Machine Learning Techniques for performing Topic Modeling and Identification on Bibliographic Datasets

William H. Mischo (w-mischo@Illinois.edu)
Elisandro Cabada

December 14, 2021
CNI Fall Briefing
Washington, DC

ILLINOIS
University Library

Environmental Scan

We have tools to design and implement interconnected services: discovery, bibliometric, and LOD services:

Robust A&I, database, vendor, and repository APIs returning JSON and XML;

DOIs as glue and other persistent identifiers;

Asynchronous and GPU parallel processing;

Open bibliographic and research datasets that provide data/content.

We want Machine Learning computational tools (clustering, classification, regression, dimension reduction) to add to the armamentarium.

ML and AI in Library Services?

We focus on systems and service frameworks that provide the scaffolding for interconnected resources and services. Bento-style discovery systems.

Want to look at ML in bibliographic, discovery, bibliometric & research impact, and visualization services and systems.

Began with a question: can we add a topic modeling component to our API-based bibliographic database service? And what about bibliometric and discovery & delivery services?

Machine Learning

ML takes sets of observations, identifies patterns and anomalies in the observations, and saves the results as a mathematical model, an n-dimensional array of vectors.

Focus here on document clustering to identify key concepts in a corpus of documents, to partition the corpus into groups of related documents.

Tools to extract words, significant phrases, and entities from sentences and article abstracts for presentation to clustering software.

In parallel: Build word and phrase indexes for vectorization and similarity measures.

Clustering and Topic Modeling

Clustering is Unsupervised; vs. Text Classification which is Supervised and uses a pre-defined set of training data to partition the corpus into related documents by assigning examined documents into one of the pre-established classes.

All ML is numerical, so need to Vectorize documents to convert the document representation to numbers to use similarity or distance measures to find related documents.

Bag of words approach vs. text as sequences (phrases).

Explainability and reproducibility are critical.

Project datasets and databases require the processing, cleaning, and preparation of data in order to use ML techniques.

Clustering and Classification Tools

Off the shelf clustering tool environments: are any ready for prime time?

Scikit-learn; TensorFlow; alteryx, Microsoft Azure and Cognitive Toolkit, RunwayML, Google AI Platform (Teachable Machine, AutoML Vision); Wolfram Mathematica; MALLET; Amazon Comprehend; Topic Modeling Tool; Stanford Topic Modeling Toolbox; Keras Deep learning.

Unsupervised algorithms: K-Means; K-Means++, K-Medoids, Latent Dirichlet Allocation (LDA); Kernel K-Means; Spectral Clustering. Choice of initial K clusters important. Dimensionality Reduction (Principal Component Analysis PAC).

Supervised algorithms: Classification (Naive Bayes, Decision Trees, K-Nearest Neighbor), Regression (Linear, Ridge, Least Squares).

AI and ML in Libraries

Notre Dame IMLS: Machine Learning, Libraries, and Cross-Disciplinary Research: Possibilities and Provocations and Investigating the National Need for Library Based Topic Modeling Discovery Systems White Paper
LoC: Machine Learning + Libraries Summit Event Summary and Ryan Cordell Machine Learning +Libraries: A Report on the State of the Field. IDEA Institute on AI, IMLS funded, first cohort.

Ryan Cordell Report, **Machine Learning + Libraries** commissioned by LoC.
Library and librarians expertise in classification, collections, and data make libraries an ideal testbed for ML.
"…libraries could become focal sites for the translation and collaboration that will be required to cultivate responsible machine learning…"

The Law of the Hammer analogy: It is important to pause and consider if artificial intelligence techniques are the best approach before trying to use them. –Hansen 164

Inside AI reports: **U.S. lawmakers, both Democrats and Republicans, said they support regulations of facial recognition software used by law enforcement agencies.** Incorrect facial recognition used by Detroit police led to a wrongful arrest.

Hype and Hope

ML has been hyped and promoted for years. –Lesk 107. Hansen lists 5 articles from 2014 to 2017 discussing the revolutionary nature of ML.

