Cedric Herman
Mentor: Srdjan Santic
Career coach: Allison Matthews

Springboard
January 2018 cohort

Capstone Project 2:
Topic Modeling on National Science Foundation Award

Contents

1) Scope

Fundamental research enables breakthrough in science and technology which in turns benefits all of us in our daily life. Research is expensive and thus funding is critical for any team of individuals wanting to give birth to their blooming ideas. Thanks to the National Science Foundation (NSF), a US federal agency, people can apply for research grants in any science and engineering fields except for medical sciences. Each award is documented and available on the NSF website per year tracing back to 1959. This project is based on the past 50 years period from 2017. Each year has its own zipped folder which contains about 10,000 xml file, one per award. Data size is about 1.5GB but it has ~400k award! (xml files).

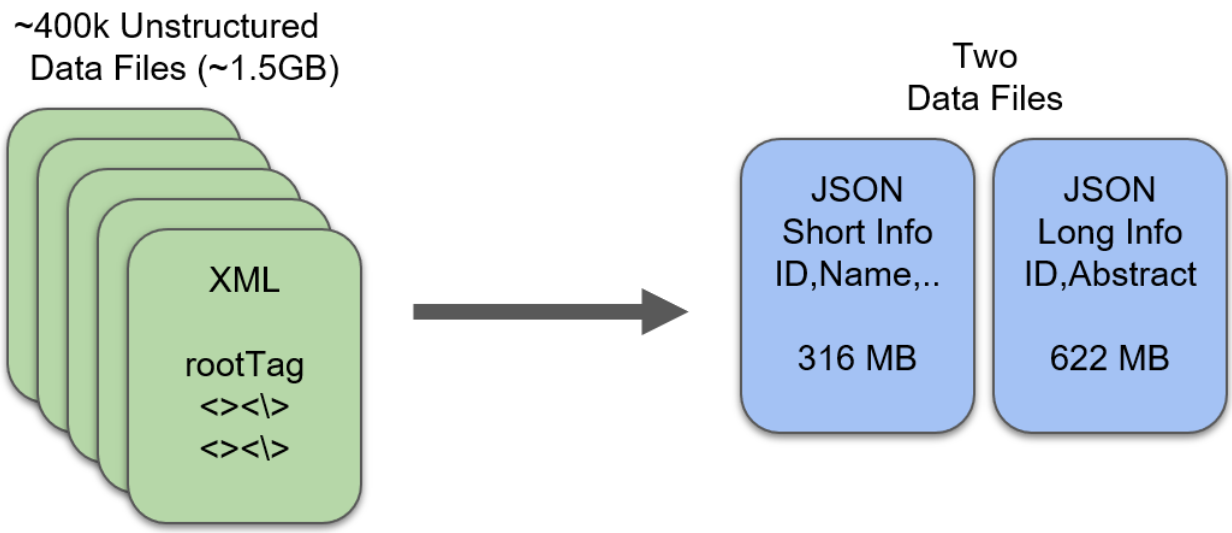
The challenge is to determine whether we can classify research topic accurately based on abstracts and other information so that we can point out what science was the most supported over the years.

2) Data Retrieval

In order to programatically download data from the NSF website, python library requests was used to connect to each data folder arranged by year. Once a zipped data folder is in memory, we will write to disk and extract its content thanks to zipfile library (part of standard library).

At this point we have all awards sorted by year ready for further processing.

Because of the high award count, we will have to parse ~400k file each time we want to carry out an analysis. IO operations are computationally costly thus it would take a lot of time just opening and closing files. Therefore, we will compute necessary information down to 2 csv files as depicted below.



We will capture all abstracts in one JSON file and all other information in another JSON files. The xml schema is available online. Using beautiful soup to extract tag information from xml files and multiprocessing to parallellize extraction on several files at once, it takes about 10 min only to transform our awards into 2 JSON files. JSON format is appropriate because the number of authors for each award varies for instance. Using RDBMS, it would take multiple tables to achieve our goal.

