How Can Topic Modeling Retrieve Information from Academic Research Papers, and Assign the Right Labels of Topic?

Xi Yang, Shiqi Ning, Wen Li, Jianing Sun,
Georgetown University Analytics Program

Motivation

- With the rapid accumulation of various scientific articles which are available online, machine learning methods such as topic modeling have been receiving much attention in those professional fields because of their interpretability.
- By analyzing "How Can Topic Modeling Retrieve Information from Academic Research Papers, and Assign the Right Labels of Topic?", we can better identify the topic classification that are trending research topics in different academic realm in an efficient way.

Data Collection

- Dataset is collected using Springer Nature API, an API that allows developers to access 13 million freely available online content published by Springer.
- Mathematics is chosen as the topic for query.
- 10000 abstracts from academic articles, papers, and book chapters are obtained from query.
- Raw query dataset is JSON format.

Data Pre-Processing

Data Cleaning

- Query results with empty abstracts are cleaned.
- Abstracts written in languages other than English are cleaned out (9766 abstracts are left).
- Content written in latex, latex math equations, and latex symbols (contents between single or double dollar signs) are cleaned out.
- Single character, number, and special signs are cleaned out using regular expression.

Data Processing

- Words in the abstracts are then tokenized and lemmatized.
- Stop words in English, as well as customized stop words for mathematics are cleaned out.
- Bigram and trigram of the abstracts' words are added.

Methodology

Latent Dirichlet Allocation (LDA)

- Created dictionary and corpus with Gensim.
- Trained LDA model with LdaMulticore and traditional LDA.
- Analyzed topics found by LDA model and generated new stop words for better results.
- Generated weighted topics .

Latent Semantic Indexing (LSI)

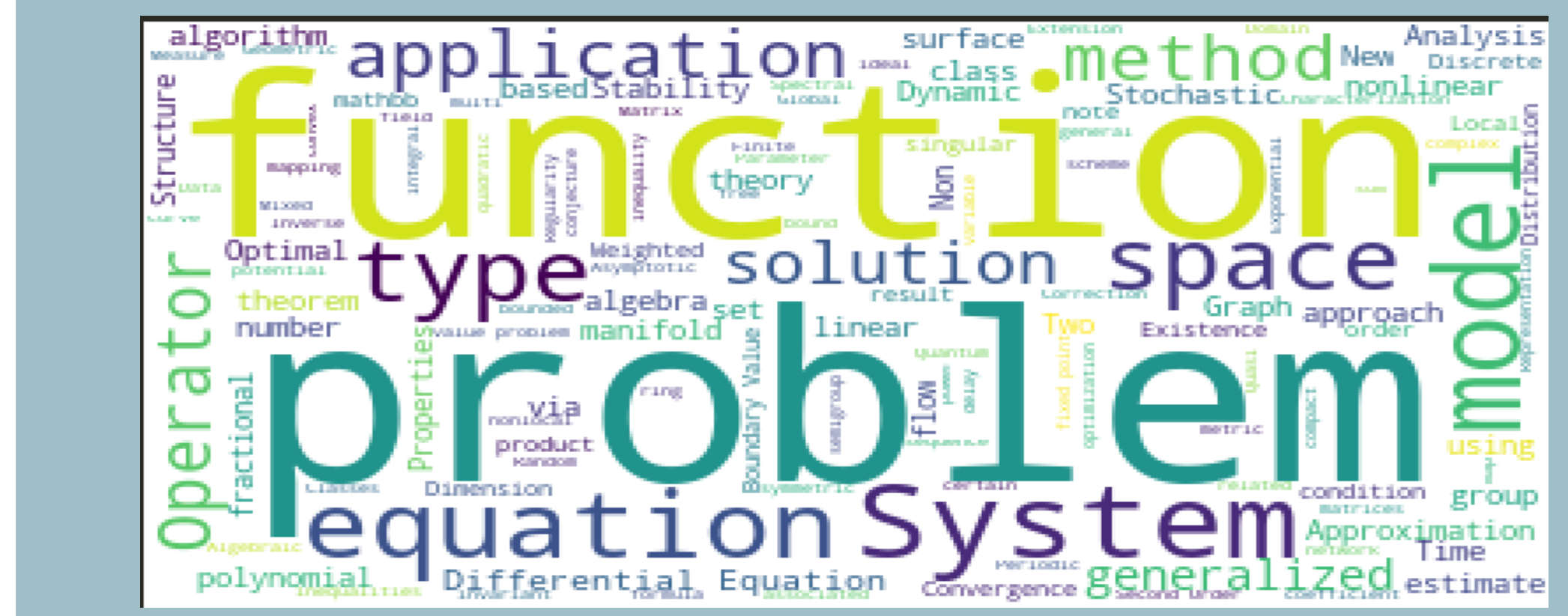
- Generate a document-term matrix of shape having TF-IDF scores using sklearn.
- Reduce the dimensions of the matrix to k (no. of desired topics) dimensions, using singular-value decomposition (SVD).
- Analyze topics found by LSI and do more data-preprocessing (add stop words, lemmatization...)

