HDSI Faculty Exploration Tool using LDA Topic Modeling

Du Xiang
Halıcıoğlu Data Science Institute
University of California
La Jolla, CA 92093
dxiang@ucsd.edu

Siddhi Patel
Halıcıoğlu Data Science Institute
University of California
La Jolla, CA 92093
srpatel@ucsd.edu

Martha Yanez
Halıcıoğlu Data Science Institute
University of California
La Jolla, CA 92093
mayanez@ucsd.edu

Sijie Liu
Halıcıoğlu Data Science Institute
University of California
La Jolla, CA 92093
sjliu@ucsd.edu

Brian Qian
Halıcıoğlu Data Science Institute
University of California
La Jolla, CA 92093
brqian@ucsd.edu

Abstract

Halıcıoglu Data Science Institute industry partners are constantly looking for faculty to work on their projects. In order to make this an easier process, we used Latent Dirichlet Allocation to find the most representative topics written by HDSI faculty. We processed the abstracts of their published works in order to run an LDA model on them. We found the topics most associated with them and created a tool that effectively shows the connection between the author and topics. 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 have also made a plan for the creation of an additional tool that allows for an easy search of faculty.

1 Introduction

The Halıcıoglu Data Science Institute (HDSI) at the 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 with 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 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 start the plan 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.

2 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 dashboard 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.

3 Literature Review

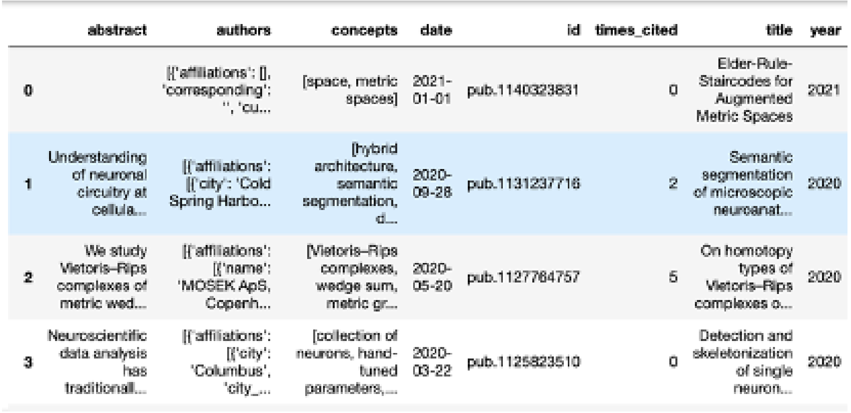
As mentioned in section a., in order to perform the replication of the exploration tool, we used texts by the aforementioned authors. "Latent Direct Allocation" is a paper by D. Blei, A. Ng and M. Jordan that explains in detail the origin and process of LDA. It explains the motivations of modeling text 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 the Towards Data Science publication. 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. Both of the articles provide a step-by-step explanation on how to perform the process, therefore the combination of them was very useful for the present report.