DeepMind, Alphabet's AI subsidiary, has partnered with the Geneva-based Drugs for Neglected Diseases initiative (DNDi) to find cures for neglected diseases like sleeping sickness.

The Library of Congress has hired computer science and digital humanities experts to explore how AI could improve its search technology. Expect new and improved online search prototypes by early next year.

Language models analyze and generate new text based on prompts. GPT-3, MT-NLG, DeepMind. Medical AI.

Google Flu Trends – great fanfare. AI in radiology problems. Sepsis early warning system missed two-thirds of sepsis cases: JAMA Internal Medicine. **YouTube's suggestion algorithms consistently direct users to videos with misinformation and sexualized content, including videos later taken offline, according to a Mozilla study.** Zillow's algorithm. Facebook's algorithms.

Augmented Intelligence is the new term.

Issues with AI

ACM Queue article on bias in AI:
https://queue.acm.org/detail.cfm?id=3466134

Coded Bias documentary, Cathy O'Neil, *Weapons of Math Destruction*, Algorithmic Justice League at MIT

News column in *Science* in 2018 in which Ali Rahimi (Google) and Ben Recht (Berkeley), and other computer scientists compare AI to alchemy. "Researchers do not know why some algorithms work and others don't, nor do they have rigorous criteria for choosing one AI architecture over another." Rahimi and colleagues document examples of what they see as the alchemy problem and offer prescriptions for bolstering AI's rigor.
https://www.science.org/content/article/ai-researchers-allege-machine-learning-alchemy

A 2021 article in Quanta Magazine building on these criticisms and arguing that science and engineering have often operated this way:
https://www.quantamagazine.org/science-has-entered-a-new-era-of-alchemy-good-20211020/

U.S. Congress and U.K. Parliament proposed bills requiring companies to provide algorithm-free content on request.

Clustering Comments

Clustering is a very important method but is only a well-designed algorithm that is not adaptive. –Hintze and Schossau

When using a topic modeler, it is important to iteratively configure and re-configure the input until the results seem to make sense.. –Lease Morgan

…with topic modeling, a scholar must interpret what a particular algorithm has defined as a statistically significant topic by interpreting a cryptic chain of words. - Janco

When it comes to topic modeling, there is no such thing as the correct number of topics. – Lease Morgan

The algorithm will not tell you what the themes or topics are but will show which articles group together. It is then up to the researcher to work out the common thread. - Altman

Attempts to use algorithmic methods to describe collections must embrace the reality that, like human descriptions of collections, machine descriptions come with varying measures of certainty. – Padilla

Illinois Example: Literature Review with ML Component

Machine Learning techniques are being used to augment bibliometric and database building services.

The Library uses the Scopus Application Programming Interface (API) to write scripts to query and retrieve bibliographic data from Scopus. This project downloaded metadata from 57,000 scientific articles on biofuels and stored the data in a custom database.

ML document clustering is done within Comparative Text Mining (CTM) algorithms developed by a CS faculty member that were used to identify the key concepts and perform a sentiment analysis from the data. Compared with K-Means.

ELSEVIER

Renewable and Sustainable Energy Reviews

Volume 133, November 2020, 110265

How do the research and public communities view biofuel development?

Qiankun Zhao [a, b], Ximing Cai [a, b], William Mischo [c], Liyuan Ma [a]

Abstract

Understanding how public and research view biofuel development, especially the unsynchronized views, can explain the gaps between the actual biofuel production and consumption and the government mandates in the US. Applying a comparative text mining technique, these views are explored using 9924 news articles and 57,849 research abstracts. It is found both the public and research communities respond actively to major policy and incentive programs and market conditions and the public has been closely following important research advances. However, research and practice should be coordinated to achieve economies of scale for regular and alternative biofuels. Priority is particularly needed for policy research to support legislation and enforcement and remove technical barriers to commercialization; meanwhile, greater attention is needed to prompt the commercialization of mature technologies. In addition, sentiment analysis shows positive public perceptions in general and negative perceptions primarily stemming from fraud in biofuel tax-credit programs with minor concerns on unintended negative impacts of biofuel development and policy implementation issues. This study contributes to methodology by using cutting-edge text-mining

