

DSC 180B A14

Project Checkpoint

Sijie Liu, Siddhi Patel, Brian Qian, Du Xiang, Martha Yanez

HDSI Faculty Exploration Tool using LDA Topic Modeling

ABSTRACT

Halicioğlu Data Science Institute industry partners are constantly looking for faculty to work in their projects. In order to make this an easier process, we used Latent Dirichlet Allocation to find the most written topics by HDSI faculty. We processed the abstracts of published works of HDSI affiliated faculty in order to run an LDA model on them. We found their most representative topics and created a tool that effectively shows the connection among them. We have made significant improvements to a previously existing tool, mainly labeling our topics and creating a more user friendly experience. In order to maintain this tool functioning, we have generated a workflow that allows for future changes. We also have created an additional tool that allows for easy search of faculty.

INTRODUCTION

The Halicioğlu Data Science Institute (HDSI) at University of California, San Diego is dedicated to the discovery of new methods and training of students and faculty to use data science to solve problems in the current world. The HDSI has several industry partners that are often searching for assistance to tackle their daily activities and need experts in different domain areas. Currently, there are around 55 professors affiliated to HDSI. They all have diverse research interests and have written numerous papers in their own fields. Our goal is to create a tool that allows HDSI to select the best fit from their faculty, based on their published work, to aid their industry partners in their specific endeavors. We will be doing this with Natural Language Processing (NLP) by managing all the abstracts from the faculty's published work and organizing them by topics. We will then discover what is the proportion of papers of each faculty associated with each of the topics and draw a relationship between researchers and their most

published topics. This will allow HDSI to personalize recommendations of faculty candidates to their industry partner's particular job.

We use Latent Dirichlet Allocation (LDA) to process our texts and obtain the most frequent words for each topic. Based on this information, a tool was created in the form of a Sankey Diagram where a specific number of topics can be selected and the relationships between authors and this number of topics displayed. The topics now have the appropriate labels that indicate the main topic related to a particular search/author. The version that this paper discusses is what we call the version 2.0 of our tool.

Since this tool will be used by Industry Partners of HDSI, who might not be familiar with Sankey Diagrams or could have difficulty interacting with it or interpreting the results, we decided to create a companion tool in the form of a search bar, which we will refer to as Easy Faculty Search Tool. This is just another way of visualizing our LDA topic modeling results.

a. Previous Work

As mentioned above, this is version 2.0 of the Sankey Diagram tool. Our previous work involved the replication of an already existing tool made by a team of data scientists at HDSI, which was the foundation for what we present now. What he had done differently in the past, was changing the number of topics that were used on each iteration of the replication by each member of our team.

Since there was not a published article related to our particular tool, our replication was based solely on the code but theoretically supported by a number of articles related to the methods utilized. Our work was possible with the use of Latent Direct Allocation and based on published work by D. Blei, A. Ng and M. Jordan as well as articles by S. Prabhakaran and S. Kapadia.

b. Literature Review

As mentioned in section a., in order to perform the replication of the allocation tool, we used texts by the aforementioned authors. “Latent Dirichlet Allocation” is a paper by D. Blei, A. Ng and M. Jordan that explains in detail the origin and process of LDA. It explains the motivations of modeling text corpora and the techniques used around LDA. The goal of the paper was to find short descriptions of the members of a collection that enable efficient processing of large collections while preserving the essential statistical relationships that are useful for basic tasks such as classification, novelty detection, summarization, and similarity and relevance judgments.¹

In order to understand the step-by-step process of the coding solution to our tool, articles on Towards Data Science and Machine Learning Eg were consulted. For an understanding of a less theoretical overview of LDA, we consulted “Topic Modeling in Python: Latent Dirichlet Allocation (LDA)” by S. Kapadia on Towards Data Science. When it comes to the coding process of our tool, this article helps us understand how to use gensim to perform LDA.

“LDA in Python – How to grid search best topic models?” by S. Prabhakaran explains how to perform LDA by using one of the most popular machine learning Python libraries: scikit learn. This article explains step by step how to perform the process, therefore it was very useful for the present report.

c. Data

¹ David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. J. Mach. Learn. Res. 3, null (3/1/2003), 993–1022.

In order to create the tool, an analysis on text data was performed. We decided to use only the abstract portions of published faculty works and perform LDA on this.

The data was obtained from Dimensions, which is a platform that allows for the search and analysis of over 150 million interlinked data items from across the research world. Such data items include publications, grants, clinical trials, patents, and policy documents.²

We manually collected each HDSI faculty member's researcher ID from Dimensions and from this, we created a csv file. Dimensions' application programming interface, or API, was used in order to collect the abstracts belonging to these particular researcher IDs beginning in the year 2015. The following image shows the first 4 rows and a sample of the columns of the data collected.

	abstract	authors	concepts	date	id	times_cited	title	year
0		[[{"affiliations": [], "corresponding": true, "city": "Columbus"}]]	[space, metric spaces]	2021-01-01	pub.1140323831	0	Elder-Rule-Staircodes for Augmented Metric Spaces	2021
1	Understanding of neuronal circuitry at cellular...	[[{"affiliations": [{"city": "Cold Spring Harbor"}]]	[hybrid architecture, semantic segmentation, d...]	2020-09-28	pub.1131237718	2	Semantic segmentation of microscopic neuronal...	2020
2	We study Vietoris-Rips complexes of metric wed...	[[{"affiliations": [{"name": "MOSEK A/S", "city": "Copenhagen"}]]	[Vietoris-Rips complexes, wedge sum, metric g...]	2020-05-20	pub.1127784757	5	On homotopy types of Vietoris-Rips complexes a...	2020
3	Neuroscientific data analysis has traditional...	[[{"affiliations": [{"city": "Columbus"}]]	[collection of neurons, hand-tuned parameters, ...]	2020-03-22	pub.1125823510	0	Detection and skeletonization of single neuron...	2020

Fig 1. Initial dataset obtained by using Dimensions API

Preprocessing the abstracts by using stemming, removing stop words and the `gensim` simple pre-process, as well as adding a column containing the actual HDSI faculty member that was included in the author list of each paper, resulted in a dataset of the following form.

