ANALYSES OF RESEARCH PROJECTS

EM624: Final Project

Abstract

Analyses and visualization of research project topics from the Naval postgraduate school's centralized archive maned Calhoun

Viraj Patel

Date: 04/27/22

Contents

Project Goals and Conditions

Objective:

The objective of my project is to extract information from the Naval postgraduate school's website where they have stored thousands of research projects. Their research material is stored in a centralized archive maned Calhoun, that is the Naval Postgraduate School's digital repository for research materials and institutional publications created by the NPS community. Materials in the Calhoun are openly accessible to anyone on the web and will be preserved for future generations.

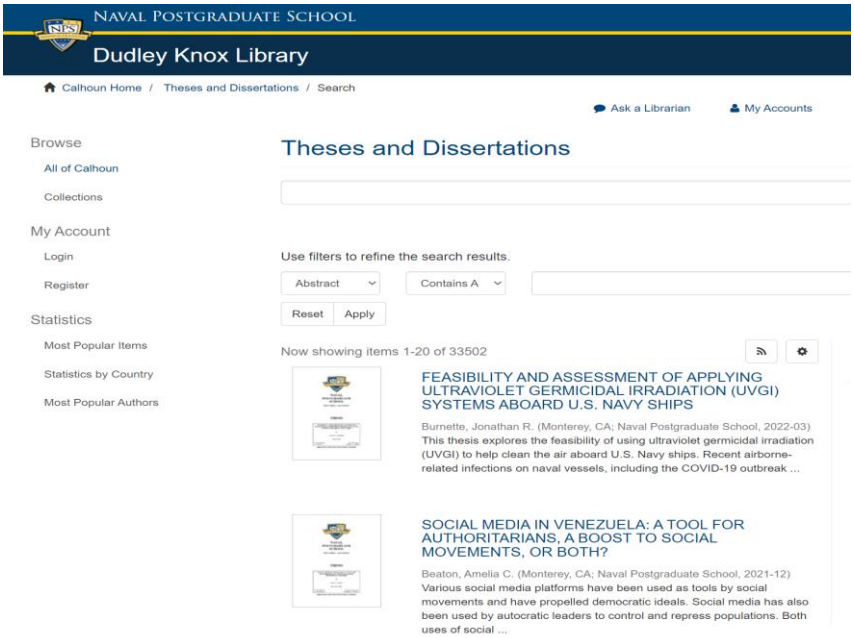
This study is focused on web mining, which will be used for extracting data from the web, and once the data is extracted than analysing that data for insights, such as what are the key areas/topics in time and what is the correlation between different authors and which authors have been more active.

Tools and Libraries:

My primary tool for this study would be python. I will also be using a lot of different libraries which python offers, hence making it one of the most powerful tools for data exploration and analyses. I will be using libraries such as requests, beautifulsoup, matplotlib, wordcloud, pandas, numpy, seaborn, nltk, mlxtend and few other supporting libraries.

The URL used for this study is:

https://calhoun.nps.edu/handle/10945/16/discover?rpp=20&etal=0&group_by=none&page=1



Website Visual

CRISP-DE

Data Collection:

The “inspect” option of the website was used to get the html content of the website, which was than used along with “requests” and “beautifulsoup” libraries of python to scrape the required content from the web. One of the difficult tasks here was to iterate through different pages to get as many projects as possible. The website had more than 1500 pages and more than 33000 reports, so code was taking long to execute, hence keeping in mind the time constraint, I choose to collect the data of first 50 pages and store it as a csv file using the “pandas” library.

The columns of the data frame are “Topic” ,”Date” ,”Authors” and “Description”. I tried getting the department name as well, but it was listed within the link of a link, and most of them didn’t have department listed, so it proved to be a difficult task keeping in mind the project duration, hence the department was excluded for now.