Biofuels 57,000 papers -- K-Means

Cluster 0:
ethanol
pretreatment
fermentation
production
biomass
lignin
hydrolysis
cellulose
lignocellulosic
bioethanol

Cluster 1:
bio
pyrolysis
oil
biomass
catalytic
temperature
fast
catalyst
yield
oils

Cluster 2:
biodiesel
oil
transesterification
reaction
catalyst
production
acid
fatty
methanol
methyl

Cluster 3:
diesel
engine
fuel
biodiesel
emissions
blends
combustion
emission
fuels
oil

Cluster 4:
biogas
anaerobic
digestion
methane
sludge
waste
production
manure
organic
treatment

Cluster 5:
biofuel
biomass
production
cell
fuel
gt
lt
high
plant
using

Cluster 6:
microalgae
lipid
algal
production
algae
biomass
growth
microalgal
cultivation
wastewater

Cluster 7:
error
energy
biofuels
production
biofuel
biomass
land
use
fuels
fuel

CTM and Results

CTM is an extension of the probabilistic latent semantic indexing model (PLSI) a model that uses clustering. Topics are generated based on a given collection of text data. Uses sampling across a number of topics, where each topic is a probability distribution of words.

The algorithm produced 12 clusters such as L*ow-Emission Diesel* from the common keywords *diesel, emission, fuel, engine, gas, air, reduction, vehicle, blend, low* and *Fuel Cells* from *cell, electrode, membrane, electron, surface, power, density, electrochemical, transfer, anode.* These clusters were evaluated with the K-Means clusters.

Research Impact Visualizations and Underlying Databases

We have developed research impact visualizations or dashboards of research indicators of individual faculty in a department or group with clickable bubbles and custom databases. Database-driven, interactive, web-based, display.

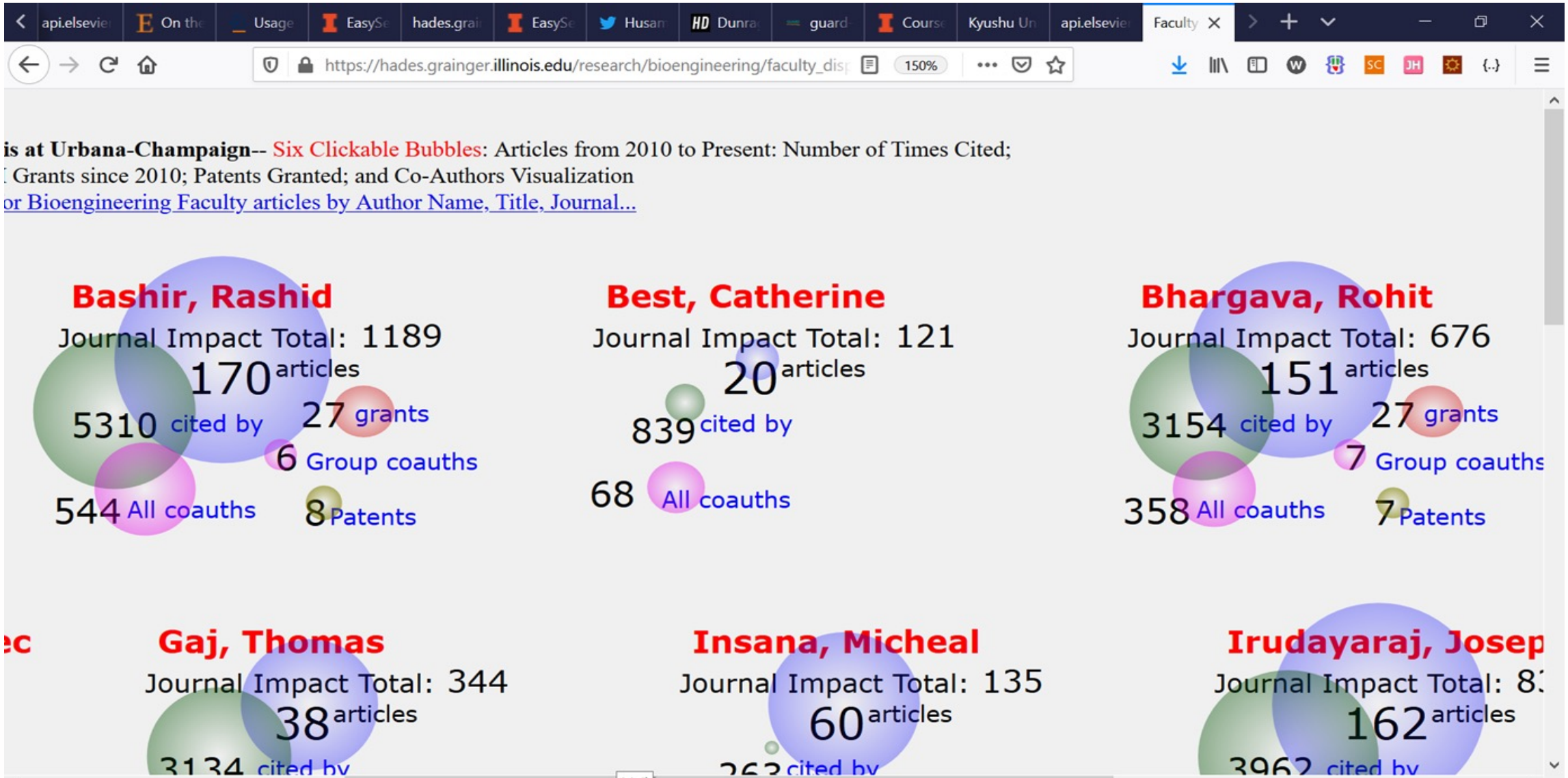
Next slide shows the visualization for the Cancer Center of Illinois 104 faculty.

