

Machine-Learning Based Identification of Emerging Research Topics Using Research & Development Administrative Data

Eric J. Oh, Kathryn Linehan, Joel Thurston, Stephanie Shipp
(Social and Decision Analytics Division)

John Jankowski, Audrey Kindlon
(National Center for Science and Engineering Statistics)

FCSM 2021 - November 3, 2021

This work is supported by National Center for Science and Engineering Statistics (49100420C0015)

Motivation

- Around 21% of all research and development (R&D) funding in the United States is provided by the federal government¹
 - What specific research areas does this public funding support?
- NCSSES Federal Funds Survey and Federal Support Survey
 - Provides information on funding by agency and broad research fields
- Can we use other types of data to supplement the surveys?

¹Table 3; <https://nces.nsf.gov/pubs/nsf21324>; 2021.

Federal RePORTER



Federal RePORTER

- Searchable database of scientific awards from agencies
 - HHS, NSF, USDA, NASA, DOD, VA, ED, EPA (most complete for HHS, NSF)
 - > 1 million grants reported in FY 2008-2019
- Includes grant abstracts and metadata
- Agencies submit project information to Federal RePORTER

Messy administrative data

General cleaning

- Fill in missing project start date
- De-duplicate abstracts

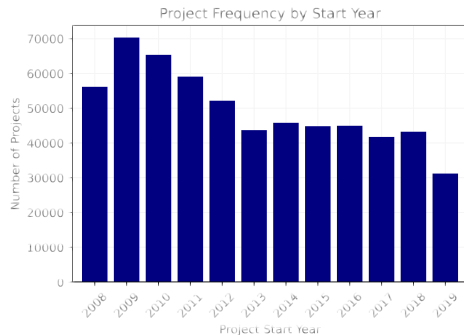
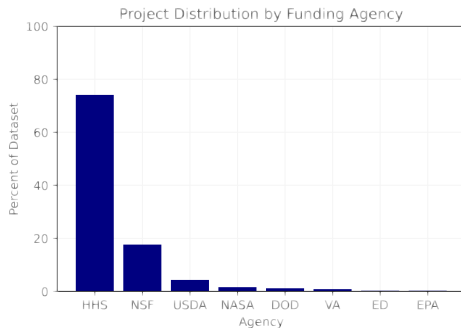
Modeling purposes

- Remove non project-specific words (e.g., "description", PI names)
- Tokenization, lemmatization, stop word removal, and the addition of bi-grams and tri-grams
- Remove most frequent words²

²Schofield; A.; Magnusson; M.; Thompson; L.; & Mimno; D. (2017). Understanding text pre-processing for latent Dirichlet allocation.

Summary of data

690,814 abstracts for unique projects in FY 2008-19



Topic modeling

- Unsupervised machine learning method for discovering latent “themes” or “topics” in text data
- Topics are defined by a group or cluster of words that share semantic relationships
- Example topics:

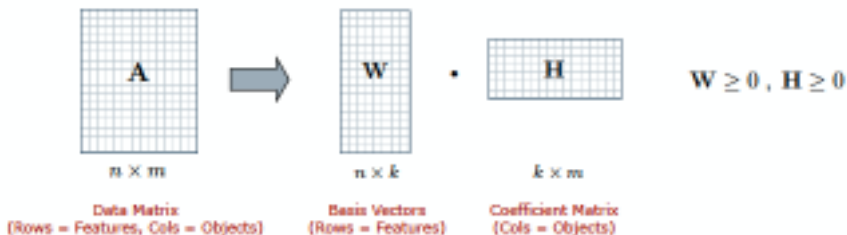
aging
age
older
adult
related
decline
life
aged
lifespan
cognitive

management
soil
crop
production
agricultural
practice
pest
farm
economic
farmer

imaging
image
mr
resolution
optical
mr
pet
contrast
microscopy
probe

Non-negative matrix factorization (NMF)

- Given non-negative data matrix **A**, NMF finds a k -dimensional approximation in terms of non-negative factors **W** and **H**