² <https://datanexus.ucsd.edu/analytic-data/dimensions.html>

abstract	times_cited	concepts	journal.title	HDSI_author	abstract_processed
nan	0	['space', 'metric spaces']	SIAM Journal on Applied Algebra and Geometry	Yusu Wang	
Understanding of neuronal circuitry at cellula...	3	['hybrid architecture', 'semantic segmentation...']	Nature Machine Intelligence	Yusu Wang	understand neuronal circuitry cellular resolut...
We study Vietoris-Rips complexes of metric wed...	5	['Vietoris-Rips complexes', 'wedge sum', 'metr...']	Journal of Applied and Computational Topology	Yusu Wang	study vietoris rip complexes metric wedge sum ...
Neuroscientific data analysis has traditionall...	0	['collection of neurons', 'hand-tuned paramete...']	bioRxiv	Yusu Wang	neuroscientific data analysis traditionally re...
Neuroscientific data analysis has traditionall...	0	['collection of neurons', 'hand-tuned paramete...']	arXiv	Yusu Wang	neuroscientific data analysis traditionally re...

Fig 2. Dataset including Processed abstracts and HDSI authors.

Figure 2 presents the first five rows of the final dataset used, excluding the columns `year`, `authors` (list form) and `title`.

We were able to perform LDA on the `abstract_processed` column and to subsequently, use the year and author data to obtain additional information required like most predominant topic document and per author.

Abstract data was used since it is a concise summary of each publication, it is easier and more efficient to process, and it contains words that are representative of the articles. Therefore, the dataset presented in Figure 2 was enough for us to perform our tasks and obtain our final tool, as well as to gain a deeper understanding of the overall data obtained from Dimensions.

In order to improve version 1.0 of our tool, we decided to extract fields from both Google Scholar and Dimensions to use as labelling for authors and articles.

Since Google does not provide any API for their Scholar product, we used [Selenium](#), which is a web-crawling framework that can be used for data extraction. Even though we faced this challenge, we still decided to use Google Scholar because it provides labels under each author that indicate their general field, as seen on Figure 3. Google Scholar also has information on faculty that Dimensions does not, as well as some of their missing articles.

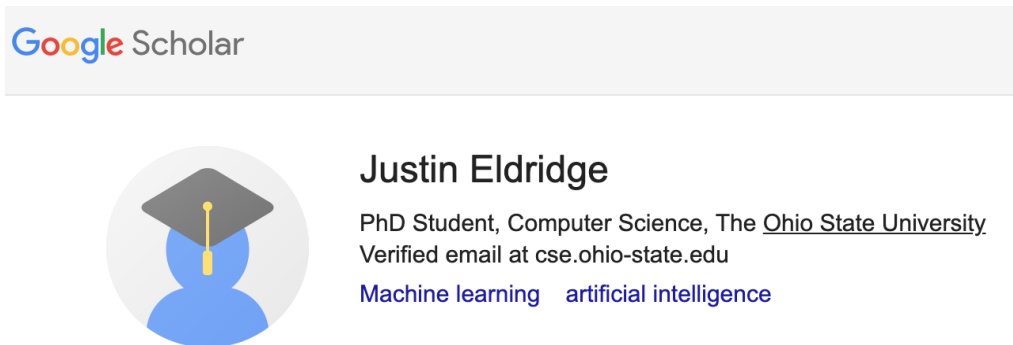


Fig 3. Google Scholar profile for an author shows general field labels. In this case, for Dr. Eldridge they are ‘machine learning’ and ‘artificial intelligence’.

As for the labeling at the article level, we decided to use Dimensions since it provides Research Categories under each article, as seen below.



Fig 4. Dimensions example of a publication and the fields of research under which it falls under.

METHODS

For version 1.0 of this tool, we had manually assigned the labeling of topics but had left out Topic 0-Topic n on the search results of the dashboard.

To allow better understanding of these topics, we used several methods to find appropriate labels for our results.

1. We scrapped the labels on the Google scholar page of faculty in HDSI and gained a histogram providing a general view of how they are distributed.