Have a larger database of 500 faculty research impact indicators from nine departments or groups which allows calculating correlations between indicators.

Database of metadata of all publications over the last 10 years for all faculty. Use of clustering techniques to derive key research fronts or topics.

Cancer Center of Illinois Faculty

https://iis-demo.library.illinois.edu/research/bioengineering/faculty_display.asp



Cancer Center Topic Extraction - 8,980 papers, K-Means

Cluster 0:	Cluster 1:	Cluster 2:	Cluster 3:	Cluster 4:	Cluster 5:	Cluster 6:	Cluster 7:
cell	data	chemical	detection	oct	imaging	membrane	error
cells	gene	inf	surface	optical	image	dna	jpc
cancer	information	binding	photonic	coherence	tissue	lipid	editorial
tumor	methods	acid	gold	tomography	method	protein	january
drug	based	redox	nanoparticles	imaging	wave	proteins	discussion
breast	model	reaction	crystal	optics	images	dynamics	periodic
expression	network	peptide	quantum	tissue	reconstruction	nanopore	general
gene	tree	society	using	tm	phase	simulations	issue
delivery	learning	structure	emission	ear	resolution	molecular	introduction
receptor	problem	polymer	based	adaptive	optical	binding	erratum

Cancer Center research topics clustering interpretation

This is the consensus topic modeling for the SciKitLearn K-Means document clustering analysis over the 8980 articles. 8 clusters, need for domain knowledge assistance:

0-- cancer cells, tumors, cell receptors, breast cancer
1-- gene models, genetic expression, gene trees
2-- redox reactions, chemical binding, peptides, polymers
3-- photonic crystals, biophotonics, nanoparticles
4-- optical coherence tomography, optics, ear
5-- tissue imaging, image resolution, optical imaging
6-- DNA, proteins, molecular binding
7-- miscellaneous and problematical

Key Point: There is often a need for domain knowledge and disciplinary experts in the interpretation of clustering results.

Text as Sequences (Phrases)

The two Clustering examples use a form of the K-Means clustering algorithm which processes individual words – text as a bag of words.

Text documents can be stored as sequences of words or phrases. The text documents are vectorized to indicate the presence or absence of the words or the number of times a word/phrase appears or a natural log representation for each term to be calculated and stored.