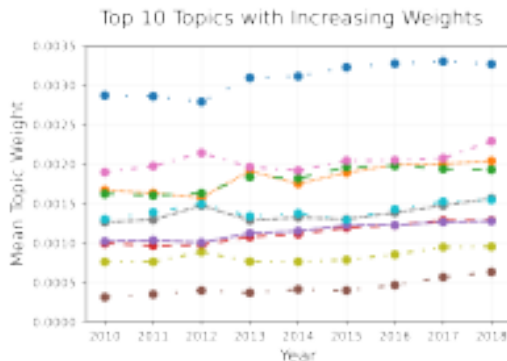


$\{\text{features, objects, basis vectors}\} = \{\text{words, documents, topics}\}$

- Each document is represented by a linear combination of topics and weights (coefficient matrix)
- k , the number of topics, must be chosen by the analyst

Figure source: Dynamic Topic Modeling via Non-negative Matrix Factorization by Derek Greene, slide 3

NMF results - increasing weights



- FR31: model, theory, problem, method, computational
- FR44: student, science, stem, school, undergraduate
- FR45: system, technology, device, design, develop
- FR33: network, social, wireless, communication, node
- FR15: data, analysis, statistical, database, management
- FR1: ad, alzheimer, tau, dementia, pathology
- FR28: intervention, behavior, treatment, social, behavioral
- FR3: aging, cognitive, age, memory, older
- FR46: training, trainee, faculty, career, mentor
- FR34: neuron, circuit, neural, neuronal, motor

Figure: Federal RePORTER, FY 2008-19; University of Virginia, Social and Decision Analytics Division computations. Due to incomplete data from 2019, results include only 2010-18 abstracts

NMF results - decreasing weights

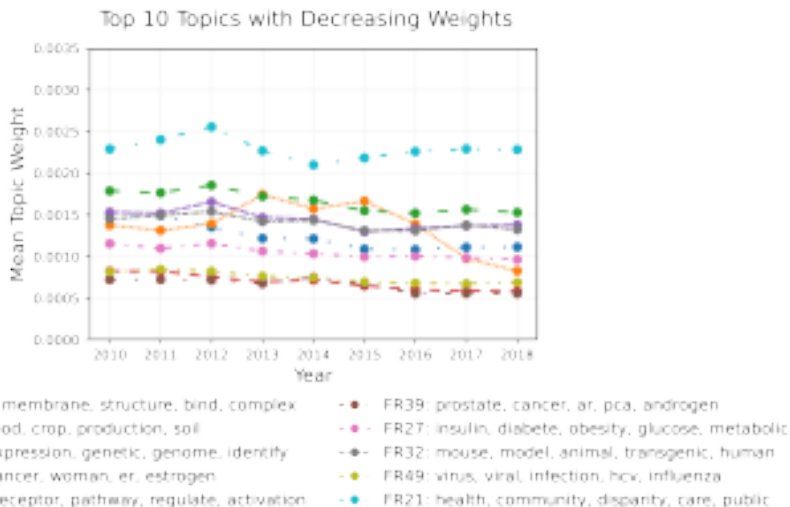


Figure: Federal RePORTER, FY 2008-19; University of Virginia, Social and Decision Analytics Division computations. Due to incomplete data from 2019, results include only 2010-18 abstracts

Identifying topics within research areas

- We want to be able to discover topics within specific research areas of interest (e.g. coronavirus, artificial intelligence)
- Fitting topic models on all Federal RePORTER abstracts is too broad - need to narrow down somehow

How can we select the abstracts from Federal RePORTER that relate to a given research area?

Embeddings

Embeddings are representations of words as vectors, with semantically similar words resulting in similar vectors

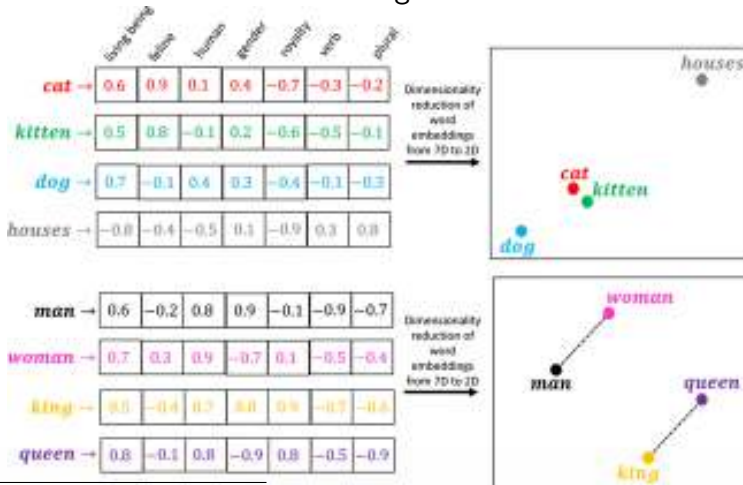


Figure source: <https://medium.com/@hari4om/word-embedding-d816f643140>

Bidirectional Encoder Representations from Transformers (BERT)

