





LDA and LSA for Topic Modeling on ORA

Joshua Uyheng
juyheng@cs.cmu.edu
CASOS Center, Institute for Software Research
Carnegie Mellon University

CASOS Summer Institute 2020




Center for Computational Analysis of
Social and Organizational Systems
<http://www.casos.cs.cmu.edu/>



Topic Models

- When we have a large amount of data, we would like to know if they can be grouped in a meaningful way
- “Topics” are a way of thinking of the clustering problem
 - Data instances are “documents”
 - Different documents use different “words”
 - When documents use similar words in similar ways, they might belong to the same “topic”




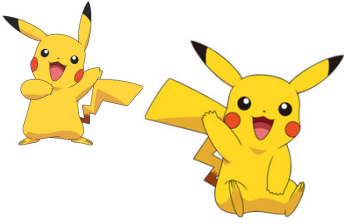
June 2020

2



Carnegie Mellon
ISRI Institute for Software Research

Some examples

Literal texts	More figurative "documents"
Dogs like to run and play.	
Dogs are people's best friend.	
Dogs like to chew on bones.	
Biology is the study of living organisms.	
Chemistry is the study of matter.	
Psychology is the study of human behavior and mental processes.	
One Direction will hold their concert next week.	
Did you buy the One Direction merchandise?	
Harry is my favorite One Direction member.	

CASOS
June 2020

3

Carnegie Mellon
ISRI Institute for Software Research

LSA vs. LDA

- Latent Semantic Analysis or Latent Semantic Indexing
 - Based on matrix factorization
 - Big difference: You can have negative values
- Latent Dirichlet Allocation
 - Based on probabilistic graphical model
 - Big difference: Scores expressed as probabilities
- Both popular

CASOS
June 2020

4



Carnegie Mellon
ISI Institute for SOFTWARE RESEARCH

Latent Semantic Analysis

The diagram shows the decomposition of a matrix X (terms $t \times d$) into three matrices: T_0 ($t \times m$), S_0 ($m \times m$), and D_0' ($m \times d$). The equation is $X = T_0 S_0 D_0'$. Below the diagram, the text defines the matrices and their dimensions.

Singular value decomposition of the term x document matrix, X . Where:

- T_0 has orthogonal, unit-length columns ($T_0^T T_0 = I$)
- D_0' has orthogonal, unit-length columns ($D_0'^T D_0' = I$)
- S_0 is the diagonal matrix of singular values
- t is the number of rows of X
- d is the number of columns of X
- m is the rank of X ($\leq \min(t,d)$)

FIG. 2. Schematic of the Singular Value Decomposition (SVD) of a rectangular term by document matrix. The original term by document matrix is decomposed into three matrices each with linearly independent components.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), 391-407.

CASOS June 2020 5

Carnegie Mellon
ISI Institute for SOFTWARE RESEARCH

Latent Dirichlet Allocation

The graphical model shows a plate for documents (outer plate) containing a plate for topics (inner plate). The outer plate contains parameters α and θ . The inner plate contains parameters z and w . A parameter β is shown outside the plates, with an arrow pointing to w . The variables N and M are also shown.

Figure 1: Graphical model representation of LDA. 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.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 993-1022.

CASOS June 2020 6



Carnegie Mellon
ISRI Institute for SOFTWARE RESEARCH

In practice...

- There is no hard and fast way to decide which model is better
- A large factor in deciding on the quality and interpretation of a topic model is human judgment
- Many will work for general purposes

