Project Final Report - Team 149

Authors: Aaishwarya Kulkarni, Kunal Chheda, Kunal Khandelwal, Ojas Mehta, Priyanka Maheshwari, Raviteja Jayanti

Introduction & Project objective

Research output has increased exponentially over the past few decades [1]. We propose to map out entireties of research networks that exist within the NCBI Pubmed database [2], the most commonly used database for healthcare research. By categorizing research articles using advanced analytical techniques, we hope to make apparent global research statistics and identify articles of influence within subject topics. We also aim to help researchers look for related influential articles within a niche area. Beyond researchers and academic institutions, stakeholders include funding bodies would benefit from a broad understanding of the field.

Methodologies and Approach

Data Extraction and Pre-processing

The PubMed database consists of over 30 million articles [3] and has been used extensively for research projects globally due to its ease of access and relative high quality. In order to capitalise information pertaining to “MeSH” terms, we selected the Baseline [4] variant over the Open Access [5] variant of the PubMed databases available. The Baseline database contained articles published from the 1970s till 2019 and contained over 1,100 zipped XML files. Each file had data for 30,000 articles expanded to 200-300 Mb of disk storage after unzipping. We also had individual XML files in the Baseline database for each day in 2020 which we individual parsed and added into the overall database.

Due to an evolving PubMed dataset over years, we found that XML files had inconsistency in its available article attributes. After experimentation with open-source tools, medic [6, 7] and pubmed parser [8, 9] we found they were unfit for our purpose and we wrote our own parser to fetch XML files, process them (each ~200-300 Mb) and write CSV files. Our total dataset had 31 million records and was ~35 Gb. After filtering for english language and “journal” articles we were left with 24 million records, of which only 6 million had a complete list of references. We used the larger dataset for computing high level trends and the filtered dataset for article clustering and citation network development.

Data Hosting

We identified Amazon Web Services (AWS - Educate Account) for hosting and processing our data and deploying our web application. We took advantage of four key products from AWS as listed below:

- Amazon S3 to enable storage of our large csv files
- Amazon EMR with PySpark for big data transformations, performed offline
- Amazon EC2 for web server hosting
- Amazon RDS for relational MySQL database that can be queried

The dataflow is depicted in figure 1.