- State of the art model for creating context aware embeddings
 - Vectors for the word “running” in “They are running a company” and “They are running a marathon” are different
- BERT provides pre-trained embeddings for use in various tasks
 - We can download the model and immediately calculate the embeddings for our corpus
- Standard BERT used Wikipedia corpus to train the model
 - Many extensions of BERT for task-specific purposes (e.g., trained on scientific articles to get better embeddings for STEM words)

Cosine similarity

- Embeddings are constructed such that semantic similarity can be captured with the cosine similarity

$$\text{similarity}(\mathbf{A}, \mathbf{B}) = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \cdot \|\mathbf{B}\|}$$

where \mathbf{A} and \mathbf{B} are vectors.

- We want to compare the semantic similarity of longer text (e.g., abstracts and search queries)
- Sentence BERT accomplishes this by constructing context-aware sentence embeddings
- Using BERT embeddings gives us a score to rank relevance/similarity to a research area
 - What query/document do we use to compare to the abstracts?

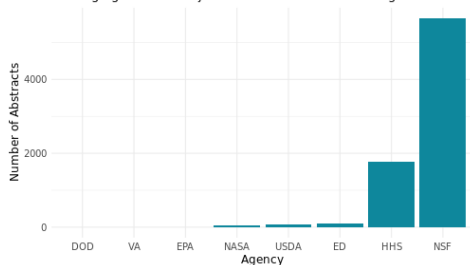
Approach for artificial intelligence (AI)

- Scrape and extract the text from the Wikipedia page for “artificial intelligence”
- Construct embeddings for each sentence in AI Wiki and Federal RePORTER abstracts
- Calculate the cosine similarity for each pairwise combination of sentences
- Take the mean of the top 10 scores for each sentence in an abstract: average the scores across sentences to obtain the similarity score for an abstract
- Use a cutoff of 2.5 SDs above the mean abstract similarity score to classify as AI

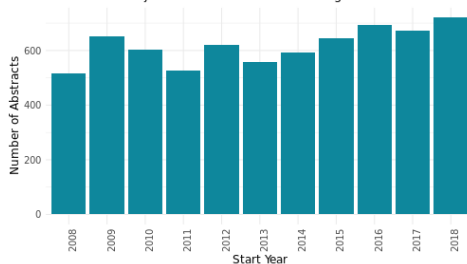
AI corpus summary

7,658 abstracts identified as AI

Funding Agencies of Projects Related to Artificial intelligence



Start Year of Projects Related to Artificial intelligence



Initial AI topic modeling results

Topic	Top 5 words
1	algorithm, optimization, computational, solution, complexity
2	biomedical, computational, phenotype, biology, biological
3	brain, neural, neuroscience, circuit, neuron
4	child, cognitive, social, developmental, mathematical
5	decision, choice, agent, uncertainty, value
6	engineering, engineer, education, nsf, student
7	language, word, processing, linguistic, text
8	learning, learn, learner, deep, machine_learning
9	network, social, agent, deep, dynamics
10	robot, human, robotics, task, robotic
11	science, social, scientific, workshop, innovation
12	software, user, code, developer, computing
13	statistical, dimensional, inference, variable, estimation
14	student, stem, teacher, thinking, skill
15	visual, object, image, vision, recognition

Discussion

- Federal RePORTER can yield very interesting results about research topics that have been funded over time
- Topic Modeling is able to discover those latent topics from just the grant abstracts
- Wikipedia and BERT embeddings can extract the relevant abstracts for artificial intelligence - hopefully, this can apply for other research areas

References

- ① Devlin, J., Chang, M. W., Lee, K., Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
- ② Lee, D. D., Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. Nature, 401(6755), 788-791.
- ③ Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).
- ④ Reimers, N., Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084.