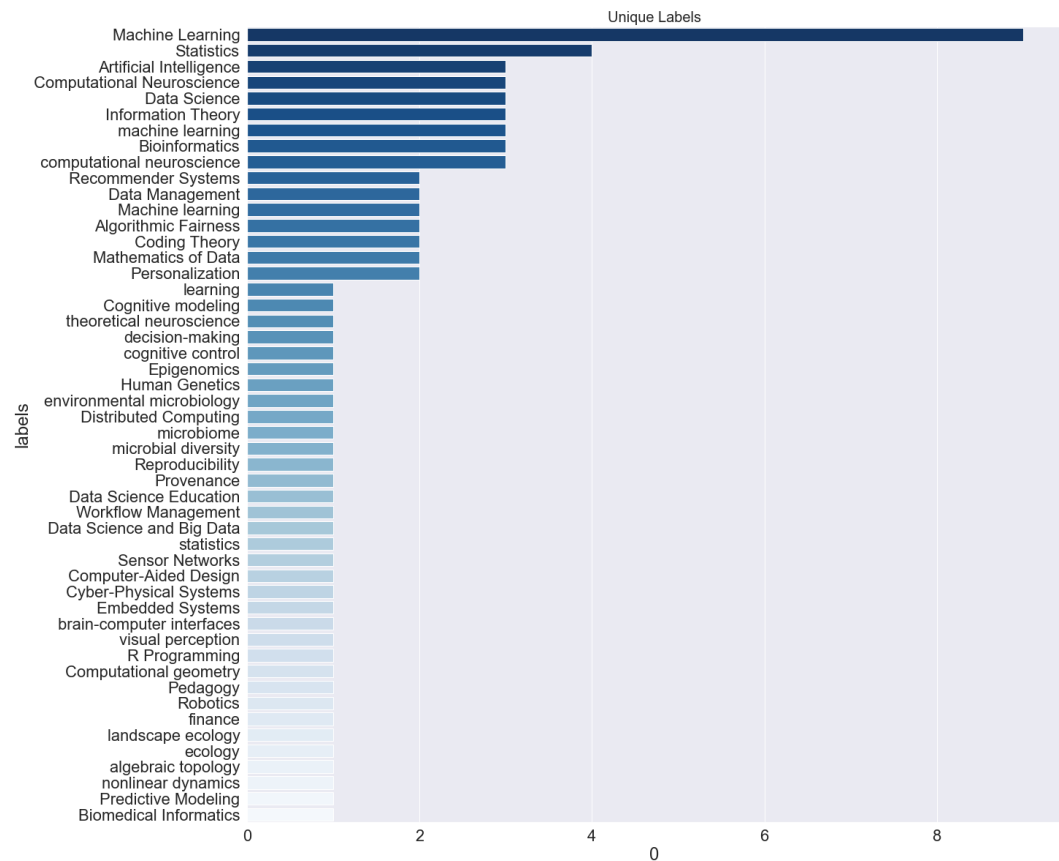
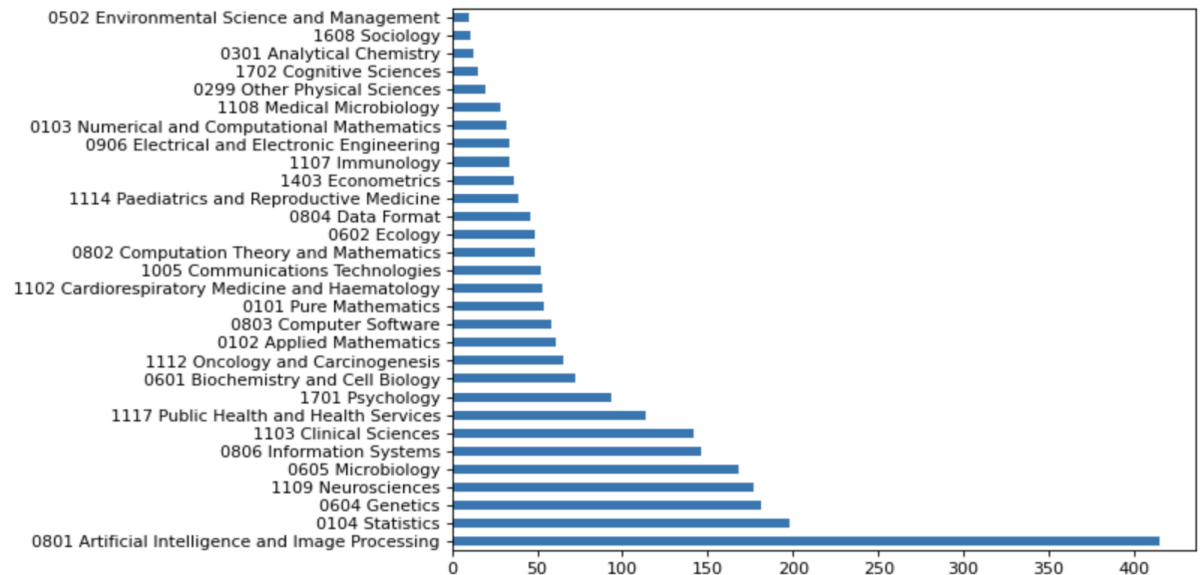


Figure x: unique labels for the labels gathered from the google scholar pages.

Since we found that unique labels included in the google scholar might not be able to provide a very clear explanation of the domain of interest for the faculty members. For example, “machine learning” can be a very broad part; a researcher can have a focus on supervised learning, unsupervised learning, reinforcement learning, or the application of a specific algorithm relating to his specific interest area. Therefore, we decided to further work on the version 2 of labeling.

- We also gathered the labels from the Dimensions API: The labels from dimensions give better interpretation at the article level. After selecting the more specific labels for the articles, we can better understand each article's main contribution. After aggregating the labels on the topic level, we can then gain a better understanding of the unsupervised results.



Now, our next step is to clean up the labels that's too vague and aggregate the labels to dynamically represent the topics we have from our topic model.

We hope to combine option 1 and option 2 to provide a better user experience for our search tool. Having labels that only work on a specific set of data would not be useful. We are currently working to improve our pipeline to be more adaptable to the data stream. For example, we would like to update our database once a year. In that case, our code must also produce meaningful results in a robust manner. Right now, our pipeline is more dynamically run on the newer version of data with specific year thresholds. However, there are still issues to be fixed and improved further in the data pipeline. Our ideal goal is that the tool can be useful and robust enough to handle any new incoming data.