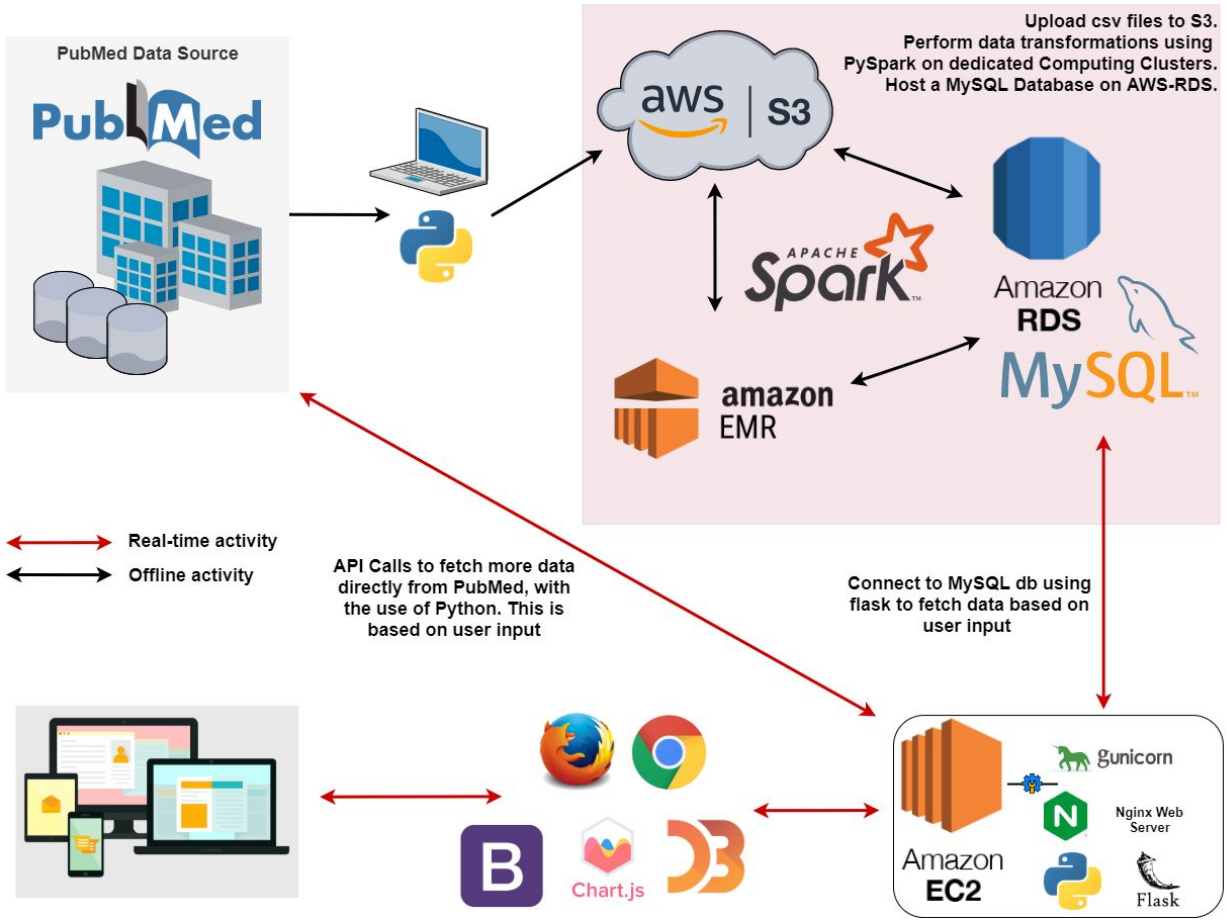


Figure 1 Data workflow. After acquiring relevant data from PubMed we used AWS products to perform big data transformations. We then hosted a web service to enable user interaction and live computations from a dedicated EC2 instance. Further minor data transformations and visualisations were performed in javascript on client browser

Analytical techniques

Data preparation for advanced analytics

In order to ensure that we could process user requests in a time-efficient manner, we shifted computationally heavy processes “offline” as seen in figure 2, which displays our analytics workflow.

Clustering

We used k-means clustering to group articles based on their MeSH terms and depict these clusters in the citation network with a distinctive color scale for the nodes. For our k-means clustering algorithm, we calculated a TF-IDF (term frequency – inverse document frequency) score [10-12] based on MeSH terms resulting in a sparse matrix with the help of the *pyspark ml* library [13]. Then on this derived sparse matrix we applied k-means to arrive at the final cluster.

Text-similarity analytics

We created a list of stopwords drawing from existing lists within the *nltk* library [14] and that provided by PubMed. We built a tokenization function along with nltk lemmatizer that would be better suited to medical terminology.

We determined the semantic similarity of articles with Cosine Similarity.[15-17] We tested different *Gensim* [18, 19] pre-trained glove models with different vector dimensions of word

embeddings. The model which best suited in terms of model size, model loading time and accuracy was *'glove-wiki-gigaword-200'* which is a 252 MB model with 200 vector dimensions. We used this model to create the doc2bow vector for each article which in turn was used to calculate cosine similarity [18].
For the heat map which represents the frequencies of co-cited articles, we used the TfidfVectorizer from sklearn [13] to provide on-click similarity computation for 2 co-cited articles.