4 Data

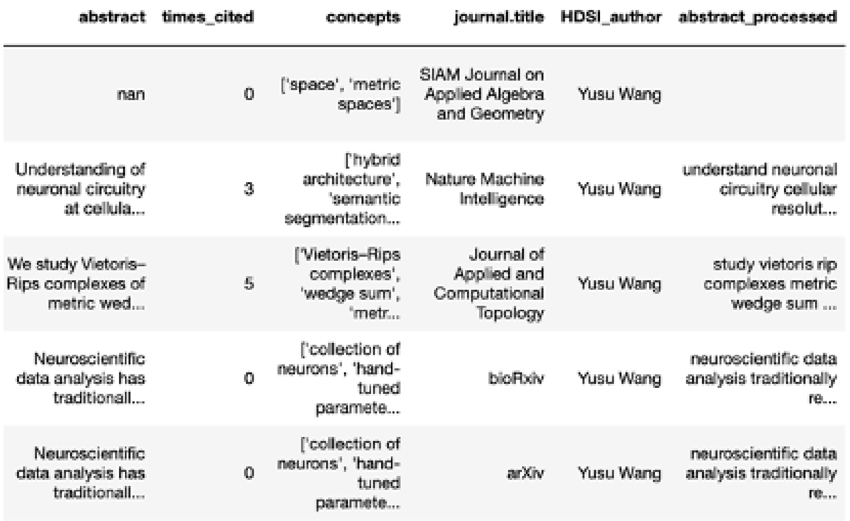
In order to create the tool, an analysis on text data was performed. We decided to use only the abstract portions of published faculty works and perform LDA on this. The data was obtained from Dimensions, which is a platform that allows for the search and analysis of over 150 million interlinked data items from across the research world. Such data items include publications, grants, clinical trials, patents, and policy documents. 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': '', 'cu...	[space, metric spaces]	2021-01-01	pub.1140323831	0	Elder-Rule-Staircodes for Augmented Metric Spaces	2021
1	Understanding of neuronal circuitry at cellula...	[{'affiliations': [{'city': 'Cold Spring Harbo...	[hybrid architecture, semantic segmentation, d...	2020-09-28	pub.1131237716	2	Semantic segmentation of microscopic neuroanat...	2020
2	We study Vietoris–Rips complexes of metric wed...	[{'affiliations': [{'name': 'MOSEK ApS, Copenh...	[Vietoris–Rips complexes, wedge sum, metric gr...	2020-05-20	pub.1127764757	5	On homotopy types of Vietoris–Rips complexes o...	2020
3	Neuroscientific data analysis has traditionall...	[{'affiliations': [{'city': 'Columbus', 'city_...	[collection of neurons, hand-tuned parameters,...	2020-03-22	pub.1125823510	0	Detection and skeletonization of single neuron...	2020

Figure 1: Initial dataset obtained by using Dimensions API

Preprocessing the abstracts by using stemming, removing stop words and the gensim simple preprocess, 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.



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...

Figure 2: Dataset including Processed abstracts and HDSI authors.

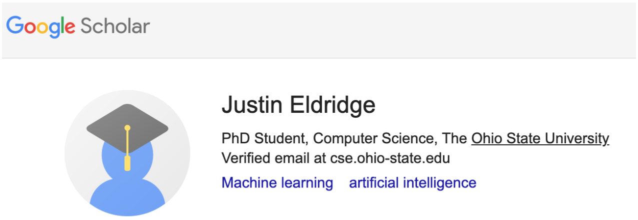
Figure 2 presents the first five rows of the final dataset used, excluding the columns *year*, *authors* (list form), and *title*.

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 the 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 labeling for authors and topics.

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 in Figure 3. Google Scholar also has information on faculty that Dimensions does not, as well as some of their missing articles.



Google Scholar

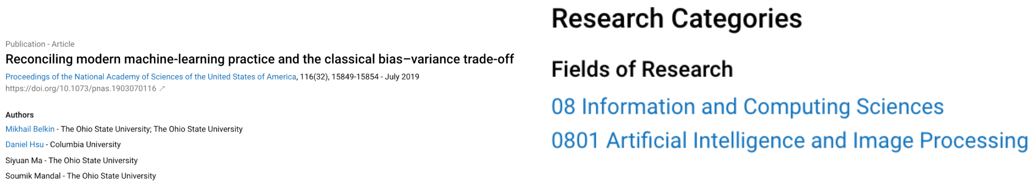
Justin Eldridge

PhD Student, Computer Science, The Ohio State University
Verified email at cse.ohio-state.edu

Machine learning artificial intelligence

Figure 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.



Publication - Article

Reconciling modern machine-learning practice and the classical bias−variance trade-off

Proceedings of the National Academy of Sciences of the United States of America, 116(32), 15849-15854 - July 2019
https://doi.org/10.1073/pnas.1903070116 ↗

Authors

Mikhail Belkin - The Ohio State University; The Ohio State University
Daniel Hsu - Columbia University
Siyuan Ma - The Ohio State University
Soumik Mandal - The Ohio State University

(a)

Research Categories

Fields of Research

08 Information and Computing Sciences
0801 Artificial Intelligence and Image Processing

(b)

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

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

1) We scraped the labels on the Google Scholar profiles of faculty in HDSI and gained a histogram providing a general view of how they are distributed.