3) Data Wrangling

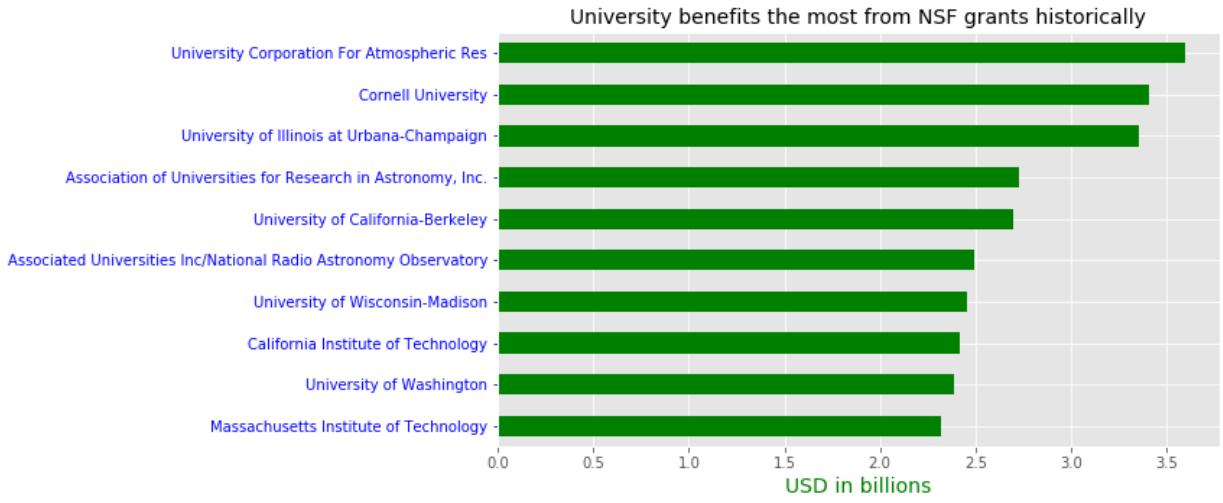
We are mostly interested in abstracts but we also need topic which at the highest level are called directorate by the National Science Foundation (NSF). There are 7 directorates:

1. Directorate for Biological Sciences
2. Directorate for Computer & Information Science & Engineering
3. Directorate for Education & Human Resources
4. Directorate for Engineering
5. Directorate for Geosciences
6. Directorate for Mathematical & Physical Sciences
7. Directorate for Social, Behavioral & Economic Sciences

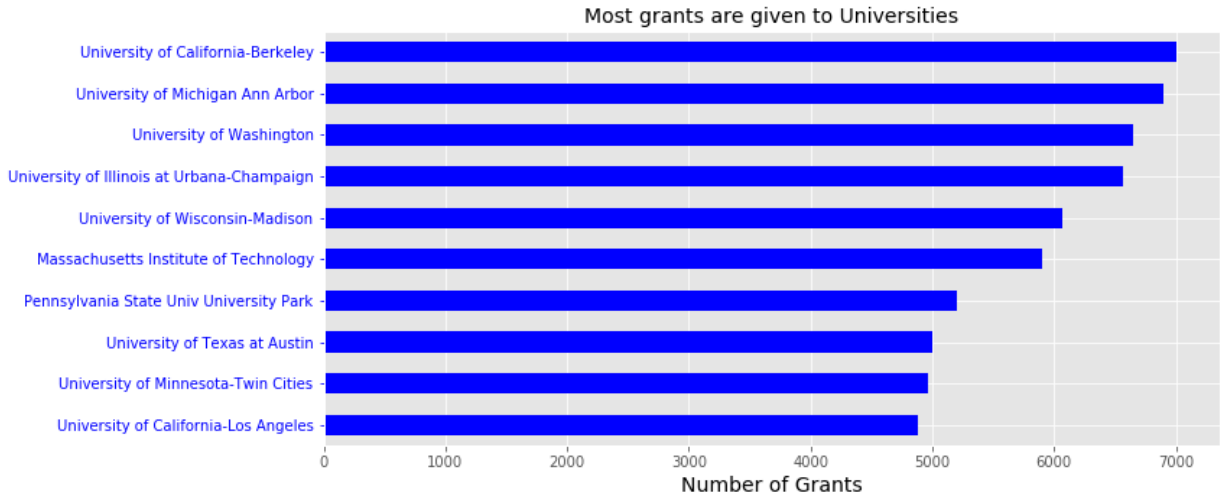
Directorate names have a multitude of abbreviations so one task is to consolidate all different spelling to a unique string for each directorate.