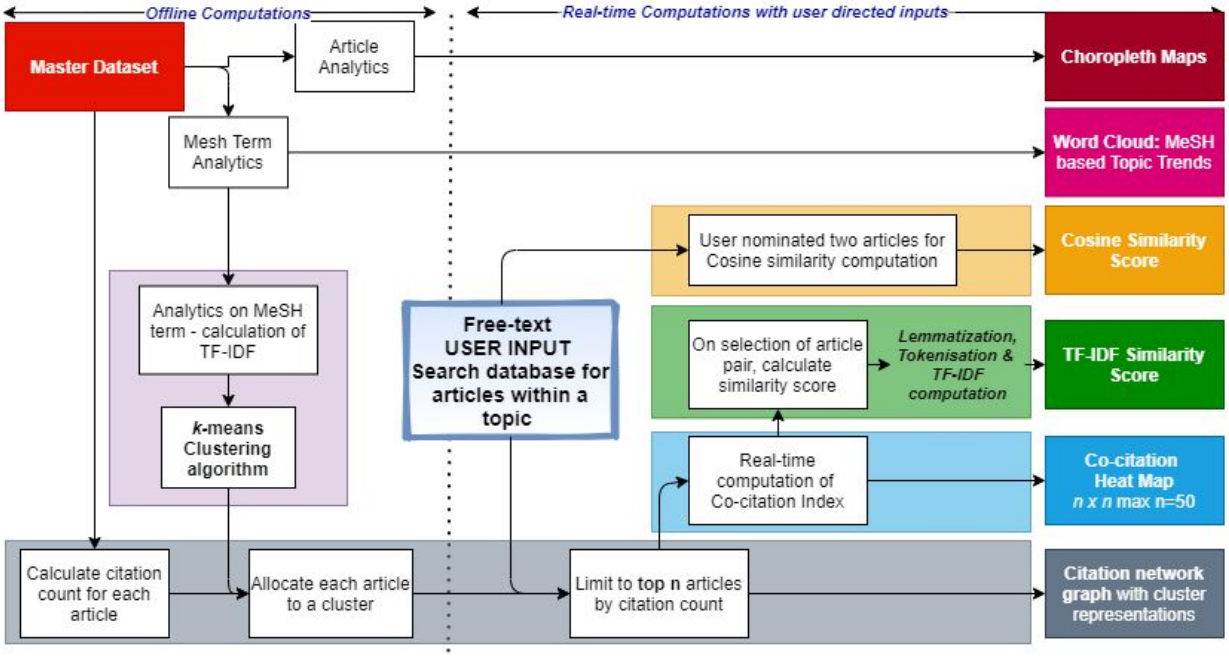


Figure 2 Analytics workflow. Our analytics is a combination between offline and real-time computations. Offline processes include the calculation of citation count for each article, clustering of all articles in the database, MeSH term and article analytics. Real-time processes stem from user inputs which serve to limit articles to a subject heading nominated by the user. This starts off a cascade of calculations which serve to create a citation network graph, co-citation heat map and text-similarity processes.

MeSH terms and Journal Analytics

We aggregated data for MeSH terms and Journal names based on the year and country wise frequency over 24 Million records. This data was used to visualize work by geography & time.These visualizations will help the user to study trends about research contribution based on article counts, MeSH search, Journal search and word cloud.

Real-time Analytics and User Interaction

We built the web application using the Flask framework for Python, which enabled us to quickly add multiple web pages and interactions. The application connects to the MySQL database to fetch real time data based on User Interaction, through RESTful APIs. The visualizations were built using tools like d3.js, chart.js and the API interactions using jQuery. The user interface styling was done using openly available CSS frameworks - bootstrap and bootswatch.

The application is run using a Gunicorn, which is an HTTP gateway configured to work with an NGinx web server. This setup is then deployed on an AWS EC2 instance, running on Ubuntu 18.04. We have used the free service provided by name.com for hosting our DNS and redirecting web requests to our server.

Visualisation

Visualisation plays a critical role in representing information, assisting the extraction of insights that may not otherwise be apparent [20, 21]. We have made use of bar charts, network graphs, word clouds, heat maps and choropleth maps.

We have created a citation graph network and co-citation heatmap to highlight relationships between articles limiting data points to prevent overcrowding [22]. We also have different charts to show journal and meSH term related analytics as seen in Figure 3. Further visualisation examples can be seen in the appendix.