	Topics	Authors	Date	Description
0	Democratic transformations in Eastern	Fuior, Corneliu	2007-03	In the process of international socialization, states are induced to adopt the constitutiv
1	Coercion, cash-crops and culture from	Merz, Andrew A.	2008-06	This thesis is a comparative study of conflict and opium in the Golden Crescent and Gc
2	Littoral Combat Ship (LCS) mission	Abbott, Benjamin P	2008-03	The threat of a large fleet engagement in the open ocean is currently over shadowed b
3	The next best alternative to an ideal	Ayhan, Serhat.	2009-03	The supply of high quality recruits is limited and services are facing a diminishing recru
4	Missile defense certification examination	Biggs, Robert C.	2008-09	The process employed by Naval Surface Forces to capture information during warfare
5	The intelligence phenomenon in a new	Filip, Valentin F.	2006-03	In the post-Cold War era, two major trends illustrate the evolution of the international
6	Density as a cost driver in Naval	Grant, Benjamin P.	2008-06	This thesis examines density reduction as an alternative to weight or size reduction wh
7	Business case analysis of the Joint Strike	Shick, BethAnn.	2007-12	The Joint Strike Fighter (JSF) program is the largest Department of Defense (DoD) milit
8	Undersea node localization using node-	Zinkhon, David C.	2009-03	Seaweb is a wide-area network interconnecting a set of distributed underwater nodes
9	North Korea's Juche ideology and the	Juetten, Stephan A.	2008-06	This thesis analyzes potential socio-cultural discord upon eventual Korean national reu
10	Security culture in times of war: how did	Reimers, Frank.	2007-06	This thesis uses a comparative case study approach to examine how security cultures
11	The Effect of Shipboard Design Decisions	Poniatoski, Jessica I	2012-06	Recently, a major decision was made by the Navy to eliminate a key capability from the
12	Employing organizational modeling and	Paskin, Matthew A.	2007-12	This research modeled and simulated the KC-135 aircraft's Programmed Depot Mainte
13	Best practices in the Navy's energy	Haley, Ryan C.; Klot	2009-12	The Department of the Navy is the second largest consumer of petroleum within the D
14	Energy regulation effects on critical	McQuinn, Matthew	2008-12	U.S. critical infrastructure includes those assets that are vital to maintaining the nation'
15	Modeling and frequency tracking of	Severson, Jared	2009-02	Marine mammal whistle calls present an attractive medium for covert underwater com
16	NOVEL SYNTHESIS AND	Maxson, Ashley R.	2013-03	With the aim to develop personal protection systems with improved mechanical prope
17	Numerical Performance Prediction of a	Chen, Bingqiang	2012-09	Using a 3-D axis-symmetric model, the cold-flow performance of a miniature ramjet in

Extracted Dataset

Data Understanding:

On exploring the extracted dataset, it is found that the dataset contains 4 columns, excluding the index, and has about 1000 rows. The dates are listed in a random order and some of the topics and descriptions is missing. Most of it is categorical data, hence I will be doing some analyses based on the categorial values.

Data Pre-Processing:

On using some common python functions like finding null values and empty rows, I found that the dataset contains few null values, firstly those were removed using the “dropna” function on the columns, then the dates were sorted in a descending order to get the latest one first.

Once that was done, I cleaned the “Topics” and “Description” columns by removing stop words, punctuations, and words less than 3 characters and stored them as an array, for extracting word clouds, bigrams, and topic modelling. The “Authors” column did not need much cleaning, but it will be used for finding correlation between different authors.

Data Analyses and Visualization:

Once I had the final cleaned dataset, I started doing the data analyses by first generating a word cloud for different topics and their descriptions. On first few runs it was seen that some of the words that were more commonly appearing, were obvious words like ‘research’, ‘analyses’, ‘study’, and some more. These words didn’t add value to our analyses, hence stop words list was extended, by adding these obvious words. The word clouds generated for topics and description can be seen below.

Word cloud for research topics over time

Word cloud for topic descriptions

Next, I explored the bigrams for the topics column of the data frame. The output is as follows:

```
The first 50 Bi-grams for the Topics are
[('Democratic', 'transformations'),
('transformations', 'Eastern'), ('Eastern', 'Europe'),
('Europe', 'efficacy'), ('efficacy', 'norm'), ('norm',
'diffusion'), ('diffusion', 'strategies'),
('strategies', 'NATO'), ('NATO', 'culture'),
('culture', 'insurgency'), ('insurgency', 'opium'),
('opium', 'belt'), ('belt', 'Littoral'), ('Littoral',
'Combat'), ('Combat', 'Ship'), ('Ship', 'mission'),
('mission', 'packages'), ('packages', 'determining'),
('determining', 'best'), ('best', 'best'), ('best',
'alternative'), ('alternative', 'ideal'), ('ideal',
'recruit'), ('recruit', 'attrition'), ('attrition',
'characteristics'), ('characteristics', 'recruits'),
('recruits', 'waivers'), ('waivers', 'educational'),
('educational', 'credentials'), ('credentials',
'Army'), ('Army', 'Missile'), ('Missile', 'defense'),
('defense', 'certification'), ('certification',
'examination'), ('examination', 'Aegis'), ('Aegis',
'warship'), ('warship', 'Army'), ('Army', 'patriot'),
('patriot', 'crew'), ('crew', 'certification'),
('certification', 'process'), ('process',
'intelligence'), ('intelligence', 'phenomenon'),
('phenomenon', 'democratic'), ('democratic',
'milieu'), ('milieu', 'Romania'), ('Romania', 'case'),
('case', 'Density'), ('Density', 'cost'), ('cost',
'driver')]
There are total 6148 Bi-grams for Topics
```

Bigrams for topics

As seen, there were a total of 6148 bigrams for the 997 rows of topics, I had in my data frame. On breaking these bigrams into chunks of 50 to 100 and analysing those, I noticed that words like ‘mission and ship’, ‘education and waiver’, ‘army and warship’, ’alternative and ideal’ are some of the more frequent bigrams. This is an important fact to analyse as it can later help me with feature extraction for topic modelling.