Since our current Sankey dashboard offers many useful features, we wanted to imagine how we could integrate a more UI focused easy search tool utilizing a similar design language as the HDSI website while retaining the core functionality of our dashboard. Thus we have currently conceptualized a demo through Figma of our faculty exploration tool that would make it very

easy for our industry partners to quickly find what they're looking for within HDSI faculty. The tool is split within three sections: Search by Topic of Interest, Search by Keywords, and a Sankey Diagram visualization. The Sankey diagram will be similar to the one we have created in the past quarter.

COVID-19 Updates

Visit UC San Diego's Coronavirus portal for the latest information for the campus community.

[View Details](#)

UC San Diego

HALICIOĞLU DATA SCIENCE INSTITUTE

About ▾

Academics ▾

Research ▾

Industry ▾

Career Services ▾


News and Events ▾

Jobs ▾

Search By Topic of Interest ▾

→

Showing results for "Machine Learning"




Justin Eldridge

Area of Research: ML Algorithmns and Systems

Relevant Publications:

- "Unperturbed: spectral analysis beyond Davis-Kahan" ([read more...](#))




Arun Kumar

Area of Research: ML and AI

Relevant Publications:

- "Towards an Optimized GROUP BY Abstraction for Large-Scale Machine Learning" ([read more...](#))



Jelena Bradic

Area of Research: ML and Bioinformatics

Relevant Publications:

Stay In Touch with HDSI

First/Last Name

Email Address

Potential Data Science Student ▾

☐ By continuing, you accept the [UC San Diego Website privacy policy](#)

Subscribe

Upcoming Events

There are no upcoming events at this time.

Search By Topic of Interest

Sankey Diagram

Search By Keywords

Search By Topic of Interest



Assistant Teaching Professor
Work Email: jeldridge@eng.ucsd.edu

Area of Research: ML Algorithmns and Systems

List of publications:

- Unperturbed: spectral analysis beyond Davis-Kahan
- Beyond hartigan consistency: Merge distortion metric for hierarchical clustering
- Robust features for the automatic identification of autism spectrum disorder in children
- Graphons, mergeons, and so on!
- Denali: A tool for visualizing scalar functions as landscape metaphors

"Unperturbed: spectral analysis beyond Davis-Kahan"

General Topic: Machine Learning
Authors: Justin Eldridge, Mikhail Belkin, Yusu Wang
Publication Year: 2017

Abstract: Classical matrix perturbation results, such as Weyl's theorem for eigenvalues and the Davis-Kahan theorem for eigenvectors, are general purpose

Stay In Touch with HDSI

First/Last Name

Email Address

Potential Data Science Student

☐ By continuing, you accept the UC San Diego
[Website privacy policy](#)

Subscribe

Upcoming Events

There are no upcoming events at this time.

The Easy Search Tool will be coded using a conjunction of Observable and RxJS. Observable is a data visualization tool that uses minimal conventions to create custom visualizations. We will use that tool in order to rebuild our Sankey diagram to match the aesthetics of the HDSI website. RxJS is a reactive programming language that makes it easy to compose asynchronous and event-based programs. We will be using this to build our "Search by Keywords" and "Search by Topic of Interest" section of the Easy Search Tool.

APPENDIX

Q2 Project Proposal: HDSI Faculty Topic Allocation Tool

0. Abstract

In this proposal, we will lay out the previous work that has been attempted by our team to answer this question and how we plan on improving our results in the next quarter. We'll start out by discussing the intention behind our Q1 project that uses topic modeling in order to assign a domain expertise to HDSI-affiliated faculty at UCSD. We'll then discuss how we plan on furthering the current project to build a version 2.0 of the project.

1. Introduction

The Halicioglu Data Science Institute (HDSI) at the University of California, San Diego (UCSD) was developed in 2018 in order to be a hub for data science pursuits on the campus. Given this purpose, HDSI is not used solely as a center for students, but it also attracts a plethora of companies who want to partner with faculty in order to pursue different topics in data science. Given that there are around 55 faculty members affiliated with HDSI who all have diverse research interests, it's difficult for partnering companies to determine who to approach with a certain task. Our goal is to create a tool that allows HDSI to select faculty who are best fit to aid industry partners in their specific endeavors.

2. Q1 Work and Methods

For quarter 1 of this capstone class, our team focused on the replication of the 1.0 version of this tool that was developed in summer 2021 under the guidance of Dr. Molly Roberts. In order to begin developing the tool, we used the Dimensions application programming interface (API) to gather information about research articles written by each faculty member. Dimensions is a platform that allows for the search and analysis of over 150 million interlinked data items from across the research world. Such data items include publications, grants, clinical trials, patents, and policy documents. With these data items, you can view them in context, analyze results in

bulk, and trace relationships.³ With this API, we created a spreadsheet that contained data on 2,194 research papers affiliated with each faculty member. An example of the first few rows in the spreadsheet is reflected below.