Our assumption is that using phrases instead of individual words will better capture the semantic meaning of the documents and help us to better do topic modeling. The primary clustering techniques that use text as sequences (phrases) typically employ either a kernel k-means or a spectral clustering technique.

Example: a Discovery Dataset

Generated a sample Discovery Set of 435 documents retrieved using the Scopus API on the search:

TITLE-ABS-KEY((web scale discovery librar*) OR (web-scale discovery librar*)) OR TITLE-ABS-KEY(("discovery service*" OR "discovery system" OR "discovery tool*") AND librar*) OR TITLE-ABS-KEY(bento AND (librar* OR discovery))
yr=2012-2021

Database has Word and phrase index representations. Phrase extraction was done with the Microsoft Azure Cognitive tools. The MS Cognitive Services Text Analysis Tool API was used to extract phrases, significant words, and entities. MS Cognitive Tools – speech synthesis, speech recognition for chatbot and this Text Analysis tool. Others from Amazon, Google.

.

Phrase Extraction from Abstracts

MS Cognitive Services Text Analysis Tool API to extract phrases, significant words, and entities.

abstract_phrase_extraction_documents_library_discovery.csv [Read-Only] - Excel

C19: "deep discovery", "end-user information discovery", "Web-scale discovery services", "libraries", "library's local", "licensed content", "revolution", "library

	B	C
19	Web-scale discovery services for libraries provide deep discovery to a library's local and licensed content and represent an evolution-perhaps a revolution-for end-user information discovery as pertains to library collections. This article frames the topic of web-scale discovery and begins by illuminating web-scale discovery from an academic library's perspective-that is, the internal perspective seeking widespread staff participation in the discovery conversation. This included the creation of the Discovery Task Force, a group that educated library staff, conducted internal staff surveys, and gathered observations from early adopters. The article next addresses the substantial research conducted with library vendors that have developed these services. Such work included drafting of multiple comprehensive question lists distributed to the vendors, onsite vendor visits, and continual tracking of service enhancements. Together, feedback gained from library staff, insights arrived at by the Discovery Task Force, and information gathered from vendors collectively informed the recommendation of a service for the UNLV Libraries.	"deep discovery", "end-user information discovery", "Web-scale discovery services", "libraries", "library's local", "licensed content", "revolution", "library collections". "web-scale discovery", "discovery conversation", "internal perspective", "topic of web", "academic library's perspective", "widespread staff participation", "article". "library staff", "internal staff surveys", "observations", "group", "Discovery Task Force", "creation", "early adopters". "substantial research", "library vendors", "addresses", "article", "services". "vendors", "onsite vendor visits", "drafting of multiple comprehensive question lists", "continual tracking of service enhancements", "work". "insights", "information", "Discovery Task Force", "service", "recommendation", "feedback", "library staff", "vendors", "UNLV Libraries".
	This article explores the basic principles of web-scale discovery systems and how they are being implemented in libraries. "Web scale discovery" refers to a class of	"basic principles of web", "scale discovery systems", "article", "libraries". "vast number of resources", "single search box", "wide variety formats", "electronic journal collections", "class of products"

