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 government\(^1\)
  - What specific research areas does this public funding support?

- NCSES 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?

\(^1\)Table 3; https://ncses.nsf.gov/pubs/nsf21324; 2021.
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 words

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:
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 $H$

$$A \rightarrow W \cdot H \quad W \geq 0, \ H \geq 0$$

- Each document is represented by a linear combination of topics and weights (coefficient matrix)

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

- $k$, 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

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

**Example Embeddings**

- **cat** → \(0.6\ 0.9\ 0.1\ 0.4\ -0.7\ -0.3\ -0.2\)
- **kitten** → \(0.5\ 0.8\ -0.1\ 0.2\ -0.6\ -0.5\ -0.1\)
- **dog** → \(0.7\ -0.1\ 0.4\ 0.3\ -0.4\ -0.1\ -0.3\)
- **houses** → \(-0.8\ -0.4\ -0.5\ 0.1\ -0.9\ 0.3\ 0.8\)
- **man** → \(0.6\ -0.2\ 0.8\ 0.9\ -0.1\ -0.9\ -0.7\)
- **woman** → \(0.7\ 0.3\ 0.9\ -0.7\ 0.1\ -0.5\ -0.4\)
- **king** → \(0.5\ -0.4\ 0.7\ 0.8\ 0.9\ -0.7\ -0.6\)
- **queen** → \(0.8\ -0.1\ 0.8\ -0.9\ 0.8\ -0.5\ -0.9\)

**Figure source:** [https://medium.com/@hari4om/word-embedding-d816f643140](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

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

where \(A\) and \(B\) 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

Number of Abstracts

Agency: DOD, VA, EPA, NASA, USDA, ED, HHS, NSF

Number of Abstracts

Start Year: 2008 to 2018
Initial AI topic modeling results

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


