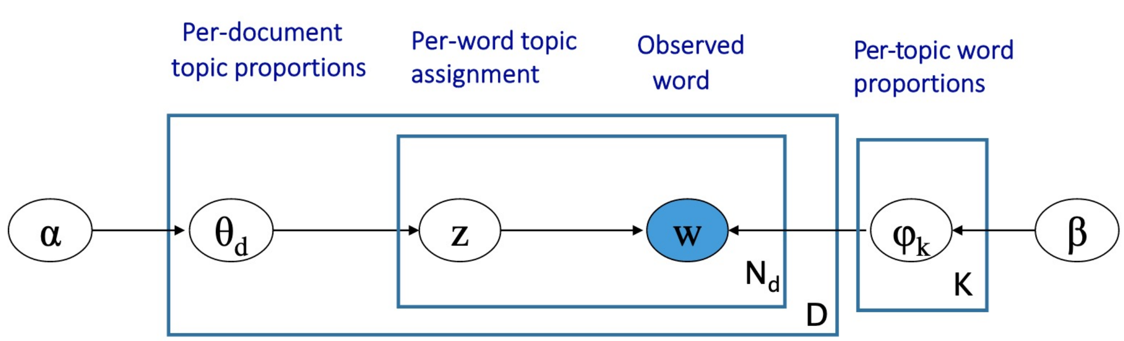
LDA paper: https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf
Blog: https://monkeylearn.com/topic-analysis/,

LDA - Latent Dirichlet Allocation

- **Generative Model**

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



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.

Code Example

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

Libraries:

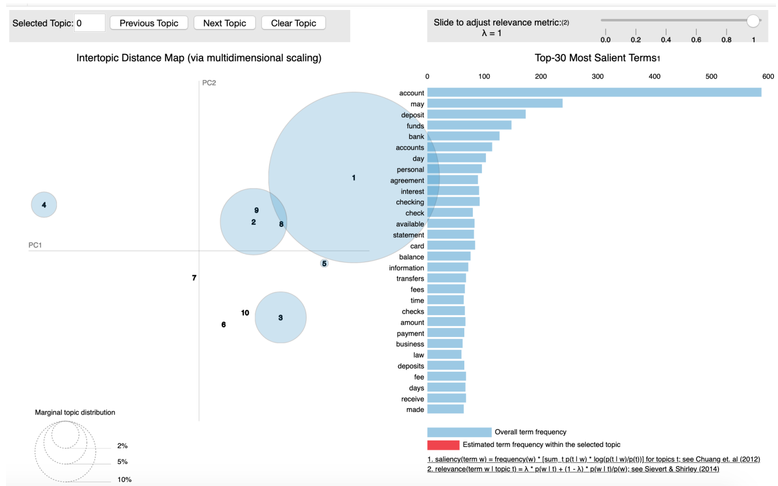
- Gensim: https://radimrehurek.com/gensim/models/ldamodel.html, https://radimrehurek.com/gensim/auto_examples/core/run_topics_and_transformations.html
- Scikit-learn: https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html

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

Visualization of Topics

- LDA: PyLDAVis - https://github.com/bmabey/pyLDAvis



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

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

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

About Next Lecture – Lecture 20

Lecture 20 Outline: Project Review

- Let
  - L1: Review #1 slides repo
  - L2: Review #2 slides repo
- Refer to your slide at L1
- Enhance it with information about
  - Current status
  - Result of actual system on test example
  - Any critical issue
- Put new slide at L2
- In class, give update within 2-3 mins.