Clustering on the Discovery Dataset

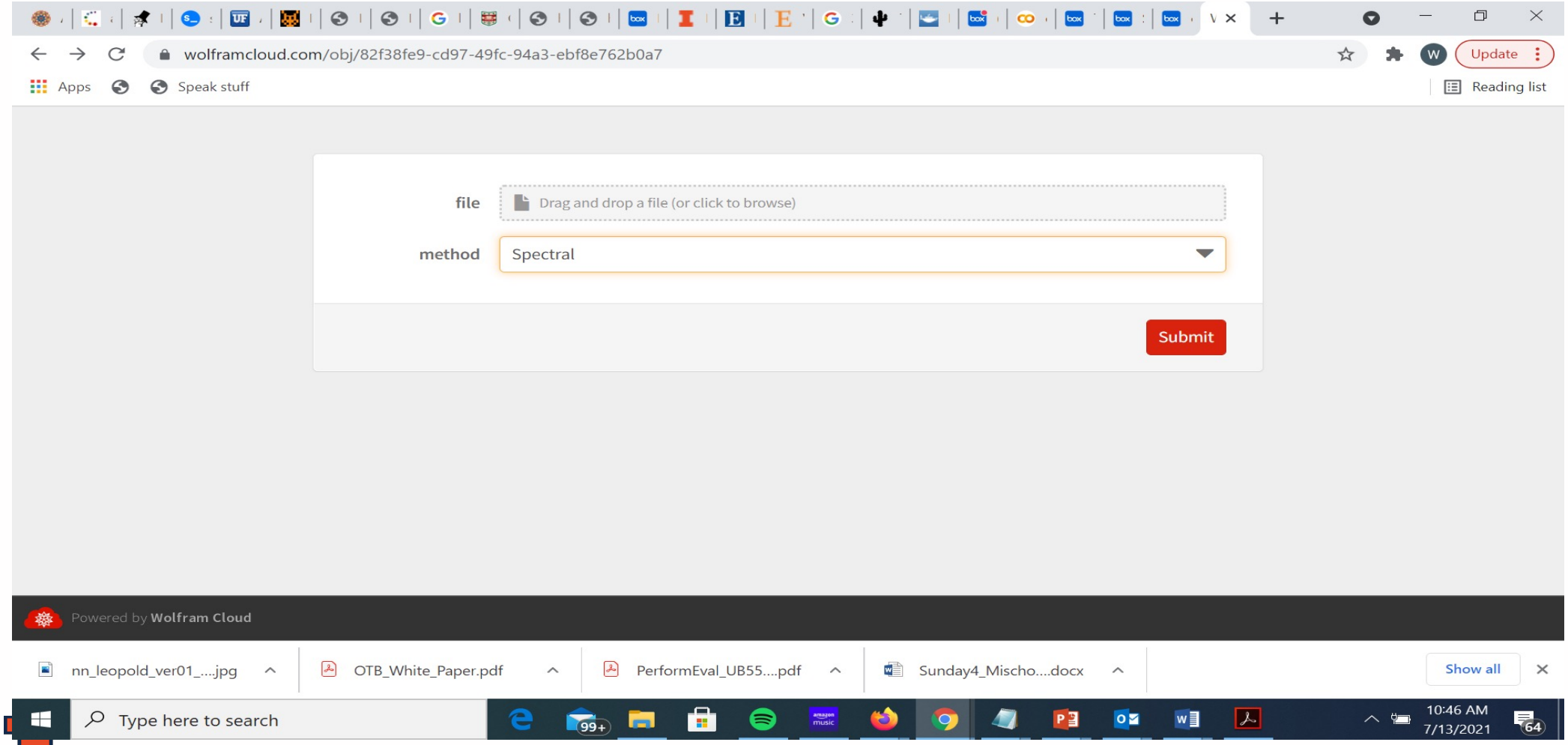
Scikit-learn K-Means algorithm on the phrases put everything in one cluster.

Scikit-learn Spectral Clustering test and that also resulted in only one cluster.

In the Scikit-learn calculations we used doc2vec which did not resolve into the bag of words but only resulted in one cluster.

Wolfram Mathematica Spectral Clustering. Broke nicely into four clusters of 142, 137, 72, and 81 documents using phrases.

.


wolframcloud.com/obj/82f38fe9-cd97-49fc-94a3-ebf8e762b0a7
Update
Apps
Speak stuff
Reading list
file
Drag and drop a file (or click to browse)
method
Spectral
Submit
Powered by Wolfram Cloud
nn_leopold_ver01_....jpg
OTB_White_Paper.pdf
PerformEval_UB55....pdf
Sunday4_Mischo....docx
Show all
Type here to search
10:46 AM
7/13/2021