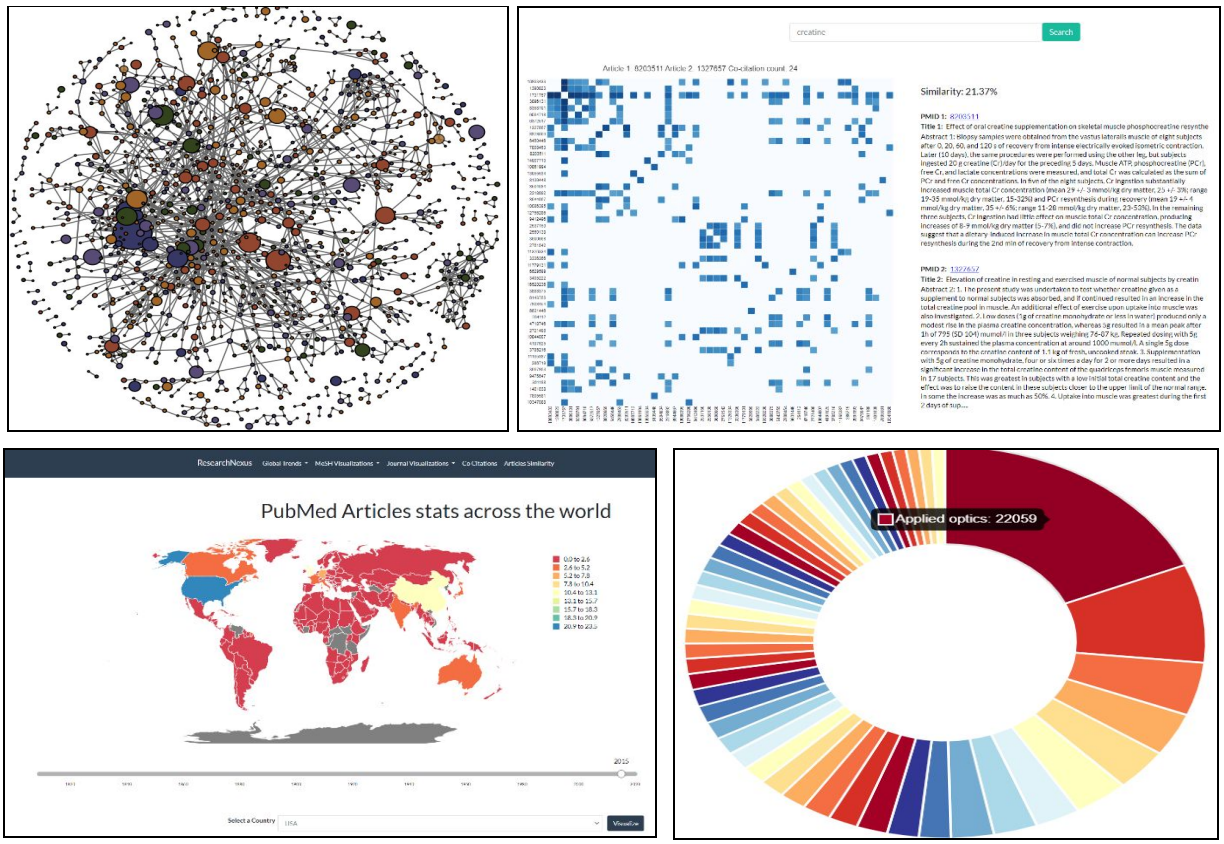


Figure 3 *Top Left) Network graph tracing relationship between top 2000 most cited articles within 'cardiovascular' field. Top Right) Heat Map (50X50) showing co-citation index and text similarity % based on abstracts between two articles within 'creatine' meSH term. Bottom Left) Choropleth map of research contribution by nations around the world in 2015. Bottom Right) Doughnut chart showing top 50 journals by the selected country (USA) in 2015*

Experiments and Evaluation

We conducted experiments and where relevant, we took the help of a medical expert to review our results at various stages.

1. ***Comparing efficiency of available vs. customised stopwords list and tokenization algorithms:*** we observed that open-source stopwords list and tokenization algorithms from the *nltk* library were insufficient due to sub-optimal similarity results. We modified the list with pubmed stopwords and created a new tokenization algorithm, which jointly led to a better set of similarity scores when reviewed by a medical expert. For example, we built a custom

tokenizer to handle words such as *BD+O(2), p-value, NF-kappaB, ALADIN(I482S)* which would otherwise lose their meaning when splitting on all joining characters (,), +,.

