Exploring Trends in CMU Grant Data
Kyra Balenzano, Melissa Dy, Michael Li, and Veda Lin
Project Advisor: Peter Freeman  Project Supervisors: Huajin Wang and Sarah Young

Background

● Professors and researchers often seek the help of the Carnegie Mellon University Libraries to acquire information about the university’s research grants and research collaborations.
● Current methods to answer these inquiries require large amounts of manual labor and analysis, as there are no tools at the CMU Libraries’ disposal to process and glean insights from the research grant data they have.
● The goal of our research is to create such a tool—specifically, an R Shiny app—that can visualize and analyze the data in many ways, including those that address these questions:
  1. What funding agencies have funded research at CMU and how has this changed over time?
  How is this related to fields/categories of research?
  2. Which authors are associated with a certain field/category of research?
  3. How do funded research topics change over time?
By addressing these questions, the CMU Libraries will be better equipped to answer grant-related questions, therefore empowering university’s active research community.

Data Description

Our dataset, pulled from Dimensions AI, a research grants database that links grants to publications, consists of 6,128 research grants affiliated with CMU and 28 primary information fields, which can be summarized in three main groups:

<table>
<thead>
<tr>
<th>General</th>
<th>Funding</th>
<th>Categorizations (System Author)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grant ID, Number</td>
<td>Funding Amount</td>
<td>Fields of Research Categories (ANZSRC)</td>
</tr>
<tr>
<td>Title</td>
<td>Funder</td>
<td>Research, Condition, and Disease Categories (NH)</td>
</tr>
<tr>
<td>Abstract</td>
<td>Funder Group</td>
<td>Health Categories (HRCS)</td>
</tr>
<tr>
<td>Researchers</td>
<td>Funder Country</td>
<td>Health Research Activities (HRCS)</td>
</tr>
<tr>
<td>Research Organization</td>
<td>Funder Grant Link</td>
<td>Cancer Types (ICRP)</td>
</tr>
<tr>
<td>Research City, State, Country</td>
<td></td>
<td>Common Scientific Outline Categories (ICRP)</td>
</tr>
<tr>
<td>Start Date, Year</td>
<td></td>
<td>Units of Assessment (REF)</td>
</tr>
<tr>
<td>End Date, Year</td>
<td></td>
<td>Sustainable Development Goals (UN)</td>
</tr>
<tr>
<td>Resulting Publications</td>
<td></td>
<td>Dimensions AI Link</td>
</tr>
</tbody>
</table>

In the pre-processing stage, we combined the categorizations into one column for easier access and created a mapping to combine similar categories.

Categories & Funding

● Our R Shiny app allows users to visualize and analyze trends in categories of research and funding sources for a selected time frame using bar graphs.

Researchers & Collaborations

● Our R Shiny app also allows users to investigate which researchers are associated with a given field and their collaborations with each other.

Topic Modeling

● Within a research category, there is no further granularity of research topics.
● Leveraging Natural Language Processing (NLP) techniques on the grant titles and abstracts proves to be promising in identifying these topics.
● Specifically, we use Latent Dirichlet Allocation (LDA) with Gibbs Sampling, an unsupervised learning model, to find topic clusters.

Advanced Searching

● Users can easily search for relevant research using keywords and phrases.
● Alternatively, users can provide a specific grant of interest and explore similar works.

Conclusions & Future Work

● Our app displays a wide variety of insights that can be used by the CMU Libraries to help professors and researchers find the resources they need to succeed.
● About 1,700 grants do not have any category labelings, therefore limiting the functionality of some of our application’s tabs. Initial attempts to categorize these uncategorized grants based on patterns in titles and abstracts were unsuccessful, so we defer the determination of optimal methods for categorizing these grants to future work.
Exploring Trends in CMU Grant Data

Kyra Balenzano, Melissa Dy, Michael Li, and Veda Lin

Project Advisor: Peter Freeman    Project Supervisors: Huajin Wang and Sarah Young

Carnegie Mellon University Libraries
Background

- Professors and researchers often seek the help of the Carnegie Mellon University Libraries to acquire information about the university’s research grants and research collaborations.
- Current methods to answer these inquiries require large amounts of manual labor and analysis, as there are no tools at the CMU Libraries’ disposal to process and glean insights from the research grant data they have.
- The goal of our research is to create such a tool—specifically, an R Shiny app—that can visualize and analyze the data in many ways, including those that address these questions:
  1. *What funding agencies have funded research at CMU and how has this changed over time? How is this related to fields/categories of research?*
  2. *Which authors are associated with a certain field/category of research?*
  3. *How do funded research topics change over time?*

By addressing these questions, the CMU Libraries will be better equipped to answer grant-related questions, therefore empowering university’s active research community.
Our dataset, pulled from Dimensions AI, a research grants database that links grants to publications, consists of 6,128 research grants affiliated with CMU and 28 primary information fields, which can be summarized in three main groups:

**General**
- Grant ID, Number
- Title
- Abstract
- Researchers
- Research Organization
- Research City, State, Country
- Start Date, Year
- End Date, Year
- Resulting Publications
- Dimensions AI Link

**Funding**
- Funding Amount
- Funder
- Funder Group
- Funder Country
- Funder Grant Link

**Categorizations (System Author)**
- Fields of Research Categories (ANZSRC)
- Research, Condition, and Disease Categories (NIH)
- Health Categories (HRCS)
- Health Research Activities (HRCS)
- Cancer Types (ICRP)
- Common Scientific Outline Categories (ICRP)
- Units of Assessment (REF)
- Sustainable Development Goals (UN)

In the pre-processing stage, we combined the categorizations into one column for easier access and created a mapping to combine similar categories.
Our R Shiny app allows users to visualize and analyze trends in categories of research and funding sources for a selected time frame using bar graphs.
Topic Modeling

- Within a research category, there is no further granularity of research topics.
- Leveraging Natural Language Processing (NLP) techniques on the grant titles and abstracts proves to be promising in identifying these topics.
- Specifically, we use Latent Dirichlet Allocation (LDA) with Gibbs Sampling, an unsupervised learning model, to find topic clusters.
- Users can freely adjust the number of clusters to find the most personally usable results, and the year range to see changes over time. Additionally, details of corresponding grants can be displayed.
Our R Shiny app also allows users to investigate which researchers are associated with a given field and their collaborations with each other.
Advanced Searching

- Users can easily search for relevant research using keywords and phrases.
- Alternatively, users can provide a specific grant of interest and explore similar works.
Conclusions & Future Work

- Our app displays a wide variety of insights that can be used by the CMU Libraries to help professors and researchers find the resources they need to succeed.

- About 1,700 grants do not have any category labelings, therefore limiting the functionality of some of our application’s tabs. Initial attempts to categorize these uncategorized grants based on patterns in titles and abstracts were unsuccessful, so we defer the determination of optimal methods for categorizing these grants to future work.
Thank you!