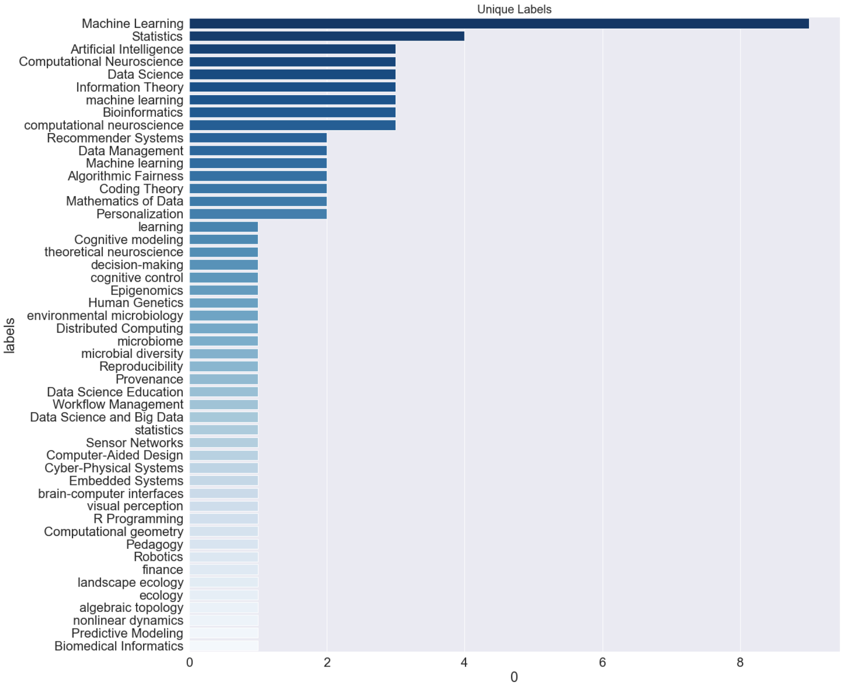


Figure 5: Unique labels gathered from the Google Scholar profiles.

As useful as these Google Scholar unique labels are for understanding in a broader context what authors publish, we found that they might not be able to provide an explanation as specific as we require for our current work. For example, "Machine Learning" is a very extensive field; a researcher can have a focus on supervised learning, unsupervised learning, reinforcement learning, or the application of a specific algorithm relating to their particular interest area. Therefore, we decided to expand our work to the second version of labeling.

2) We also gathered labels from the Dimensions API and we found that they provide a 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.

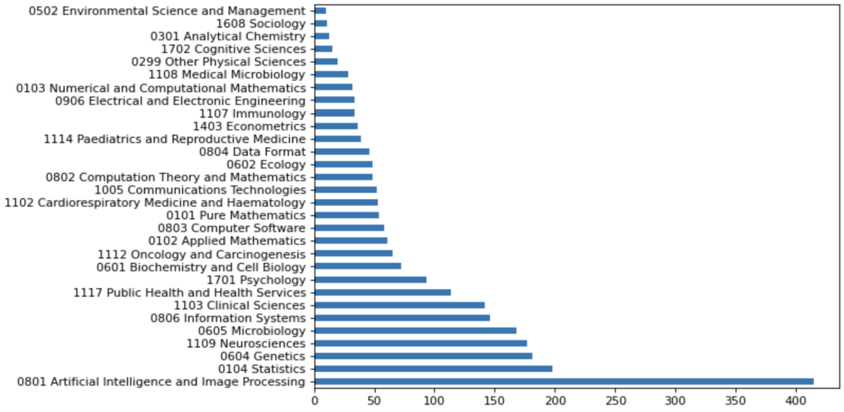


Figure 6: Distribution of labels obtained from Dimensions at the article level.

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

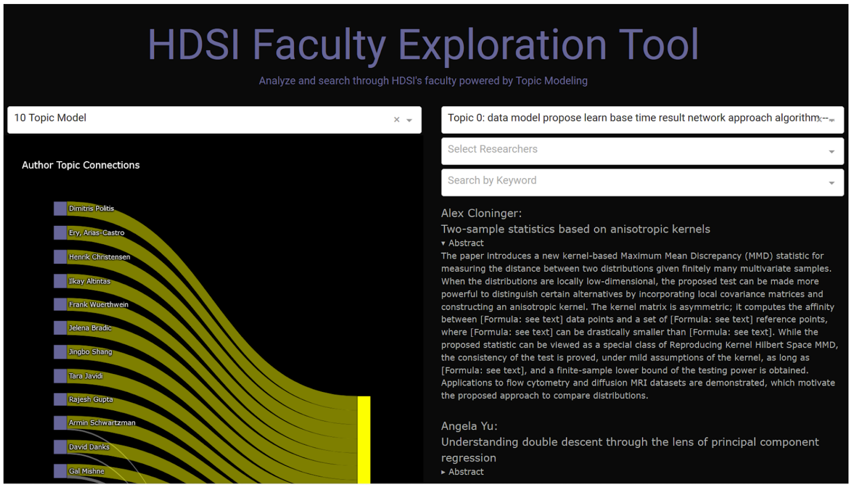
For version 2.0 of our dashboard, we combined option 1 and option 2 to provide a better user experience for our search tool. To make our results easier to understand, we provide the article-level labels as related fields next to our generated topics. This is helpful because it does not replace the generated words but provides a broader context to them. It is easier for the user to make sense of the topics without losing the essence of them.