2. ***Comparison of citation network development techniques on our dataset:*** We compared three broad techniques of developing a citation network: direct citation [22], co-citation [23] and bibliographic coupling [24]. These can be built with a directed or undirected graph [22, 25]. We determined that the direct citation network was the most intuitive method of display in a network graph. However, we also determined that the co-citation network also provided information regarding the strength of the relationship between two articles. We felt that the best display of this would be via a heat map, where articles that have a greater co-citation score would have a darker colour. The method of bibliographic coupling did not provide any further insight and so we discontinued its pursuit.
3. ***Comparison of k-means clustering to Louvain algorithm:*** When testing the Louvain algorithm [26, 27], we quickly found there existed a large number of isolated communities containing articles that had few edges. We decided not to go ahead with the Louvain algorithm for three reasons: computational time was extensive (A search result of 100,000 articles took up to 60minutes); due to isolated nodes, a large number of communities were formed; the algorithm was built on undirected edges and did not distinguish between an article that cited many other articles, compared to an article that had been cited by many other articles (which is of greater interest). For visual representation, we also had to limit our search to avoid a cluttered network graph. This would substantially affect the formation of communities with the Louvain algorithm. With our k-means clustering methodology based on MeSH terms, we were able to compute clusters for every article prior to user-based search, speeding up the process dramatically.
4. ***Evaluation of text similarity algorithms when applied to abstracts:*** Medical experts were able to confirm that the articles that were retrieved based on highest text similarity scores (TF-IDF) were in fact correlated. Whilst this was a subjective assessment, we were able to repeat this process for a number of articles to gain confidence in our methodologies. We found that the use of the *gensim* library performed slightly better than TF-IDF, at the cost of significantly more expensive computational power. We therefore decided to use TF-IDF when retrieving a large number of articles and the *gensim* library for focused pairwise article comparison.
5. ***Sentence embedding model – BioSentVec:*** BioSentVec is a pre-trained model on PubMed+MIMIC-III data used for calculating similarity between articles. As the size of the model is around 21GB, it was computationally difficult to load the model and compute similarity against the whole dataset.

Team Member contribution

Whilst overall all team members contributed to the project evenly, each member, often in groups, took a lead in certain areas of the project. Details seen below.

Data processing	**NLP & Clustering**	**Visualisation**	**Web interface**	**Writeup**
RT/KC	KK/AK	PM/OM	All	All

Discoveries & Insights

Throughout the process, we made abundant discoveries relating to data quality, efficient methods for data computation, visualisation and types of citation networks and various NLP techniques and their limitations. Some key highlights include the following:

- Even a highly regarded database such as PubMed, has missing information relevant for our project. This is amplified as the database evolves over time with more fields added

- Code optimisation was critical to our project in order to process user requests and real-time computations over a large dataset in a reasonable timeframe
- Paired articles with a higher co-citation count also tend to have greater similarity scores
- TF-IDF was significantly quicker than vectorization based NLP algorithms such as cosine similarity. This allowed rapid text-similarity computations over a large number of article pairs.

Discussion of challenges and next steps

Two significant challenges faced in the project, within the timeline provided, was hosting a large, shared database and interaction methodologies that would result in the least financial impact and performing real-time analytics on the large volume of data in the most computationally efficient way.

The challenge we had during the data processing phase was a lot of AWS services were unavailable for the Educate accounts. For example, the entire data transformation stage could have been configured as a data pipeline service, but AWS Educate accounts were not authorised to access these services. So we had to resort to manually loading the processed data files into AWS S3 buckets before processing them into the Database, which was highly time intensive and consumed significant internet bandwidth.

Due to the sheer volume of data, a significant challenge in this project was to strike the right balance between real-time computation and offline-processing. Our aim, as we initially set out, was to maximise user-driven analysis and exploration of the vast pubmed database to maximise the value of the project to the user. Challenges we met related to the computational demands of advanced analytics. In particular, on-demand cosine-similarity computation for a matrix of over 10 by 10 articles was in the order of several minutes, which we deemed too long a wait time for any user. However, with millions of articles, even offline computation of an *n by n* matrix of cosine similarity scores of the entire article database would be impractical and would require several high powered computing clusters. We tackled this challenge by using an approximate but rapid computation of text-similarity with the TF-IDF methodology as well as offering a separate two-article cosine similarity comparison for finer detail.