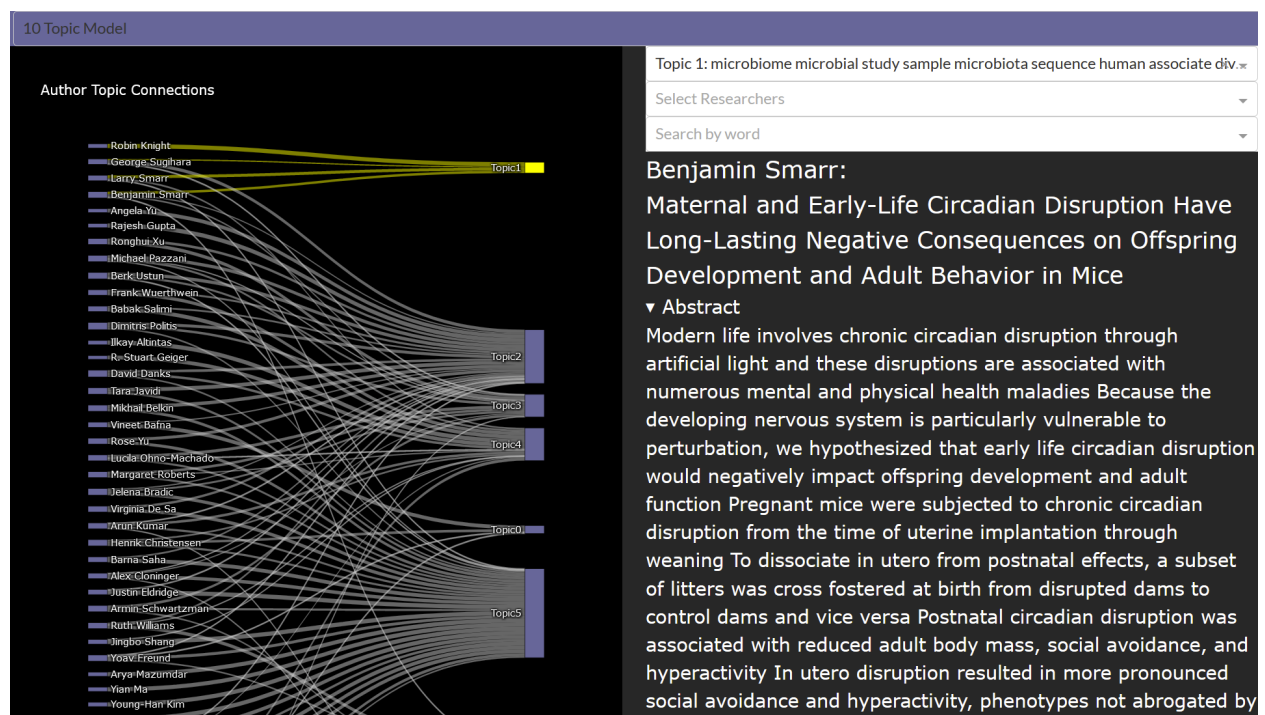
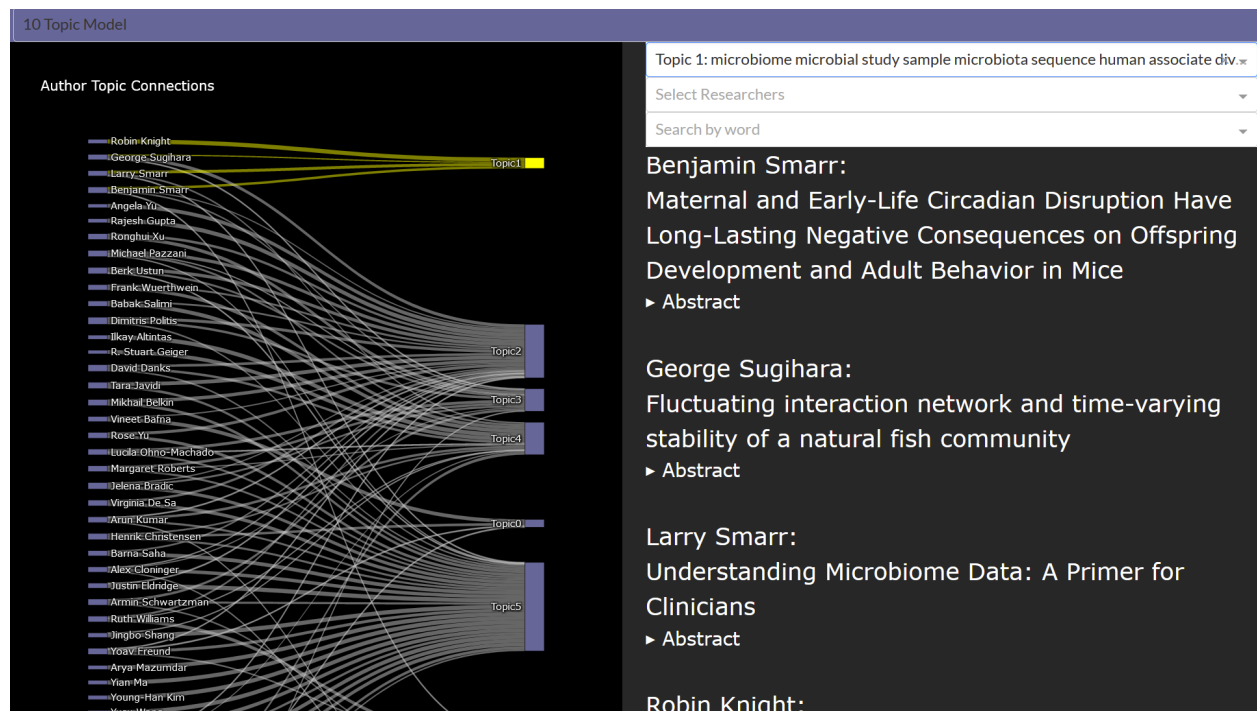
	year	authors	title	abstract	times_cited	concepts	journal.title	HDSI_author
0	2021	[{'raw_affiliation': [], 'first_name': 'Chen', ...}]	Elder-Rule-Staircodes for Augmented Metric Spaces	NaN	0	['space', 'metric spaces']	SIAM Journal on Applied Algebra and Geometry	Yusu Wang
1	2020	[{'raw_affiliation': ['Cold Spring Harbor Labo...'], 'first_name': 'Chen', ...}]	Semantic segmentation of microscopic neuroanat...	Understanding of neuronal circuitry at cellula...	3	['hybrid architecture', 'semantic segmentation...']	Nature Machine Intelligence	Yusu Wang
2	2020	[{'raw_affiliation': ['MOSEK ApS, Copenhagen, ...'], 'first_name': 'Chen', ...}]	On homotopy types of Vietoris–Rips complexes o...	We study Vietoris–Rips complexes of metric wed...	5	['Vietoris–Rips complexes', 'wedge sum', 'metr...']	Journal of Applied and Computational Topology	Yusu Wang
3	2020	[{'raw_affiliation': ['Computer Science and En...'], 'first_name': 'Chen', ...}]	Detection and skeletonization of single neuron...	Neuroscientific data analysis has traditional...	0	['collection of neurons', 'hand-tuned paramete...']	bioRxiv	Yusu Wang