Since the user can also select a particular researcher on our Sankey diagram, we decided to use the labels obtained from Google Scholar in this section.

Having labels that only work on a specific set of data would not be useful. We worked on improving our pipeline to be more adaptable to the data stream. Ideally, the data that we are using to generate our dashboard will be updated once a year, not only to account for new faculty additions but to include the most recent publications. This means that our code must also produce meaningful results in a robust manner. Our pipeline was previously tailored to the newer version of data with specific year thresholds. However, we fixed this issue and improved our data pipeline, making our code maintainable and easy to modify and expand. Our ideal goal was that as mentioned previously, the tool will be useful and robust enough to handle any new incoming data.

6 Results

As shown below, our current dashboard has many useful functions that provide a great amount of information. We are certain that this tool will be exceptionally helpful towards our industry partners since it has features such as the topic selection where the proprietary topics are listed to demonstrate the variety of areas of research our HDSI faculty has to offer. As a topic is selected, they will be able to find a list of the faculty with their respective papers, along with their abstracts, to give a general sense of what each faculty member is working on.



10 Topic Model

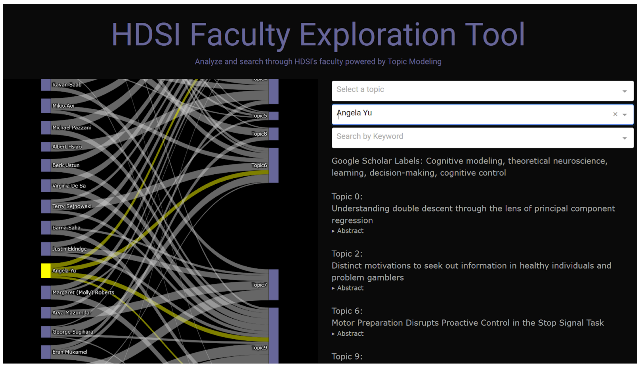
Topic 0: data model propose learn base time result network approach algorithm--

Alex Cloninger:
Two-sample statistics based on anisotropic kernels
▾ Abstract
The paper introduces a new kernel-based Maximum Mean Discrepancy (MMD) statistic for measuring the distance between two distributions given finitely many multivariate samples. When the distributions are locally low-dimensional, the proposed test can be made more powerful to distinguish certain alternatives by incorporating local covariance matrices and constructing an anisotropic kernel. The kernel matrix is asymmetric; it computes the affinity between [Formula: see text] data points and a set of [Formula: see text] reference points, where [Formula: see text] can be drastically smaller than [Formula: see text]. While the proposed statistic can be viewed as a special class of Reproducing Kernel Hilbert Space MMD, the consistency of the test is proved, under mild assumptions of the kernel, as long as [Formula: see text], and a finite-sample lower bound of the testing power is obtained. Applications to flow cytometry and diffusion MRI datasets are demonstrated, which motivate the proposed approach to compare distributions.

Angela Yu:
Understanding double descent through the lens of principal component regression
▸ Abstract

Figure 7: Topic Selection feature and list of faculty and their related articles

The researcher selection function is another interesting way of discovering where a specific faculty member's line of research actually lies. As an example, in Angela Yu's case, with a 10 topic model, her papers demonstrate an expansive variety of topics from Applied Mathematics, to Psychology and Neuroscience.