Next steps in our project include…

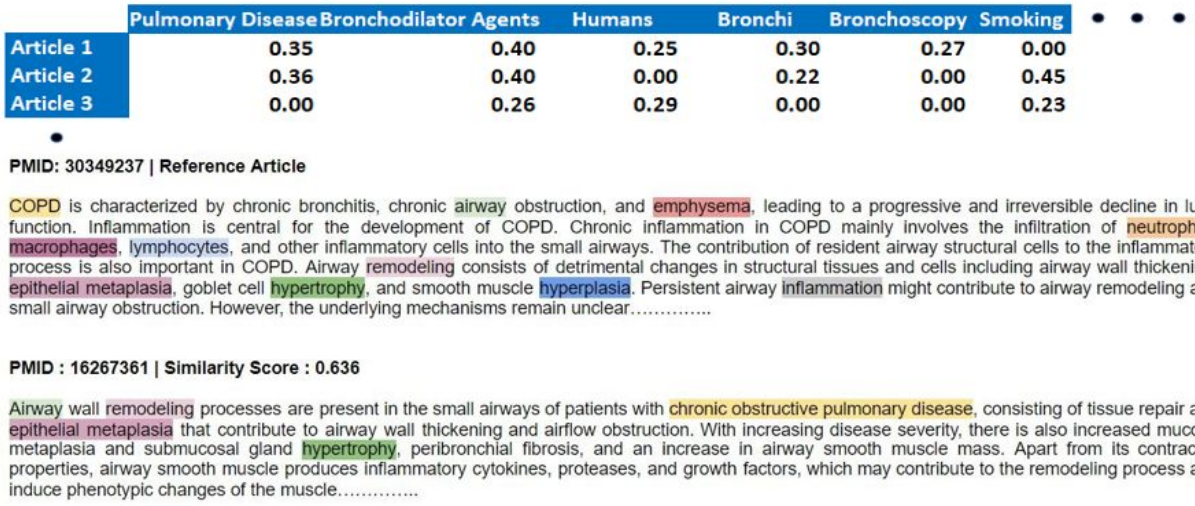
- Computation of impact factor for each journal
- Creation of a unique list of authors and analysing co-authorship networks. Duplicate author names are commonly seen but need to be separated by estimates of likelihood based on other parameters such as relevant department, country, subject heading and co-authors.
- Finer detail in our geographical maps & creation of network maps to demonstrate strength of relationships between different institutions, assessed by number of co-authored papers
- Natural Language Processing on full articles (this is limited by the availability of only PMC articles)

Overall, we have created a tool for research exploration within the largest body of medical articles in the world, the PubMed database. We hope that this tool can better enable students, academics and other stakeholders to get a better sense of trending research topics, recognise key journals, explore the geographical variation in research output and identify articles of overwhelming influence in particular niche areas.

Appendix

Column Name	Description	DataType
PM Id	PubMed ID	String
Title	Title of the article	String
Abstract	Abstract of the article	String
Has Abstract *(added)*	Introduced to reduce bias against older articles	Boolean
MeSH terms	List of Medical Subject Headings	String
Has MeSH terms *(added)*	Introduced to reduce bias against older articles	Boolean
Authors	Names of the authors	String
Author Affiliations	Affiliations of the authors	String
Publishing Date	Year/Date when the article was published	Date
Journal	Journal of publishing	String
Journal Abbreviation	Journal Abbreviation	String
Journal Country	Country where the journal was published	String
Grant Agencies	Agencies that sponsored grants	String
Grant Country	Country where agencies sponsored grants	String
Reference List	List of PM IDs of articles that have referenced index article	String
DOI	Digital Object Identifier information	String

Figure A: Data Fields extracted from Pubmed Baseline XML files using custom created parser



	Pulmonary Disease	Bronchodilator Agents	Humans	Bronchi	Bronchoscopy	Smoking
Article 1	0.35	0.40	0.25	0.30	0.27	0.00
Article 2	0.36	0.40	0.00	0.22	0.00	0.45
Article 3	0.00	0.26	0.29	0.00	0.00	0.23

Figure B: Above image demonstrates the TF-IDF score matrix and example of similarity scores derived on article abstracts with related MeSH terms