4) Data Exploratory Analysis

Historically, as far back as 1960, universities have received the most fund from the National Science Foundation (NSF). There are few instances where businesses were rewarded by the NSF.

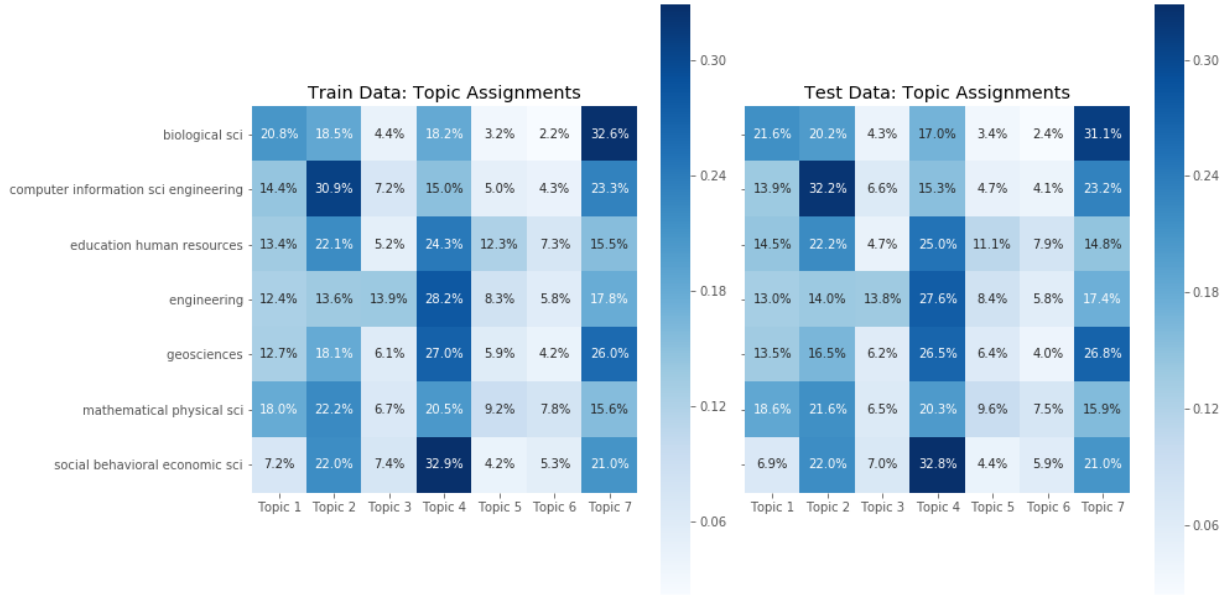


Interestingly, Cornell University does not appear in the top 10 based on the number of grants. University of Michigan comse second in terms of number of grants but it does not appear on total funds received.



5) Topic Modeling

Given our list of directorate, it would be interesting to know if we can effectively classify each field or topic based on their abstract information. The confusion matrix below shows we are having a hard time to differentiate between directorate. It simply means there is a lot of overlap between science fields.



On the other hand, thanks to the LDA visualization available as a standalone interactive html page. One can explore the top 30 words that is the most representative of each topic.