HDSI Faculty Exploration Tool

Analyze and search through HDSI's faculty powered by Topic Modeling

Select a topic

Angela Yu

Search by Keyword

Google Scholar Labels: Cognitive modeling, theoretical neuroscience, learning, decision-making, cognitive control

Topic 0:
Understanding double descent through the lens of principal component regression
▸ Abstract

Topic 2:
Distinct motivations to seek out information in healthy individuals and problem gamblers
▸ Abstract

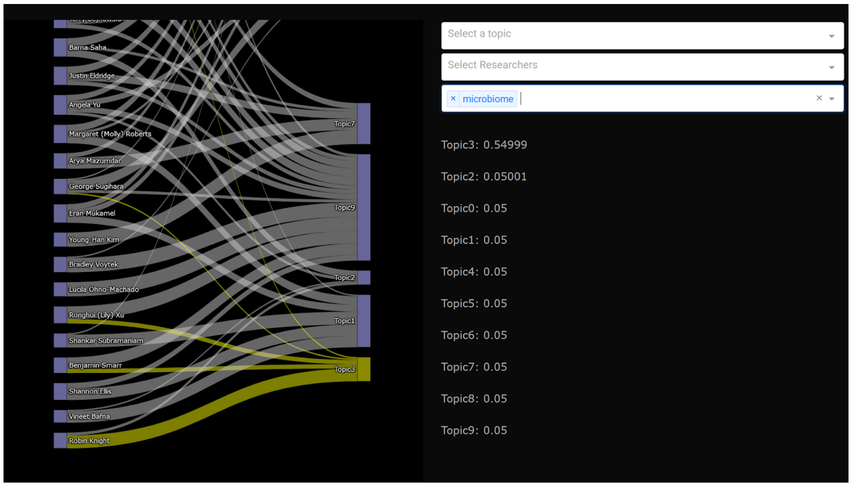
Topic 6:
Motor Preparation Disrupts Proactive Control in the Stop Signal Task
▸ Abstract

Topic 9:

Figure 8: Faculty Selection feature and list of the topics they have written on as well as their respective articles.

As can be seen in Figure 8, as Angela Yu's name is selected, you can also see the added Google Scholar Labels for her fields of research. It is important to note that not all authors had these labels on their profiles, so a minority of them will be blank.

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 the user is looking for researchers that have experience with microbiomes and microbiology, Topic3 would be the most likely to pertain to those subject matters, as seen on Figure 9. Thus the researchers that are connected to Topic3 would be the best fit for our industry partner's particular task. In this case, we can observe that the highlighted flow points to George Sugihara, Ronghui Xu, Benjamin Smarr, and Robin Knight.


Select a topic

Select Researchers

× microbiome

Topic3: 0.54999

Topic2: 0.05001

Topic0: 0.05

Topic1: 0.05

Topic4: 0.05

Topic5: 0.05

Topic6: 0.05

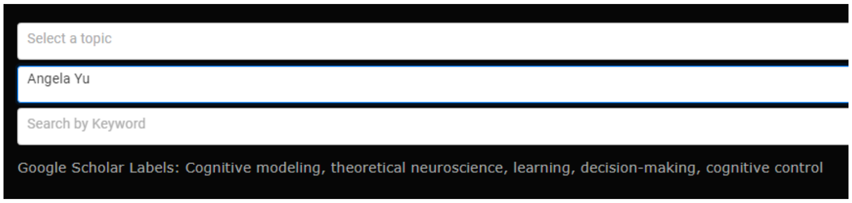
Topic7: 0.05

Topic8: 0.05

Topic9: 0.05

Figure 9: Word Search feature and list of the topics that are more associated with it. On the left-hand side, the authors that are connected to this topic.

In addition to the dropdown navigation, the enhancement on the labels allow the users to explore the content of the HDSI faculty a step further. As mentioned, the Google Scholar fields of each author are extracted to help users to have a broad understanding of the author's work. One benefit of using this field is that these labels are created by the researchers themselves, which makes them a reliable source and a good representation of their general work. In the following example, we can clearly see that Angela Yu is a researcher dedicated to the cognitive sciences and neuroscientific domains.


Select a topic

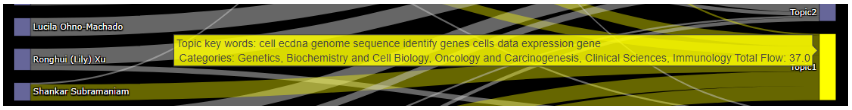
Angela Yu

Search by Keyword

Google Scholar Labels: Cognitive modeling, theoretical neuroscience, learning, decision-making, cognitive control

Figure 10: Google Scholar labels as you select a particular researcher.