I than used the nltk library to print the frequency of the most commonly occurring words in the topic description column. The output is as follows:

```
[('systems', 62), ('purpose', 53), ('states', 49),
('military', 49), ('current', 49), ('model', 47),
('united', 45), ('data', 42), ('process', 41),
('security', 41), ('defense', 36), ('determine', 33),
('examines', 32), ('high', 32), ('surface', 31),
('developed', 31), ('management', 30), ('department',
30), ('national', 30), ('investigation', 29),
('marine', 28), ('potential', 26), ('provide', 26),
('information', 25), ('flow', 25), ('computer', 25),
('control', 25), ('effects', 24), ('combat', 24),
('major', 23)]
```

Frequent words in topics

The information obtained above can be useful when doing NLP analyses, and feature extraction. It also helps us to understand the context of the research topics and what are the focus areas.

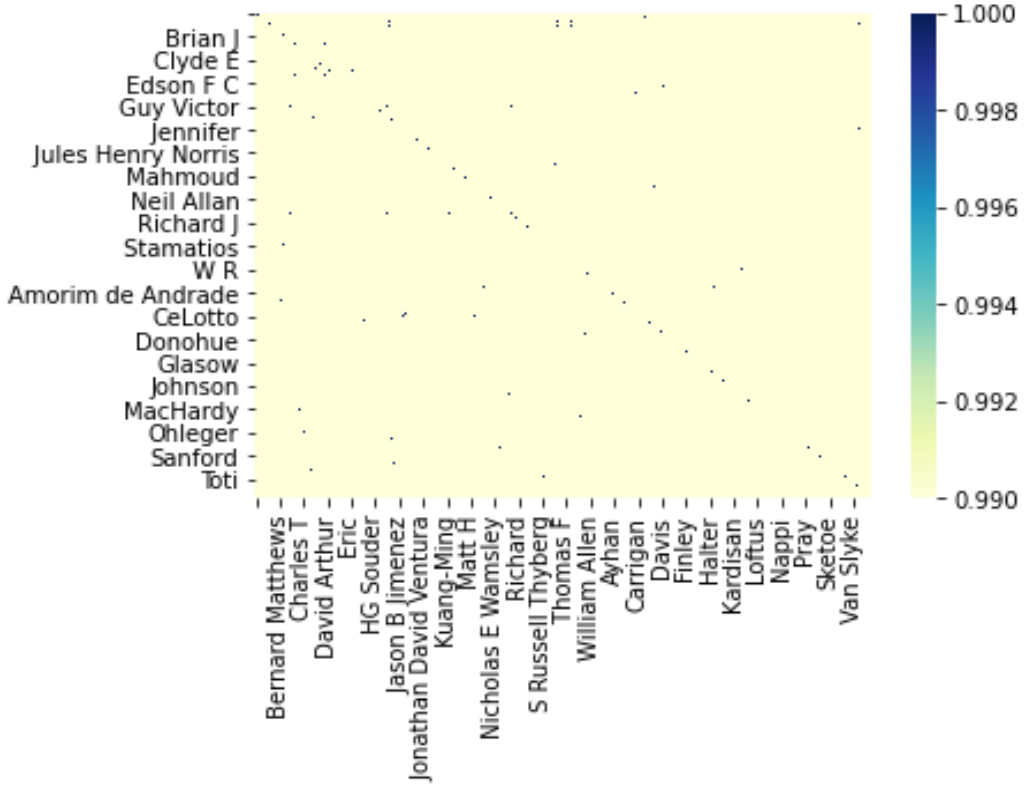
Next, I tokenized the 'authors' column, again to extract bigrams to see which two authors worked together more frequently. The output is as follows:

```
The first 20 Bi-grams for authors are
[('Corneliu', 'Andrew'), ('Andrew', 'Benjamin'),
('Benjamin', 'Robert'), ('Robert', 'Valentin'),
('Valentin', 'Benjamin'), ('Benjamin', 'David'),
('David', 'Stephan'), ('Stephan', 'Jessica'),
('Jessica', 'Matthew'), ('Matthew', 'Alice'),
('Alice', 'Ryan'), ('Ryan', 'Roy'), ('Roy', 'Shane'),
('Shane', 'Matthew'), ('Matthew', 'Jared'), ('Jared',
'Ashley'), ('Ashley', 'Bingqiang'), ('Bingqiang',
'Jessica'), ('Jessica', 'Benjiman'), ('Benjiman',
'Ceying')]

There are total 1352 Bi-grams for Authors

Authors that worked together more frquently are:
[(('Michael', 'Thomas'), 3), (('Robert', 'Joseph'),
3), (('Robert', 'William'), 3), (('Francis', 'John'),
3), (('William', 'Robert'), 3), (('William',
'Edward'), 3), (('James', 'Edward'), 3), (('Jonathan',
'Steven'), 2), (('Michael', 'Paul'), 2), (('Matthew',
'Michael'), 2), (('Steven', 'William'), 2),
(('William', 'David'), 2), (('Ryan', 'Peter'), 2),
(('John', 'Michael'), 2), (('Thomas', 'Edward'), 2),
(('Charles', 'James'), 2), (('William', 'Francis'),
2), (('Thomas', 'John'), 2), (('Geoffrey', 'Keith'),
2), (('Donald', 'William'), 2)]
```