CASOS
June 2020 7

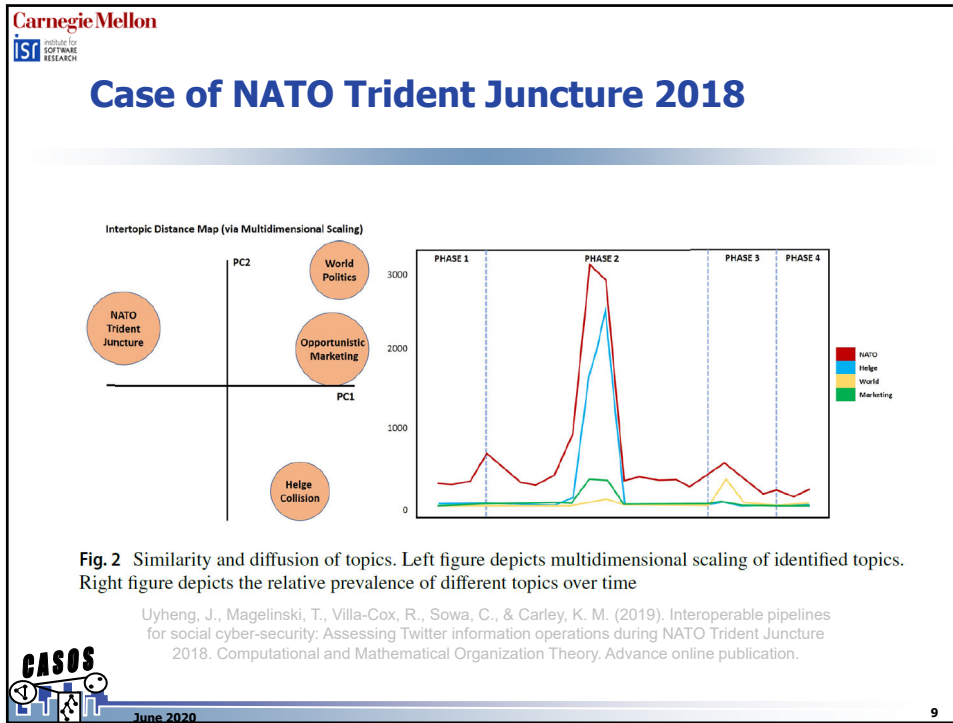
Carnegie Mellon
ISRI Institute for SOFTWARE RESEARCH

In a network setting

1. Documents and words don't have to be literal documents and words
 - People can serve as "documents"
 - Hashtags can serve as "words"
 - Topics can represent tendencies between certain agents to invoke certain hashtags
2. We can visualize multiple kinds of connections between agents and concepts

CASOS
June 2020 8





Topics extracted

Table 1 Summary of topics identified using LDA. While some tweets specifically focused on NATO and events during the Trident Juncture Exercise, others focused more broadly on world politics or exploited public attention surrounding NATO to market various commodities online

Topics	Key words	Sample tweets
NATO Trident Juncture	Nato, exercise, trident juncture, norway, russia, military, maneuver, soldier, large, participate	JORSTADMOEN, Norway—A U.S. Army AH-64 Apache assigned to the 1st Battalion, 3rd Aviation Regiment, 12th Combat Aviation Brigade departs Rena Leir Airfield, Norway, during Exercise Trident Juncture, Nov. 5, 2018. TridentJuncture2018. WeAreNATO
Collision of Helge Ingstad	Russian, ship, photo, october, navy, frigate, sink, tanker, great, take	NATO has to learn from his TridentPUNCTURE. A fire on Canadian frigate HMCS Halifax. The storm nearly sank ship of USNavy GunstonHall. CanadianNavy ship #Toronto lost its turn. A collision with tanker. Navy frigate Norwegian KNM 'Helge Ingstad'
World politics	Go, prepare, need, medium, thank, want, leave, Ukraine, man, injure	There might be more truth here than you can handle... Just sayin' USA Syria Palestine MidEast Yemen Iraq Libya Afghanistan Kushner Zionism Nazi Wahhabism NATO NewWorldOrder UN AIPAC Genie-Energy Trump Obama Hillary GeorgeBush Obama BillClinton
Opportunistic marketing	Say, member, test, do, amp, defense, war, time, big	New on ebay: Arc Touch Wireless USB Receiver Mouse Slim Optical Flat Microsoft Touch Mouse KZ

Uyheng, J., Magelinski, T., Villa-Cox, R., Sowa, C., & Carley, K. M. (2019). Interoperable pipelines for social cyber-security: Assessing Twitter information operations during NATO Trident Juncture 2018. Computational and Mathematical Organization Theory. Advance online publication.

CASOS June 2020 10



Carnegie Mellon
ISI Institute for SOFTWARE RESEARCH

Topics for social cyber-security

Table 4 Summary of bot activity for each topic identified with LDA

Topic	Bot activity	
	Number of bot tweets	(%)
Collision of Helge Ingstad	2385	31.97
NATO Trident Juncture	42512	25.63
World politics	3018	20.30
Opportunistic marketing	3799	7.82

We organize rows by percentage of tweets in each topic associated with a predicted bot. The collision topic featured the highest level of predicted bot activity

Uyheng, J., Magelinski, T., Villa-Cox, R., Sowa, C., & Carley, K. M. (2019). Interoperable pipelines for social cyber-security: Assessing Twitter information operations during NATO Trident Juncture 2018. Computational and Mathematical Organization Theory. Advance online publication.

CASOS
June 2020 11

CASOS

LDA and LSA for Topic Modeling on ORA

Joshua Uyheng
juyheng@cs.cmu.edu
CASOS Center, Institute for Software Research
Carnegie Mellon University

CASOS Summer Institute 2020

Carnegie Mellon Center for Computational Analysis of
Social and Organizational Systems
<http://www.casos.cs.cmu.edu/>