Aside from the researcher level labels, the article level labels were extracted from Dimensions to generate automatic labels for the topics. We did this by first aggregating the articles on the topic level and then ranking them by the frequency of the categorical labels. By picking the top frequent labels in each topic, we can generate a series of labels for each topic that is more readable than the key words from LDA. The topic labels are formulated to be represented here in our UI:



Topic key words: cell ecdna genome sequence identify genes cells data expression gene
Categories: Genetics, Biochemistry and Cell Biology, Oncology and Carcinogenesis, Clinical Sciences, Immunology Total Flow: 37.0

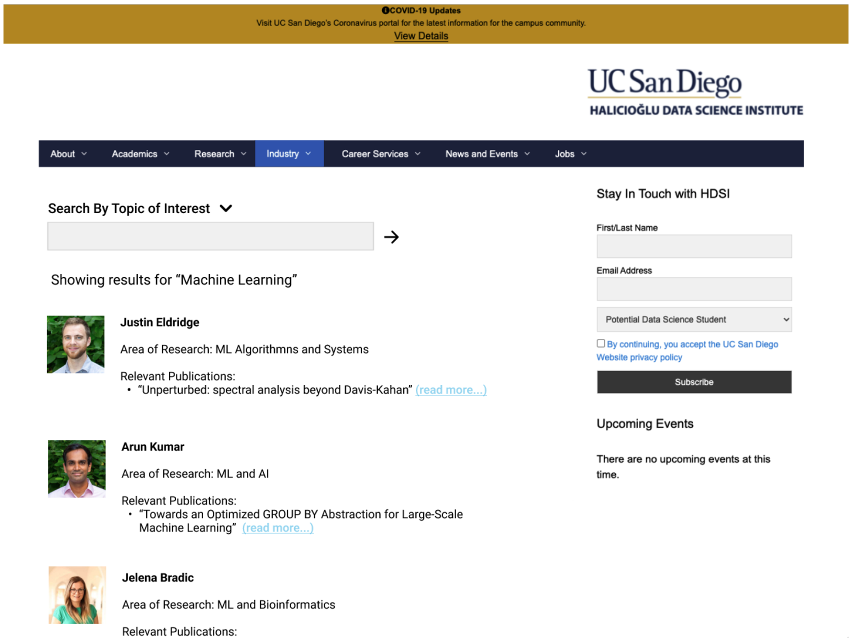
Figure 11: Category labels shown as you hover over each topic on the Sankey diagram.

7 Discussion and Future Work

LDA allowed us to obtain a proportion of topics for each author and it was helpful to understand where their expertise lays on. We believe that HDSI industry partners will find our tool helpful and accessible, and will be able to get familiarized with it quickly. The final version tool allows for word search, topic and researcher selection. When searching for authors it displays the topics they have written on and their abstract corresponding to the particular topic.

We developed a workflow that will aid future HDSI team members in obtaining the data required for the tool easily. We also worked on automating the labeling process by obtaining categories from both Dimensions and Google Scholar.

For future endeavors, we have imagined a different tool that will aid industry partners that might not be familiar or are having difficulties with the Sankey Diagram tool. 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's faculty. The tool is split into three sections: Search by Topic of Interest, Search by Keywords, and Sankey Diagram. While this is still a work in progress, in our Figma demo we demonstrate some of the key features that we are further interested in developing along the line.


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
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.

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...)