Important packages and tools

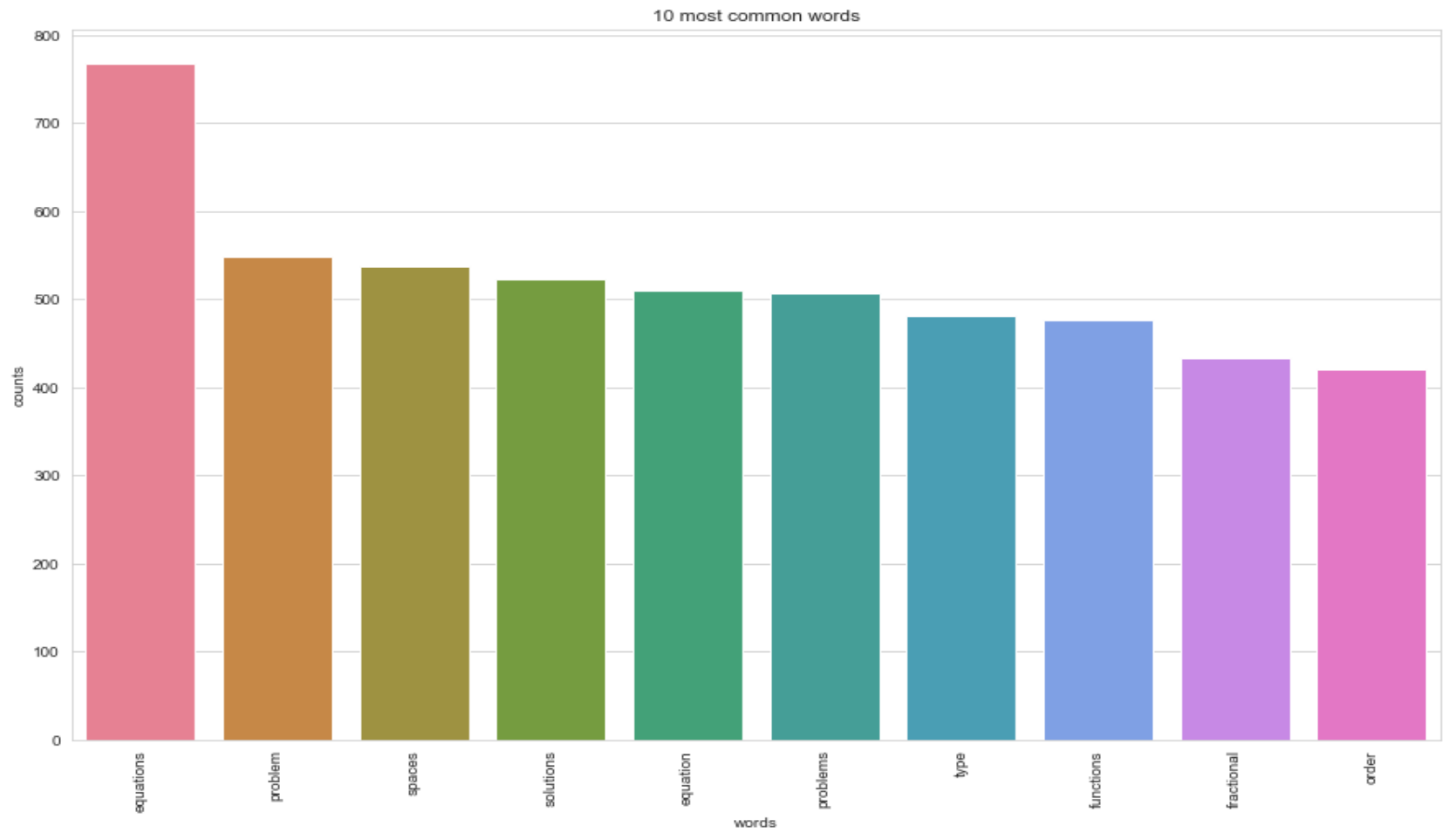
- Natural language related: regex, nltk, genism and sklearn.
- Plotting related: matplotlib, bokeh, wordcloud, and seaborn.
- Data related: requests, pandas, numpy,

