NATIONAL SCIENCE FOUNDATION: TOPIC MODELING

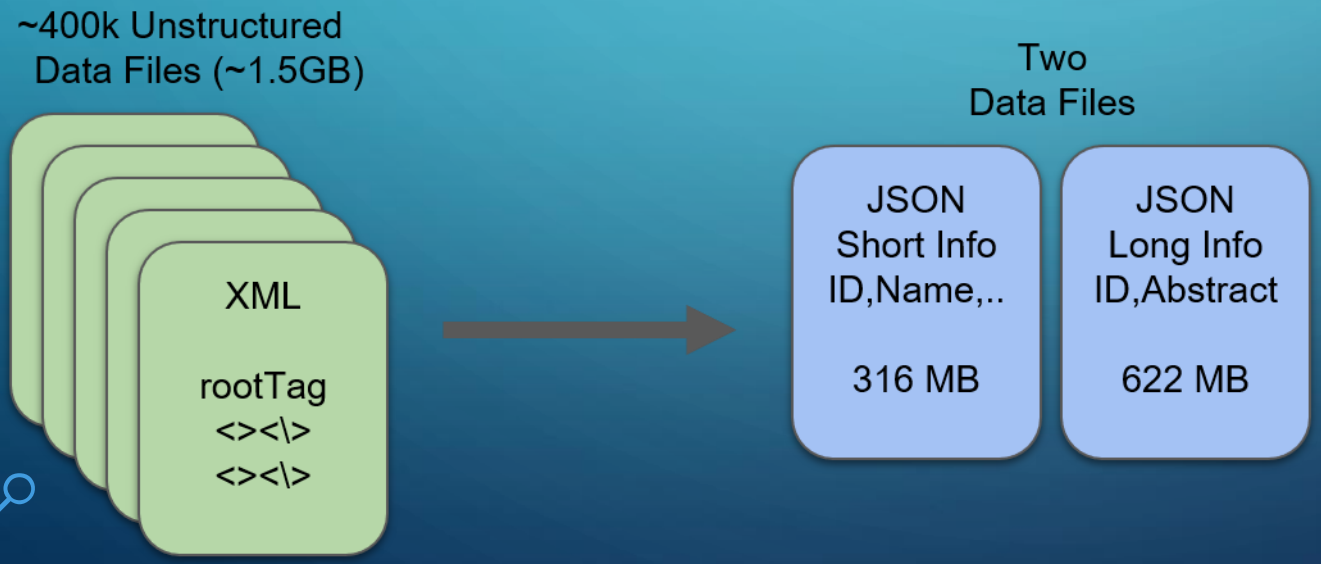
THE NATIONAL SCIENCE FOUNDATION IS A U.S. FEDERAL AGENCY THAT SUPPORTS ALL FIELDS OF SCIENCE & ENGINEERING EXCEPT MEDICAL SCIENCES. EACH YEAR, THEY DISTRIBUTE ABOUT 10,000 AWARDS.

PROBLEM DESCRIPTION

- Our goal is to predict what field each award belongs to thanks to its abstract. Each award has an short summary (abstract) of what it is trying to achieve along with institution, investigators, dollar amount,… (see Award.xsd for complete list)
- We can also help describe what kind of research each field is focused on by extracting the most relevant terms.
- Science fields are organized in 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

~400K AWARD TO PROCESS

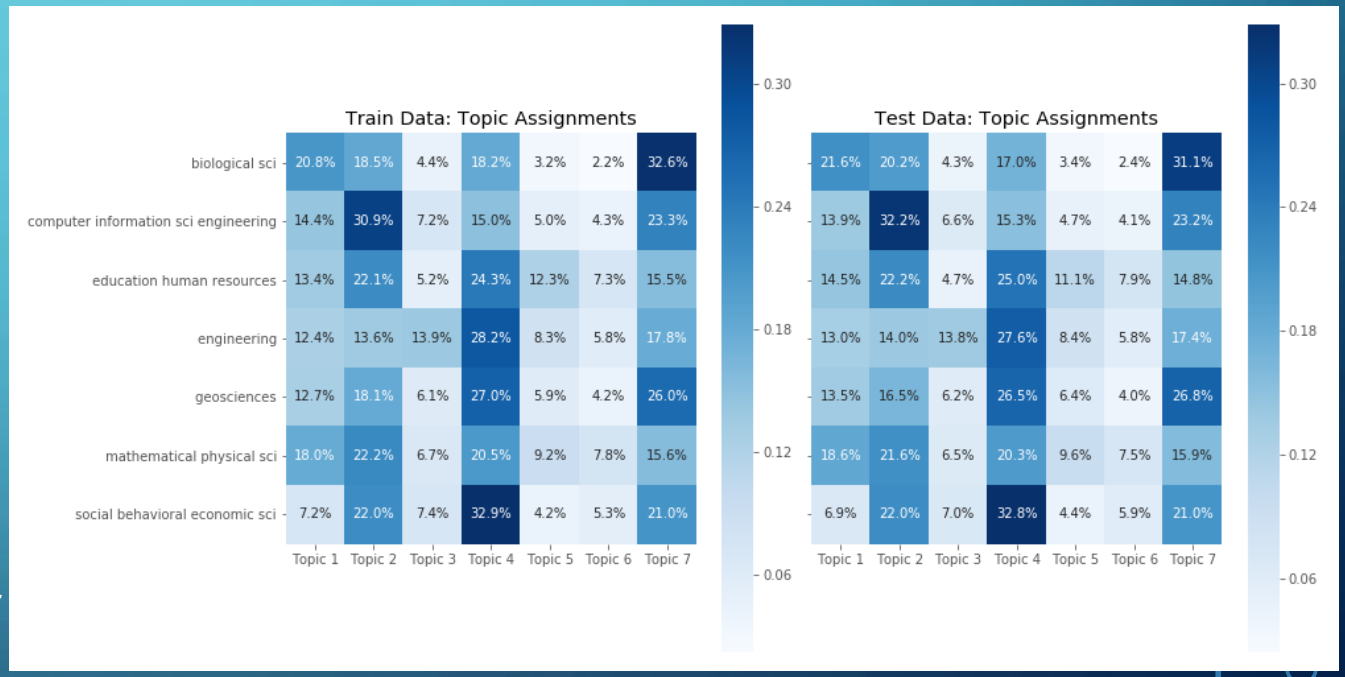
- Awards available for download on the NSF website as far back as 1959.
- A pipeline was built to consolidate all 400k xml files to two json files. JSON object can accommodate varying number of attribute (i.e. there could be one or more investigators) and faster to parse.
- One JSON file dedicated to abstracts and another one for short information to keep their respective size low



TOPIC MODELING

- Text cleaning steps:
  - Remove html tag (e.g. <a>)
  - Remove links starting with http/https
  - Set to lower case
  - Remove stop words
  - Remove accents
- Create document-term matrix
- Use Generative Probabilistic algorithm Latent Dirichlet Analysis (LDA) with 7 components.

NOTE: LDA surpasses LSA and pLSI as described in paper by David Blei and al.



Social sciences and biology are topics that distinguish themselves the most from all other topics.