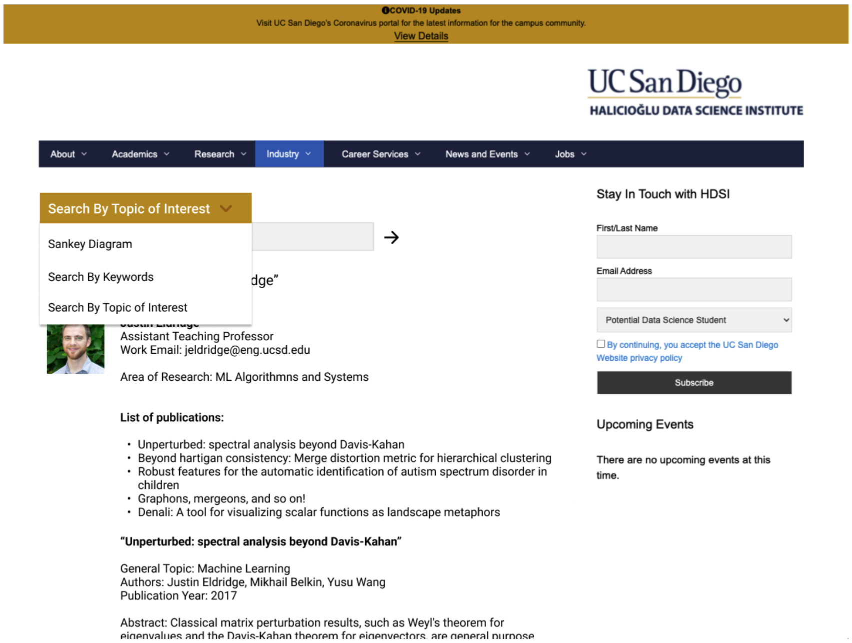
Jelena Bradic

Area of Research: ML and Bioinformatics

Relevant Publications:

Figure 12: A diagram of the general look of a future Easy Search Tool.

Our main focus was keeping the overall layout as simple as possible for ease of accessibility especially for those who may not be familiar with topic modeling. Thus our easy search tool resembles a search engine-like format that matches the current theme of the HDSI website. This search bar tool will be useful to explore the topics and interests that pertain to HDSI faculty similar to the search by keyword function within the current dashboard. The resulting information however is laid out in a more intuitive structure with a profile picture alongside the faculty member's name, area of research, and their most relevant publication. Our goal was to provide a more polished experience following a more modern aesthetic. Additionally, we conceptualized a profile style page that expanded upon each faculty member's publications, abstracts, and their contact information.


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
Sankey Diagram
Search By Keywords
Search By Topic of Interest
dge"
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.

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

Figure 13: Faculty profiles including research and previously obtained labels on Easy Search Tool.

We explored many different pathways of accomplishing this concept, from testing out Observable, RxJS, and HTML, but with the limited time schedule we decided to put more focus on revamping the current dashboard. Our original plan was to use a conjunction of Observable and RxJS. Observable is a data visualization tool that uses minimal conventions to create custom visualizations. We planned on using Observable in order to rebuild our Sankey diagram to match the aesthetics of the HDSI website. While we thought RxJS would be ideal to rebuild the "Search by Keywords" and "Search by Topic of Interest" section since it's a reactive programming language that makes it easy to compose asynchronous and event-based programs. Future work would be focused on integrating our easy search tool on the HDSI website itself which would give access to a wider audience to explore and play around with topic modeling in a simple to use interface.

References

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet Allocation". In: *J. Mach. Learn. Res.* 3.null (Mar. 2003), pp. 993–1022. ISSN: 1532-4435.

[2] Prabhakaran, Selva. "LDA - How to Grid Search Best Topic Models? (with Examples in Python)." Machine Learning Plus, 9 Oct. 2021, https://www.machinelearningplus.co m/nlp/topic-modeling-pythonsklearn-examples/.

[3] Kapadia, Shashank. "Topic Modeling in Python: Latent Dirichlet Allocation (LDA)." Medium, Towards Data Science, 29 Dec. 2020, https://towardsdatascience.com/endto-end-topic-modeling-in-pythonlatent-dirichlet-allocation-lda35ce4ed6b3e0.

8 Appendix

Our Project Proposal: https://docs.google.com/document/d/1yKhGqM42MS7HRj8AdxZsOiU5SWhWauZCBOY-DacxsD8/

Our Github Repository: https://github.com/MarthaY01/hdsi_faculty_tool

Our Project Website: https://marthay01.github.io/hdsi_faculty_tool/

Our raw data can be found on: https://github.com/IreneLiu2018/capstone_a14/blob/master/New_Vis/data/raw/final_hdsi_faculty_updated.csv

Figma demo for future Easy Search Tool Press the 'Play' button on the upper right, you are able to hover over to 'Industry' > 'Faculty Exploration Tool' > search arrow > Dr. Justin Eldridge's profile https://www.figma.com/file/WO5QJnJrALVwI2xu1NfgAg/Main?node-id=4%3A50