Exploratory Data Analysis

Word Cloud for Title



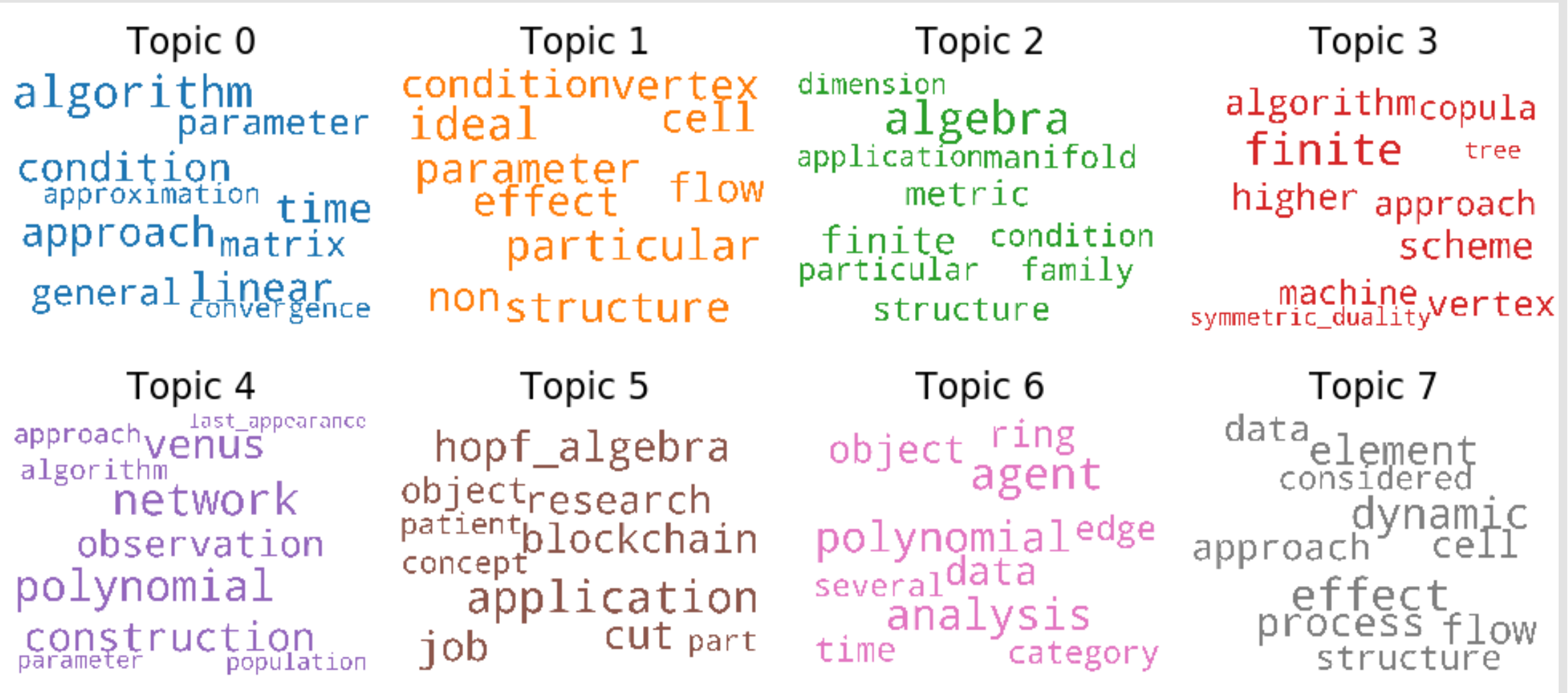
- Operation: Work on transform textual data into vector representation (Bag of Words)
- Convert lists of titles to lists of vectors
- Plot the ten most frequent words based on above