After gathering this data, we utilized natural language processing (NLP) in the form of Latent Dirichlet Allocation in order to build topics associated with each faculty member based on the words in abstract sections of their research papers. More specifically, we obtained the proportion of papers of each faculty associated with different topics and drew a relationship between the researcher and their most published topics. This will allow HDSI to personalize recommendations of faculty candidates to their industry partner’s particular job and demonstrate a general representation of the variety of current skills that HDSI has to offer. In the next section, we’ll discuss the use of a sankey diagram in creating our final visualization that relays our findings for each faculty member.

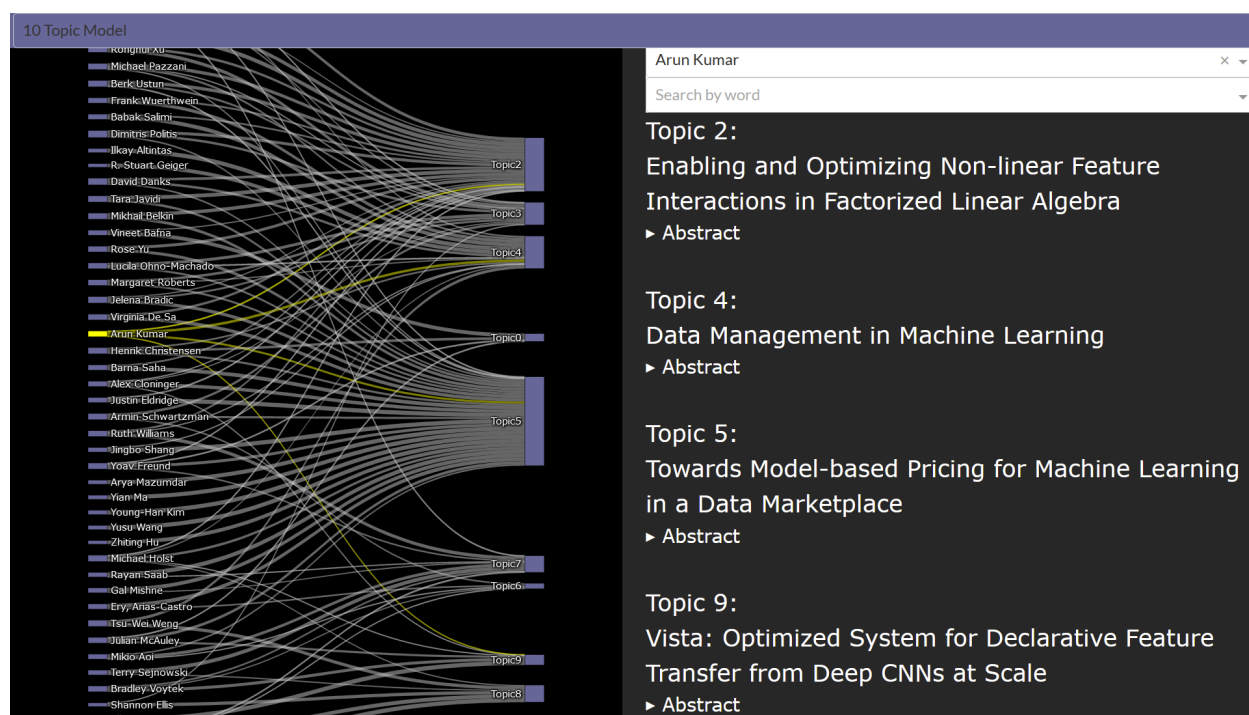
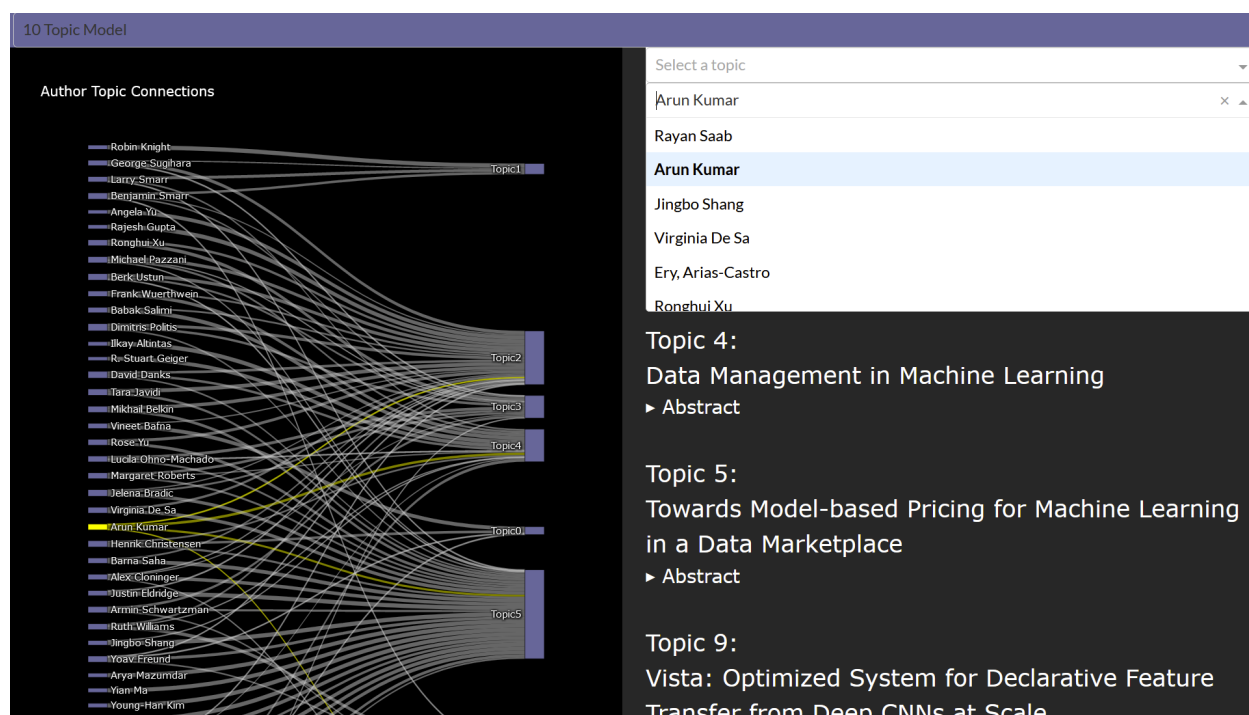
3. Q1 Results