Data Represented	Meaning	Tool
Co-citation index	Represents the frequency with which a pair of articles have been cited by a common third article. This is a description of the closeness of these two articles	Heat Map
Direct citation	Network graph with nodes representing articles and edges representing a citation between two articles	Network Graph
Text-similarity	TF-IDF similarity performed on articles with a high number of co-citations (seen in heat map). Cosine similarity allows a user-nominated article (each article represented by a PMID) for calculation of a similarity score	Similarity Scores
Research output by Geography & Time	Displays which countries have contributed the most articles over the past centuries	Choropleth Article Maps
Research contribution based on Journals	Displays the most productive journals in terms of research output over time	Choropleth Journal Maps
MeSH Analytics	MeSH term trends for a country in a year	MeSH country
	MeSH search and its trends over the years	MeSH search
	MeSH cloud and frequencies for a year	Word Cloud
Journal Analytics	Top research journals for a country in a year	Journal Country
	Journal search and research work over the years	Journal search
	Top journals based on total publications for a year	Journal trends

Figure C: Different Visualisation tools used to represent data

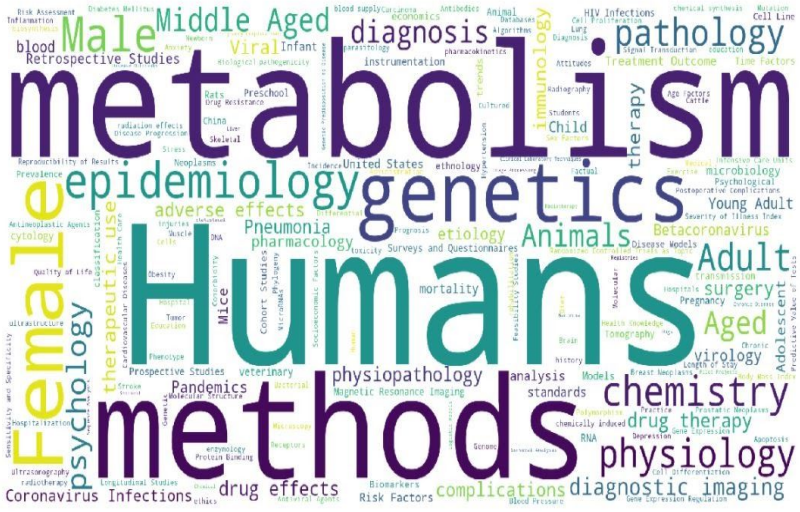


Figure D: Word Cloud of meSH terms used in 2020

References

[1] P. Fontelo and F. Liu, "A review of recent publication trends from top publishing countries," *Systematic Reviews,* vol. 7, 12/01 2018, doi: 10.1186/s13643-018-0819-1.

[2] W. S. Canese K, *The NCBI Handbook.* , Second ed.: National Center for Biotechnology Information (US), 2002. [Online]. Available: https://www.ncbi.nlm.nih.gov/sites/books/NBK153385/.

[3] N. L. o. M. NCBI. "PubMed Overview." NCBI. https://pubmed.ncbi.nlm.nih.gov/about/ (accessed 31 October 2020).

[4] "MEDLINE/Pubmed Data." https://www.nlm.nih.gov/databases/download/pubmed_medline.html (accessed 31 October 2020).

[5] "Open Access Subset." NCBI. https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/ (accessed 31 October 2020).

[6] "PyPI - Medic." https://pypi.org/project/medic/ (accessed 31 October 2020).

[7] F. Leitner. "Medic: A Command-Line Tool To Manage A Mirror Of MEDLINE." https://zenodo.org/record/9968#.X50cS4gzbIU (accessed 31 October 2020).

[8] T. Achakulvisut, D. Acuna, and K. Kording. "Pubmed Parser: A Python Parser For Pubmed Open-Access XML Subset And MEDLINE XML Dataset. ." https://zenodo.org/record/3648715#.X50ehIgzbIU (accessed 31 October 2020).

[9] Titipata. "Pubmed Parser: A Python Parser For Pubmed Open-Access XML Subset And MEDLINE XML Dataset — Pubmed Parser 0.2.2 Documentation. ." https://titipata.github.io/pubmed_parser/ (accessed 31 October 2020).