Bigrams for authors

There are a total of 1352 bigrams for authors. Since the website has more than 33000 documents and I am working with just 1000 of those, I didn't find much collaboration pairs of authors but the most common was 3 times that any two unique authors worked together, their names can be seen in the picture above. Bigrams only show us how many of them worked together how many times, whereas I also wanted to see how many different authors one author collaborated with, hence I had to plot a correlation matrix for all unique authors.

I tried to find how are different authors related to each other using the “One hot encoder” library from “mlxtend”. Using this library, I converted all the unique authors in each row to ones and zeros, for plotting the correlation matrix. On the first attempt the heat map wasn’t very clear as there were so many unique authors and most of them didn’t have any collaboration, so I tweaked the “vmin” function to 0.99 to include only the authors who collaborated the more. The heat map is as follows:



Heatmap for top author collaborations

As you can see the heatmap was still not quite significant, so I extracted data out of this heatmap, to find the authors with highest correlation and I also used the “apriori” function to find the support level for different authors. The authors with the highest collaboration and support level for few other authors is as follows:

```
Alan Riva          Amie Wiborg       1.0
William O          Powell            1.0
Curtis J Hosoglu   Michael A Hall    1.0
Joseph Patrick     Heil              1.0
FM Coleman         LA Nasipak        1.0
dtype: float64
```

Authors with more collaboration

```
0.006  ( Michael J)
0.006   ( Thomas J)
0.006    (Anderson)
0.006      (Harris)
0.007     (Johnson)
0.005      (Porter)
0.008       (Smith)
0.006    (Williams)
```

Authors and their support level

As we can see from the support level numbers that, those were very low indicating very few collaborations between authors.

Project Summary

Conclusion:

On doing the analyses of the extracted word clouds, bigrams and commonly appearing words in the description, it can be concluded that most of the research projects involved designing a purposeful system using computers and complex modelling techniques, which could help defence and military in the US. The projects involved investigating a problem, developing a potential solution keeping in mind the risks involved and cost associated with it. Most projects belonged to mechanical and systems engineering departments.

Since my dataset only comprised of 1000 documents out of 33000 available documents, from the heatmap it can be concluded that there wasn't much collaboration between authors, the most being thrice that any two authors worked together, like 'Thomas and Michael' and 'Robert and Joseph' being a couple of them, whereas authors like 'Joseph Patrick', 'Amie Wiborg' collaborated the most. The overall support level for each author was very low indicating very low collaboration overall.

Reflection:

One of the most important skills I was able to acquire while working on this project was web mining and extracting textual data from the web using 'requests' and 'beautifulsoup' libraries. I also explored on how to work with categorical variables, and plot correlation matrix for categorical variables using one hot encoder and "apriori" function. I also tried an approach to topic modelling, but due to its complex algorithm and time constraint I couldn't execute it, but I would surely, add it to this report for the future.

References:

- Zychlinski, S. (2019, December 26). *The Search for Categorical Correlation*. Medium. https://towardsdatascience.com/the-search-for-categorical-correlation-a1cf7f1888c9
- *Search*. (n.d.). Calhoun.nps.edu. Retrieved April 25, 2022, from https://calhoun.nps.edu/handle/10945/16/discover?rpp=20&etal=0&group_by=none&page=1&sort_by=dc.date.issued_dt&order=desc
- Kapadia, S. (2020, December 29). *Topic modeling in Python: Latent dirichlet allocation (LDA)*. Medium. Retrieved April 24, 2022, from https://towardsdatascience.com/end-to-end-topic-modeling-in-python-latent-dirichlet-allocation-lda-35ce4ed6b3e0