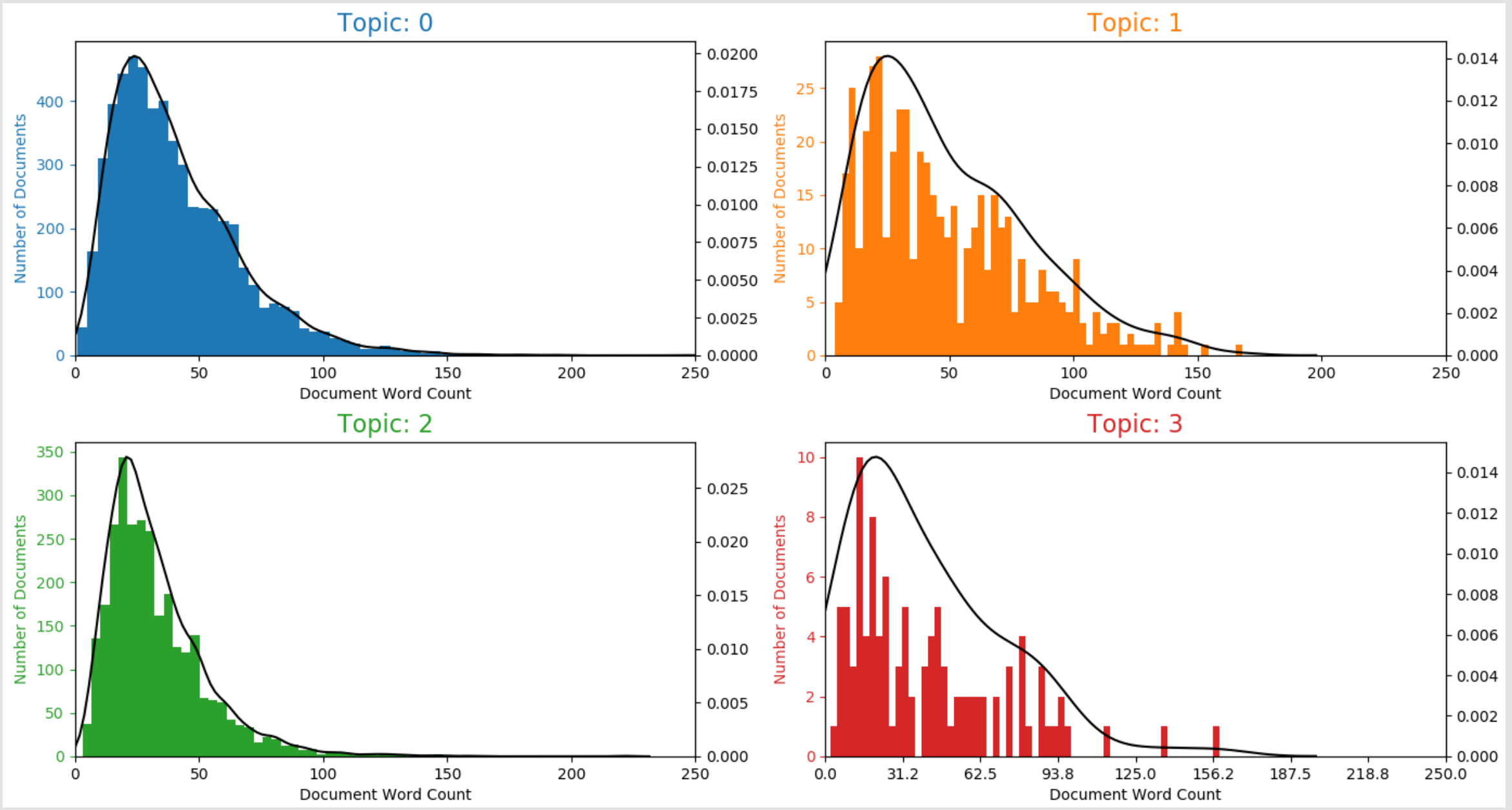
Top ten:
- Equations
- Problem
- Spaces
- Solutions
- Type
- Function
- Factorial
- Order

