Analysis of research trends using Latent Dirichlet Allocation for geologic subdisciplines in South Korea

Taeyong Kim[1] [xodyd123123@pukoyng.ac.kr] and **Minjune Yang**[2*] [minjune@pknu.ac.kr]

[1]*Division of Earth Environmental System Sciences, Pukyong National University, Busan 48513, South Korea*

[2]*Department of Earth and Environmental Sciences, Pukyong National University, Busan 48513, South Korea*

Environmental Hydrogeology Lab

1. Study rationale

Since the mid-twentieth century, geology in South Korea has considerably advanced as a scientific discipline. Over the past few decades, geology has interacted with physical or engineering viewpoints. So, modern geology needs to be interpreted with an interdisciplinary perspective. This study aimed to classify geology's academic subdisciplines in Korean and analyze the evolutionary trend of each subdiscipline in South Korea for 54 years from 1964 through 2019.

2. Latent Dirichlet Allocation (LDA)

- Generative probabilistic model^
- Using a Dirichlet distribution, LDA can Build a topic per document model and words per topic model.
- Assign each word in a document to one of $\boldsymbol{k}$ **topics randomly**
- For each topic $\boldsymbol{k}$ assign the word $\boldsymbol{w}$: $P(w|k) * P(k|d)$

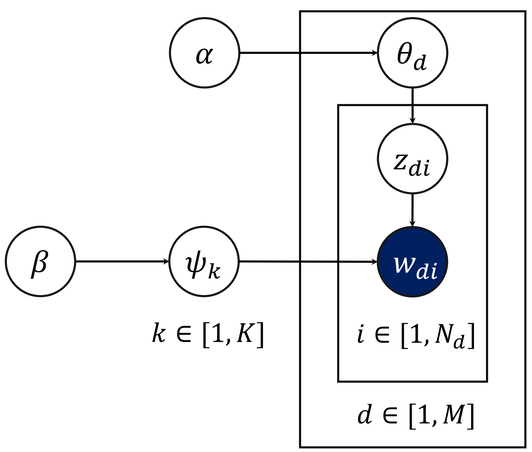

Fig. 1. Graphical model representation of LDA.

Table 1. Notations of variables and parameters in LDA

Notation	Description
d	Index of documents
k	Index of topics
i	Index of words
j	Index of journals
t	Index of years
α	Dirichlet prior on the per-document topic distributions (hyperparameter)
β	Dirichlet prior on the per-topic word distributions (hyperparameter)
θ_d	Topic distribution of document d
ψ_k	Word distribution of topic k
w_{di}	Word i in corpus of document d
z_{di}	Topic assignment for word w_{di} from document d
K	Number of topics
M	Number of documents
N_d	Number of words in document d

3. Data set and Preprocessing

- **Use 11,148 of Korean geological title, keyword, and abstract for LDA**
- **Tokenization**
- **Stop words**
  - [korea, korean, lt, gt, sup, sub…]
- **Remove non-English accounts**
- **Stemming**
  - [groundwat, earthquak, heavi, concentr, format….]
- **Remove words that appears less than 5 times in whole corpus**

Table 2. publication year and number of articles collected by geological journals in South Korea

Journal titles	First issue (year)	Number of articles
Economic and Environmental Geology	1968	2758
Journal of the Geological Society of Korea	1964	2044
The Journal of Engineering Geology	1991	1109
Journal of the Mineralogical Society of Korea	1988	1039
Journal of Soil and Groundwater Environment	1994	1002
Geophysics and Geophysical Exploration	1998	830
Geosciences Journal	2002	786
The Journal of the Petrological Society of Korea	1985	668
Mineral Science and Industry	1988	472
Journal of the Paleotological Society of Korea	1985	382
Korean Journal of Petroleum geology	1993	58

4. Results

Table 3. Results of LDA. In this study, we chose the number of topics K = 20

Topic #1	Topic #2	Topic #3	Topic #4	Topic #5
granit, rock, magma, biotit	sp, talc, asbestos, cave	volcan, island, tuff, erupt	age, zircon, granit, massif	fault, zone, fold, deform
Topic #6	**Topic #7**	**Topic #8**	**Topic #9**	**Topic #10**
bentonit, cement, acid, leach	adsorpt, uranium, arsen, radon	metal, heavi, mine, pb	soil, remove, contamin, remedi	shale, format, sediment, deposit
Topic #11	**Topic #12**	**Topic #13**	**Topic #14**	**Topic #15**
ga, inject, storag, basin	fault, earthquak, basin, peninsula	ore, miner, deposit, vein	Groundwat,landslid, soil, ranfal	stone, sediment, climat, ice
Topic #16	**Topic #17**	**Topic #18**	**Topic #19**	**Topic #20**
water, groundwat, co2, aquif	Slope, rock, Fractur, strength	Miner, phase, X-ray, oxid	Resourc, tunnel, Well, system	model, resist, data, seismic

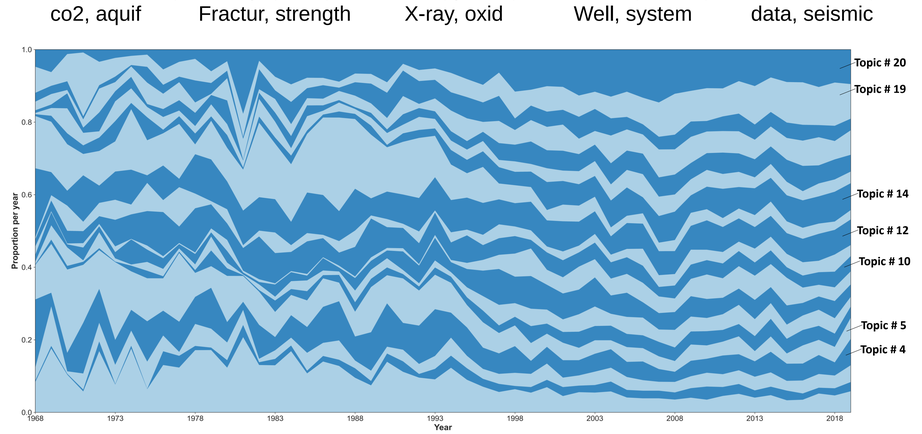

Fig. 3. Geological topic distribution over time in South Korea.

5. Conclusions

- The highest proportion of Korean geological topic in the past(~1990) is **Topic #13.**
- The hottest topic in Modern geology is **Topic #19** and **Topic #20.**
- The results of this study fill an important gap in understanding the research trends of geologic subdisciplines in South Korea, showing their emergence, growth and diminution.

6. Principal references

- Lijun Sun, Yafeng Yin. 2017. Discovering themes and trends in transportation research using topic modeling. Transportation Research Part C: Emerging Technologies, Pages 49-66.
- David M. Blei, Andrew Y. Ng and Michael I. Jordan. 2003. Latent Dirichlet Allocation. Journal of Machine Learning Research, 3:993-1022.
- McCurley, K. L. and J. W. Jawitz. 2017. Hyphenated hydrology : Interdisciplinary evolution of water resource science, Water Resour. Res.,53, s2972–2982.