wolframcloud.com/obj/82f38fe9-cd97-49fc-94a3-ebf8e762b0a7

WOLFRAM NOTEBOOK

Make Your Own Copy | Download | Share | Info | Sign In

1 2 3 4

```
{bioinformatics services available web years particular service Generalist on-line catalogs pool particular datatype textual search catalogs domain
  time structural bioinformatics resources Structural Bioinformatics Sema, publishers libraries continuing budget concerns economic downturn time
  uncertainty vendors planning strategic changes high-level of services norm Creative thinking views needed changes organizations electronic,
  social-type Web social discovery systems users tags features Purpose content of bibliographic records ratings reviews catalogue records primary
  underlying principles of cataloguing user mind user convenience prin, library instructors Online Computer Library Center WorldCat Local library
  catalog discovery layer OCLC tool teaching educational benefits WorldCat Local's faceted searching unexpected challenges classroom patron groups
  te, English-language refereed journals education psychology study open-access availability current status format publishers publications Gale
  Academic OneFile EBSCO Academic Search Complete ERIC ProQuest Central DOAJ, recent economic downturn budgets environment norm needs institutions
  librarians fewer players hands of fewer control scholarly content turn consolidation of publishers challenges dominant metrics content Us,
  generation library catalogue interfaces presumed advantages user experience user documentation patrons imbedded help assumption of ease best
  practices experience familiarity Web interfaces library Web sites inter, Ex Libris Primo Central discovery service North Carolina-Piedmont
  Automated Library System NC-PALS implementation consortium's selection purpose of highlighting considerations concerns case study challenges
  article planning, browser-based search tool unique implementation Stevens Institute of Technology creation article Internet browser JavaScript
  book mark applet users search of library resources oneSearch bookmarklet importance of convenience, dissatisfaction discovery tools range
  Discussions celebration internal library cataloging realities of shrinking library budgets students sustainability problems setbacks true
  information navigation skills metadata challenge, Davidson College different discovery systems course years selection implementation product
  evaluative processes implementation experiences needs analyses study projects potential bumps potential enhancements prod, Seminole State College
  of Florida Library discovery tool adoption case study CCLA Library Automation Primo beta test site College Center release transition library's
  research discovery tool instructional support stu, parallel libraries parallel code domain experts supercomputers complex analyses of huge
  datasets KDT high-level language Knowledge Discovery Toolbox difficulties graph expert small set of high-level graph operations exibl, complex
```

nn_leopold_ver01_....jpg | OTB_White_Paper.pdf | PerformEval_UB55....pdf | Sunday4_Mischo....docx | Show all

Type here to search

10:48 AM 7/13/2021

Under the Hood: Vectorising

Number of words total=72,171 Unique words=7337 (removal of stop words)
Phrases Total=17,879 Unique phrases=9711

Documents are similar if they are "close" At 430 documents, this is a combination of 430 objects taken two at a time so this is 92,235 initial comparisons to get the "closeness" values for any two documents.

Combination of n things taken r at a time: n! / r! (n-r)!

$$Cos\theta = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \, \|\vec{b}\|} = \frac{\sum_1^n a_i b_i}{\sqrt{\sum_1^n a_i^2} \sqrt{\sum_1^n b_i^2}}$$

where, $\vec{a} \cdot \vec{b} = \sum_1^n a_i b_i = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n$ is the dot product of the two vectors.

Vectorising Issues

Document corpus with *n* documents and *d* different terms D is an *n x d* document-term matrix. Representation in a relational database problem with limits on the number of columns.

Dimensionality Reduction (matrix factorization) techniques are employed: linear algebra

Clustering vs. Classification. Can Topic Modeling benefit from initial identification and classification of a few articles?

Important to reduce noise in the universe of terms *d*. Sparseness problem. This is a closeness measure, not a search. Elimination of terms, phrase extraction.

.

Examples

words_count_ID_final

words	Wcount	wnumber
abandonment	1	5
abate	1	6
abcd	1	7
abfe	1	8
abfes	1	9
abi3	1	10
ability	21	11
able	25	12
abor	2	13
about	90	14
above	4	15
absolute	2	16
absorption	1	17
abstract	5	18
abstraction	2	19
abstractions	2	20
abstract-only	1	21
abstracts	3	22
abubu.js	1	23
Abundant	3	24
Abusive	1	25
Academia	1	26
Academic	166	27
Academics	4	28