Results from LDA Model

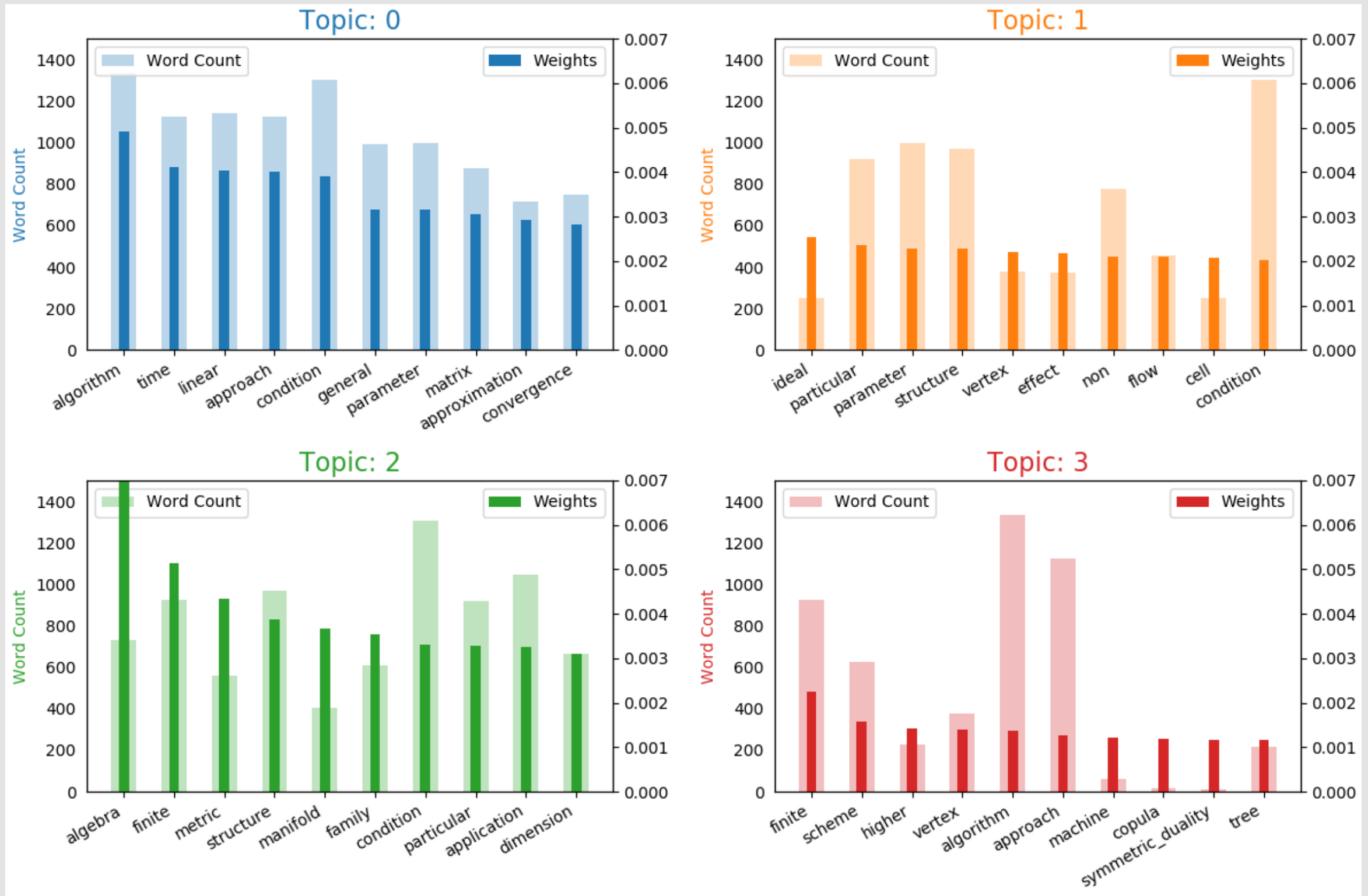
Word Clouds of 8 Topics Built from LDA Model


Topic 0
algorithm parameter condition approximation time approach matrix general linear convergence
Topic 1
condition vertex ideal cell parameter effect flow particular non structure
Topic 2
dimension algebra application manifold metric finite condition particular family structure
Topic 3
algorithm copula finite tree higher approach scheme machine symmetric_duality vertex
Topic 4
approach venus algorithm network observation polynomial construction parameter population
Topic 5
hopf_algebra object research patient blockchain concept application job cut part
Topic 6
object ring agent polynomial edge several data analysis time category
Topic 7
data element considered dynamic approach cell effect process flow structure