As shown below, our current dashboard has many useful functions that provide a large amount of information that can be extremely helpful towards our industry partners, such as the topic selection feature where the proprietary topics are listed to demonstrate the variety of areas of research our HDSI faculty has to offer. Below this the faculty are listed with their respective papers, along with their abstracts, to give a general sense of what each faculty member is working on.

³ <https://datanexus.ucsd.edu/analytic-data/dimensions.html>

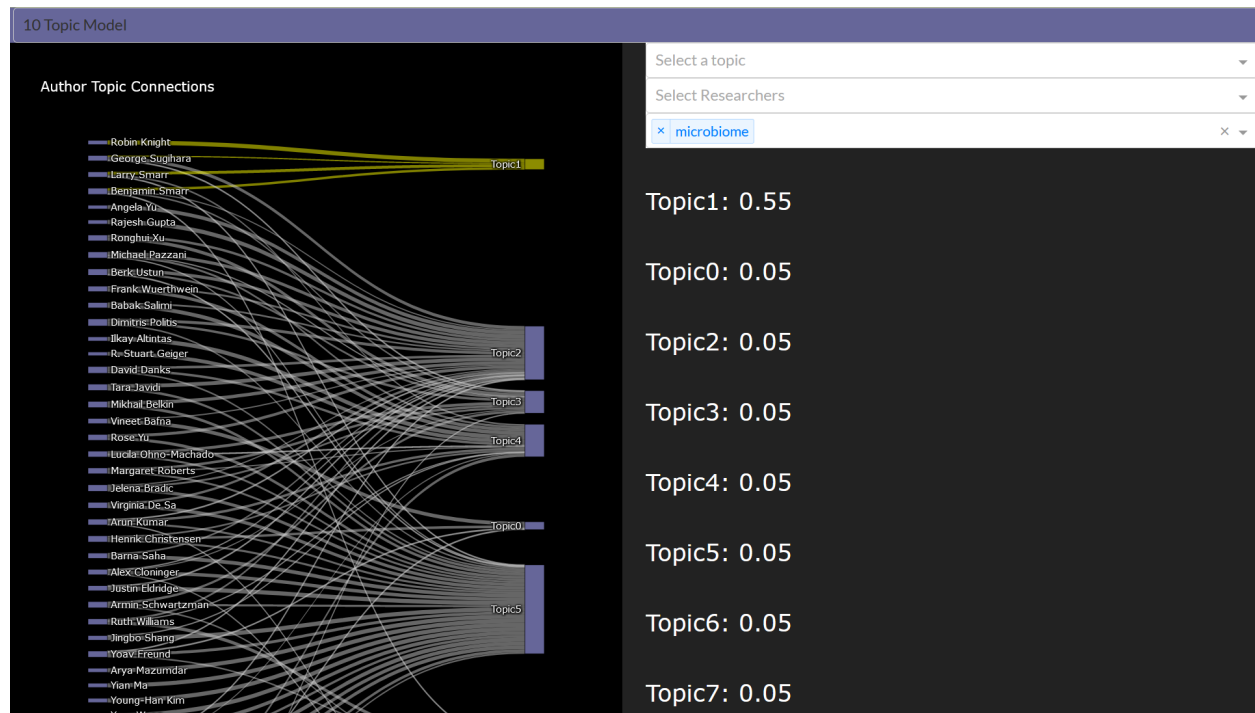


The researcher selection function is another interesting way of discovering where a specific faculty member's line of research actually lies. In Arun Kumar's case, with a 10 topic model, his papers demonstrate an expansive variety of topics from data workflow to neural networks and models.



Lastly the third main function is a general word search, where you can input any common word and see which topic is the most likely to address or contain the specific word. This function provides an easy and quick way for our industry partners to discover what they are looking for with their specific job or task in mind. For example, if our industry partners are looking for

researchers that have experience with microbiomes and microbiology, Topic1 would be the most likely to pertain to those subject matters. Thus the researchers that are connected to Topic1 would be the best fit for our industry partner's particular task.