cresults

scopnums	vcosine
SCOPUS_ID:84856896596 !! SCOPUS_ID:84858263175	0.03892
SCOPUS_ID:84856896596 !! SCOPUS_ID:84858673471	0.13852
SCOPUS_ID:84856896596 !! SCOPUS_ID:84859512946	0.14044
SCOPUS_ID:84856896596 !! SCOPUS_ID:84859722371	0.18086
SCOPUS_ID:84856896596 !! SCOPUS_ID:84859814612	0.08443
SCOPUS_ID:84856896596 !! SCOPUS_ID:84859843292	0.0606
SCOPUS_ID:84856896596 !! SCOPUS_ID:84860147358	0.04147
SCOPUS_ID:84856896596 !! SCOPUS_ID:84861307071	0.08067
SCOPUS_ID:84856896596 !! SCOPUS_ID:84861309490	0.10591
SCOPUS_ID:84856896596 !! SCOPUS_ID:84861399622	0.16697
SCOPUS_ID:84856896596 !! SCOPUS_ID:84863097993	0.03024
SCOPUS_ID:84856896596 !! SCOPUS_ID:84864334008	0.12781
SCOPUS_ID:84856896596 !! SCOPUS_ID:84864365182	0.19531
SCOPUS_ID:84856896596 !! SCOPUS_ID:84864546394	0.10446
SCOPUS_ID:84856896596 !! SCOPUS_ID:84865228460	0.32615
SCOPUS_ID:84856896596 !! SCOPUS_ID:84865230643	0.14736
SCOPUS_ID:84856896596 !! SCOPUS_ID:84865281369	0.1646
SCOPUS_ID:84856896596 !! SCOPUS_ID:84866364871	0.1434
SCOPUS_ID:84856896596 !! SCOPUS_ID:84866367855	0.10258
SCOPUS_ID:84857523725 !! SCOPUS_ID:84857752538	0.1677
SCOPUS_ID:84857523725 !! SCOPUS_ID:84857775830	0.2152
SCOPUS_ID:84857523725 !! SCOPUS_ID:84858263175	0.15609
SCOPUS_ID:84857523725 !! SCOPUS_ID:84858673471	0.17429

phrases2

phrase	sentence	scopnum	vwhe
absolute binding free energies	5	SCOPUS_ID:85096814224	Abstr
abstract	2	SCOPUS_ID:85006489750	Abstr
abstraction	6	SCOPUS_ID:85084076448	Abstr
abstracts	4	SCOPUS_ID:85076353473	Abstr
abstracts	10	SCOPUS_ID:85080834723	Abstr
abundant motifs	5	SCOPUS_ID:84874803448	Abstr
abundant resources	2	SCOPUS_ID:85023166170	Abstr
Academia	1	SCOPUS_ID:84877587781	Abstr
academia	10	SCOPUS_ID:84929667151	Abstr
academic community	4	SCOPUS_ID:85026904125	Abstr
academic data aggregators	4	SCOPUS_ID:85025156368	Abstr
academic development trends	5	SCOPUS_ID:85027563760	Abstr
academic disciplines	5	SCOPUS_ID:84866377209	Abstr
academic e-book discovery	7	SCOPUS_ID:85021695823	Abstr
academic enterprise	3	SCOPUS_ID:85091374427	Abstr
academic enterprise	34	SCOPUS_ID:85091374427	Abstr
academic health science setting	2	SCOPUS_ID:84873712888	Abstr
academic health sciences library home pages	5	SCOPUS_ID:85044301294	Abstr
academic information	12	SCOPUS_ID:85025156368	Abstr
academic intuitions	2	SCOPUS_ID:84877587781	Abstr

UNIVERSITY

Lessons from IR Work

ML clustering algorithms are similar to text mining and search.

We've learned a lot from experience with IR systems using inverted file structures that record word/phrase document number, paragraph number, sentence number, and word position in the sentence.

Can we use or adapt what we know about proximity searching and field limiting, use of controlled vocabularies, and complex search arguments? Augmented Intelligence.

As Cliff mentioned in his plenary, we need to see a generation of tools that will open this up. Information professionals can play a role here.

More Questions than Answers

- Contact me with any questions w-mischo@Illinois.edu