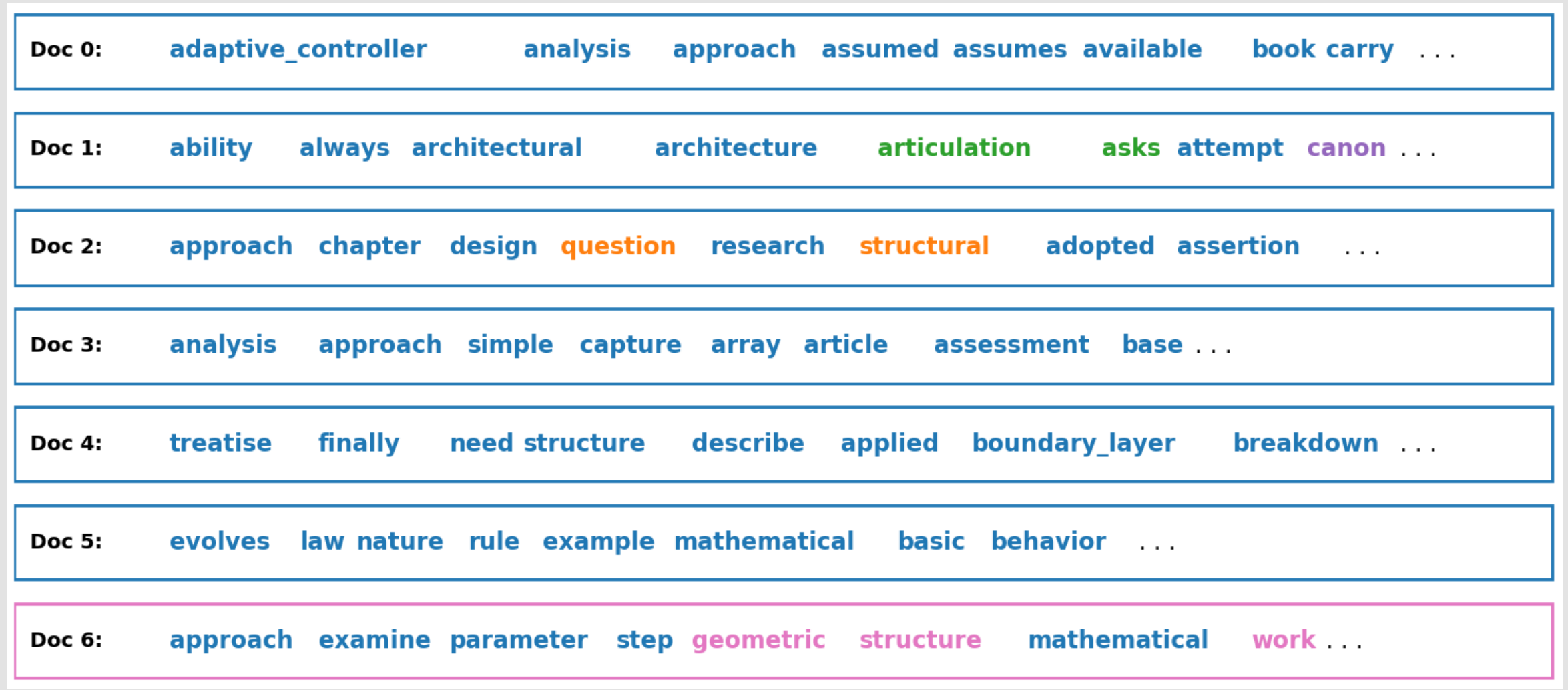
Useful Word Count Distribution of Top 4 Topics



Keywords' Count and Importance Weights of Top 4 Topics



Word Coloring by Topic Samples



Doc 0: adaptive_controller analysis approach assumed assumes available book carry . . .

Doc 1: ability always architectural architecture articulation asks attempt canon . . .

Doc 2: approach chapter design question research structural adopted assertion . . .

Doc 3: analysis approach simple capture array article assessment base . . .