However with our current interface, there are details that we noticed that can be improved upon especially in the areas of flow, navigation, and layout, such as the differing text sizes that can be distracting for viewers especially when there aren't clear sections between the faculty member, their works, abstracts, and more. With so much going on, the interface can become very cluttered, and we don't want our industry partners to be overwhelmed or even confused. Thus these are areas that we attempted to address in our current iteration of our dashboard by separating specific sections through pull down menus but need extra attention as we move forward.

With our current interface in mind, we ask ourselves: how can we improve upon the sankey diagram in a way that captures the information intuitively and efficiently? In addition, how can we make this tool useful not only for industry professionals, but also for people that want answers as quick as possible without the hassle of playing around with the diagram? Thus, as our previous work was addressed with a focus on functionality, our goals for next quarter focus on

making our dashboard more user friendly since our replication process this quarter revealed some minor deficiencies in how our interface works in real life testing.

4. Q2 Proposal

Since the resulting product of our project is a single tool, our capstone section along with our mentor, Dr. Roberts, has decided to work jointly as a team. This means that we will be creating subteams in order to divide tasks, however in the end we will combine the work that we all did into one large project. More specifically, we will be dividing our tasks amongst these teams:

- (1) Front end team: works on the resulting data visualization tool, as well as a presentation of HDSI faculty that is easier to understand for an individual that is not familiar with the tool.
- (2) Back end team: focuses on the correct labelling of topics for each HDSI faculty member.
- (3) Workflow team: Both sub-teams will conjointly work on establishing a workflow that allows for upkeep of the resulting tool.

4a. Q2 Proposal: Back-End Team

As for the back-end team, the problems are that we want to give accurate and appropriate labels for the papers of HDSI's faculty and improve the quality of our search tool by showing the essential information in the abstracts of related papers.

In Q1's study, the project has already addressed the basic labeling problems using the Latent Dirichlet Allocation (LDA) model, which assigns a topic for each document and we can use the topic-author relation to develop our dashboard. However, since the Q1 project only outputs the top words for each topic without giving a more informative and general label, we plan to work from there and introduce more information about the field of study.

Our focus is similar to Q1 in that we desire to provide labels to help users better understand faculty's fields of study. The previous labels or the topics provided by the LDA model are purely unsupervised. We worry that the model would not be robust enough when the database updates with new articles. Now, instead of only focusing on the LDA model, we will also introduce the

research fields from Google Scholar (Fig. 1) for each faculty by web scraping using Selenium plus Scrapy, and make use of the category features from the Dimensions API to cross-reference the labels for documents. In this case, the faculty's domain expertise will be well captured by the research fields provided in Google Scholar as well as the topic terms obtained from our topic model.

Yusu Wang

Professor, HDSI, Univ. California, San Diego.

Verified email at ucsd.edu - [Homepage](#)

[Computational geometry](#) [computational topology](#) [data analysis using geomet...](#)

Figure 1: the fields of study from Google Scholar (words in blue)

Another goal for the back end is that we want to make particular useful information stand out from the abstracts of papers. In Q1's study, they provide the abstracts for those most related papers. But the situation can be that it might take the industry partner a while to look through all the abstracts of the papers, especially if they might not have the expertise in those domains. However, they know what is the task of their industrial project and what can be the named-entity information related to the project. If we highlight that important information, we might better interpret and understand the work of a faculty.

Thus, we want to use Named-entity recognition (NER) from nltk to identify special information for the abstracts. For example, a specific term like the location of study, the time for investigation, product name, etc. After having those words, we are going to highlight them in the abstracts of papers, so that in our interface the industry partner can scan the abstract of most related papers and know the specific circumstances for a study in a more efficient and quicker manner.

4b. Q2 Proposal: Front-End Team

The current UI for our dashboard simply contains vital functionality and features that we want to remain, but lacking in visual appeal and a general sense of intuitiveness. We will solve this issue in our Q2 project by color coding the topics on our present tool, changing the fonts and

modifying the search bar by hiding topic descriptions, making the diagram less confusing. This will allow both HDSI workers and industry partners to produce a search that is both visually appealing and user-friendly. When it comes to the front end aspects, we also want to expand upon our tool by incorporating visual aids that allow our industry partners to become more familiar with our faculty. We have considered the fact that a number of people do not have a technical background and there is a possibility that they have never seen a Sankey diagram before. Using Figma, a collaborative tool, we will design interactive visual aids that present our faculty along with their pictures and fields of study. We will eventually implement our Figma design using HTML or CSS during the course of the next quarter.

4c. Q2 Proposal: Workflow Team

In terms of workflow, faculty publishes articles, reports and papers yearly. Our intention for this team within the Q2 Project is to facilitate the acquisition of this new data easily for future HDSI workers and to create a tool that is usable for HDSI when working with industry partners. We will be expanding our Q1 outcomes by designing a workflow that automates the process and a tool that is hopefully implemented and displayed to aid not only HDSI members, but industry professionals as well. Within this workflow, we'd like to incorporate available cloud services provided by UCSD in order to allow for the quick and easy insertion and deletion of faculty within our database. Also on the cloud, we'd like to host our files in order to make it easier to run the file that derives our topics.

5. Conclusion

In conclusion, we will be expanding upon our Quarter 1 project by automating the labeling of our diagram using research fields on Google Scholar using Selenium plus Scrapy, cross-referring labels for documents continuing the use of Dimensions API. We will also modify the Sankey diagram to make it more visually appealing: changing color, fonts, modifying the search bar, etc. The front-end team will also be working on a visualization that can be used directly by industry partners to become more familiar with our faculty along with developing more visualization to accompany the Sankey diagram. We strive to make this tool deployable and for it to be kept up-to-date, so we will be working on a pipeline that facilitates future work on our tool.