[10] J. Ma, W. Xu, Y. Sun, E. Turban, S. Wang, and O. Liu, "An Ontology-Based Text-Mining Method to Cluster Proposals for Research Project Selection," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans,* vol. 42, no. 3, pp. 784-790, 2012, doi: 10.1109/TSMCA.2011.2172205.

[11] K. Nguyen, S. Byung-Joo, and Y. Seong Joon, "Hot topic detection and technology trend tracking for patents utilizing term frequency and proportional document frequency and semantic information," in *2016 International Conference on Big Data and Smart Computing (BigComp)*, 18-20 Jan. 2016 2016, pp. 223-230, doi: 10.1109/BIGCOMP.2016.7425917.

[12] S. Qaiser and R. Ali, "Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents," *International Journal of Computer Applications,* vol. 181, 07/16 2018, doi: 10.5120/ijca2018917395.

[13] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research,* vol. 12, no. 85, pp. 2825-2830, 2011.

[14] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python.* . O'Reilly Media, Inc, 2009.

[15] W. Gomaa and A. Fahmy, "A Survey of Text Similarity Approaches," *international journal of Computer Applications,* vol. 68, 04/18 2013, doi: 10.5120/11638-7118.

[16] A. Lahitani, A. Permanasari, and N. Setiawa, "Cosine similarity to determine similarity measure: Study case in online essay assessment," presented at the International Conference on Cyber and IT Service Management (CITSM), Bandung, Indonesia, April 2016, 2016, Institute of Electrical and Electronics Engineers

[17] M. B. Magara, S. O. Ojo, and T. Zuva, "A comparative analysis of text similarity measures and algorithms in research paper recommender systems," in *2018 Conference on Information Communications Technology and Society (ICTAS)*, 8-9 March 2018 2018, pp. 1-5, doi: 10.1109/ICTAS.2018.8368766.

[18] G. Sidorov, A. Gelbukh, H. Gomez Adorno, and D. Pinto, "Soft Similarity and Soft Cosine Measure: Similarity of Features in Vector Space Model," *Computación y Sistemas,* vol. 18, 09/30 2014, doi: 10.13053/cys-18-3-2043.
[19] P. Stefanovič, O. Kurasova, and R. Štrimaitis, "The N-Grams Based Text Similarity Detection Approach Using Self-Organizing Maps and Similarity Measures," *Applied Sciences,* vol. 9, p. 1870, 05/07 2019, doi: 10.3390/app9091870.
[20] D. Henri, V. Léveillé, S. Manullang, Jm, and D. Jr, "Patent analysis for competitive technical intelligence and innovative thinking," *Data Science Journal,* vol. 4, 01/01 2005, doi: 10.2481/dsj.4.209.
[21] M. Khan and S. Khan, "Data and Information Visualization Methods, and Interactive Mechanisms: A Survey," *International Journal of Computer Applications,* vol. 34, pp. 1-14, 12/01 2011.
[22] B. M. Henrique, V. A. Sobreiro, and H. Kimura, "Building direct citation networks," *Scientometrics,* vol. 115, no. 2, pp. 817-832, 2018/05/01 2018, doi: 10.1007/s11192-018-2676-z.
[23] H. Small, "Co-citation in the scientific literature: A new measure of the relationship between two documents," *Journal of the American Society for Information Science,* vol. 24, no. 4, pp. 265-269, 1973, doi: 10.1002/asi.4630240406.
[24] M. M. Kessler, "Bibliographic coupling between scientific papers," *American Documentation,* vol. 14, no. 1, pp. 10-25, 1963, doi: 10.1002/asi.5090140103.
[25] L. Egghe and R. Rousseau, "Introduction to Informetrics. Quantitative Methods in Library, Documentation and Information Science," 01/01 1990.
[26] L. Šubelj, N. J. van Eck, and L. Waltman, "Clustering Scientific Publications Based on Citation Relations: A Systematic Comparison of Different Methods," *PLOS ONE,* vol. 11, no. 4, p. e0154404, 2016, doi: 10.1371/journal.pone.0154404.
[27] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment,* vol. 2008, no. 10, p. P10008, 2008/10/09 2008, doi: 10.1088/1742-5468/2008/10/p10008.