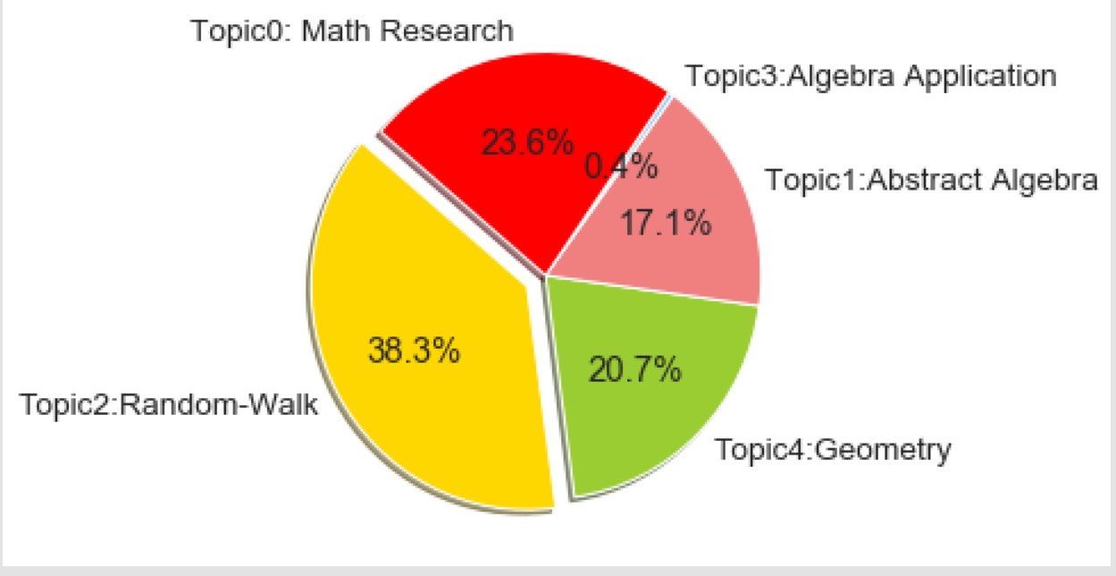
Doc 4: treatise finally need structure describe applied boundary_layer breakdown . . .

Doc 5: evolves law nature rule example mathematical basic behavior . . .

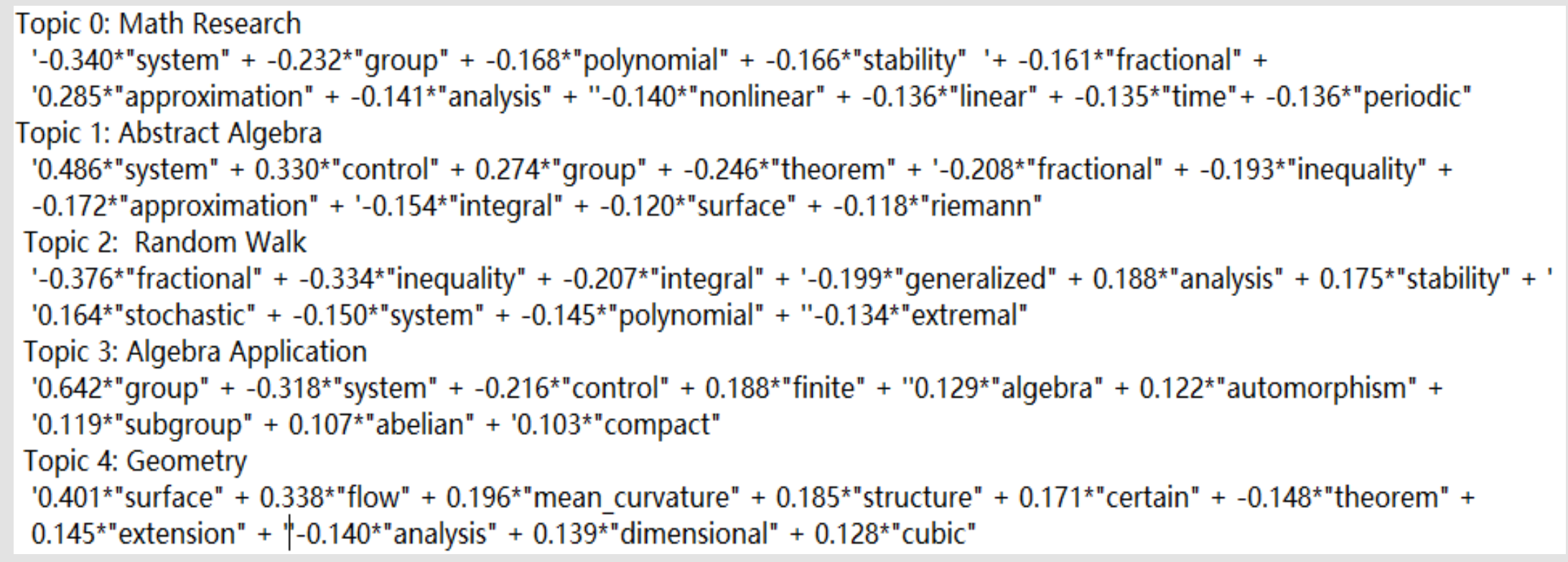
Doc 6: approach examine parameter step geometric structure mathematical work . . .

Results from LSI Model

Pie Chart of 5 Topics Built from LSI Model



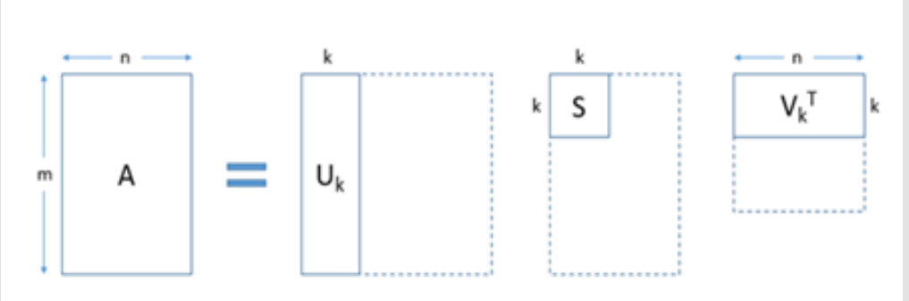
Term-Topic Matrix



Topic 0: Math Research
'-0.340*"system" + -0.232*"group" + -0.168*"polynomial" + -0.166*"stability" '+ -0.161*"fractional" + '0.285*"approximation" + -0.141*"analysis" + ''-0.140*"nonlinear" + -0.136*"linear" + -0.135*"time"+ -0.136*"periodic"
Topic 1: Abstract Algebra
'0.486*"system" + 0.330*"control" + 0.274*"group" + -0.246*"theorem" + '-0.208*"fractional" + -0.193*"inequality" + -0.172*"approximation" + '-0.154*"integral" + -0.120*"surface" + -0.118*"riemann"
Topic 2: Random Walk
'-0.376*"fractional" + -0.334*"inequality" + -0.207*"integral" + '-0.199*"generalized" + 0.188*"analysis" + 0.175*"stability" + ' '0.164*"stochastic" + -0.150*"system" + -0.145*"polynomial" + ''-0.134*"extremal"
Topic 3: Algebra Application
'0.642*"group" + -0.318*"system" + -0.216*"control" + 0.188*"finite" + ''0.129*"algebra" + 0.122*"automorphism" + '0.119*"subgroup" + 0.107*"abelian" + '0.103*"compact"
Topic 4: Geometry
'0.401*"surface" + 0.338*"flow" + 0.196*"mean_curvature" + 0.185*"structure" + 0.171*"certain" + -0.148*"theorem" + 0.145*"extension" + '-0.140*"analysis" + 0.139*"dimensional" + 0.128*"cubic"

$$A = USV^T$$

SVD :



Conclusion

- Latent Dirichlet Allocation (LDA) model is useful to perform topic modelling task on abstracts of academic articles and papers.
- By tuning LDA model and ameliorate data preprocessing workflow, different math-related keywords are gathered into different topics: linear algorithm (Topic 0), geometry (Topic 1), algebra with finite method (Topic 2), and etc.
- Latent Semantics Indexing (LSI) helps to figure out the hidden concepts or topics behind the words. The documents are gathered into 5 different topics, 2 of which are combined manually.

References

[1] R. Rubin. Foundations of Library and Information Science. 2nd ed. New York: Neal-Schuman, 2004.

[2] Michael Paul and Roxana Girju. Topic Modeling of Research Fields: An Interdisciplinary Perspective, 2009 Retrieved from: https://www.aclweb.org/anthology/R09-1061/

[3] Akira Murakami. Getting to know your corpus: applying Topic Modelling to corpus of research articles. University of Cambridge, 2016

[4] Gensim topic modeling for humans API Reference. Retrieved from: https://radimrehurek.com/gensim/apiref.html