Research Nexus

Motivation & Objective

Whilst globalisation has enabled an enriched collection of scientific literature, growing complexity in research networks makes it challenging for stakeholders to understand intricacies of these relationships. Currently, users search for medical articles on databases via simple text word searches on platforms such as Pubmed and Google Scholar. These methods may rely heavily on features reported by journals that are not standardised due to cultural differences, individual journal reporting protocols, and type of article. Our tool enables 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 medical research articles of overwhelming influence in particular niche areas.

Research Nexus - A Novel Approach

Research Nexus is a unified platform underpinned by advanced analytics and visualization techniques which attempts to demystify the growing complexity in research networks.

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.

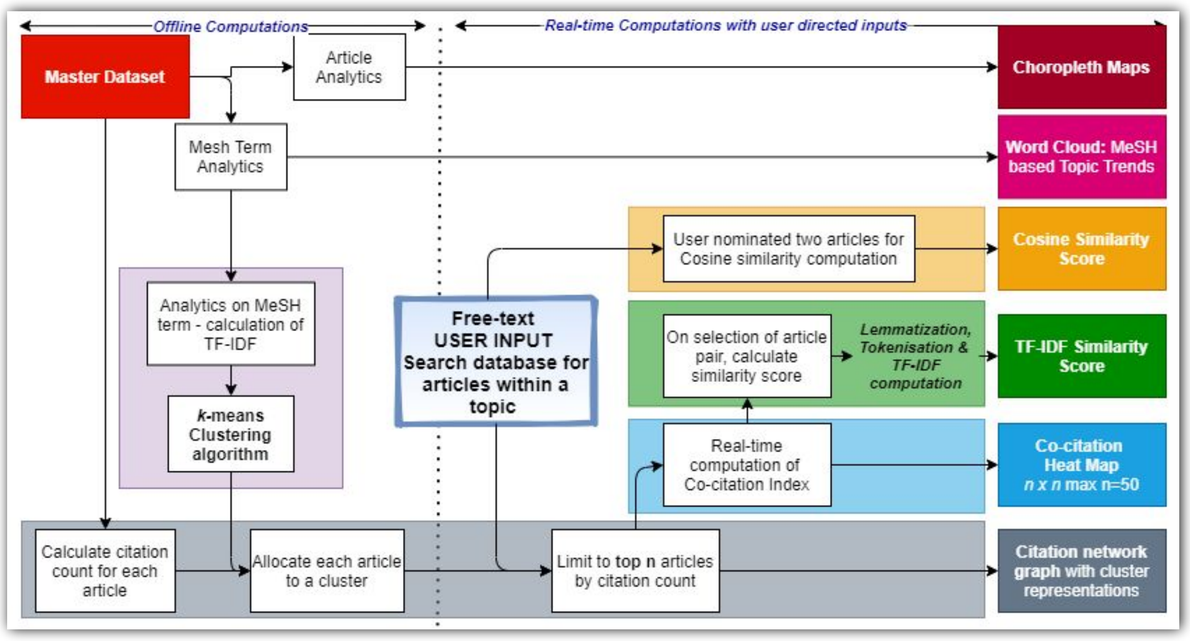

Figure 1: Analytical workflow

Data Workflow

The PubMed database consists of over 30 million articles and offers two database variants, the *Open Access* database and the *Baseline* database. However, Research Nexus is built using Baseline database as it contains critical information pertaining to "MeSH" terms (*Medical Subject Headings*). The Baseline database contains articles published from 1970s till 2020 distributed across 1000+ zipped XML files. Due to ever evolving PubMed data and unavailability of sophisticated parser, we built our own parser to perform following tasks:

1) Fetch Baseline XML files from PubMed ftp server and save on local machine amounting to 35 million records and occupying ~35GB disk space.

2) Processes each file (size ~300MB) to extract relevant fields.

3) Filter out articles where source language is not English plus type is other than *journal.* Final set obtained amounted to 24 million records. This data set was used for MeSH and Journal analysis.

4) Filter out articles where references are not present. This set of ~6 million records was used for citation and co-citation network.

The workflow of data from the main source to the consumption layer is as given in Fig 2.

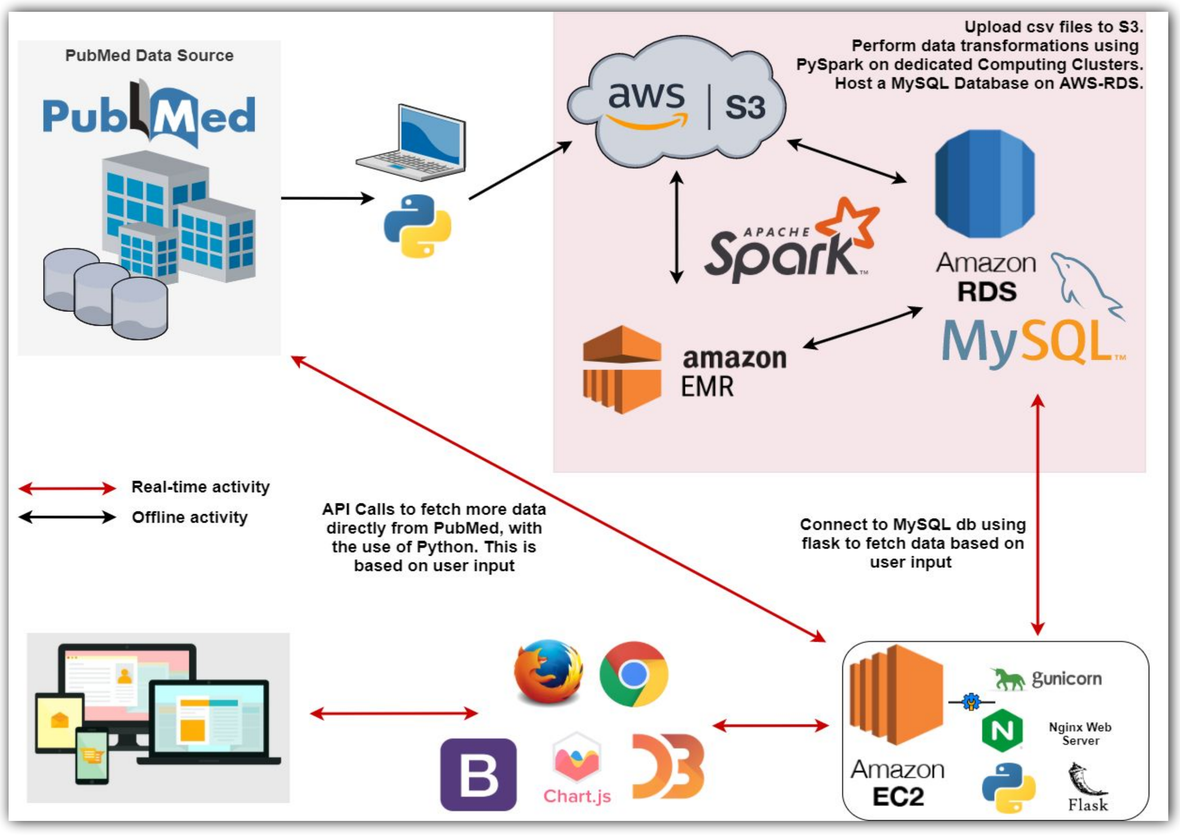

Figure 2: Data workflow. After acquiring relevant data from PubMed we used AWS products to perform big data transformations. We then hosted a web service to enable user interaction and live computations from a dedicated EC2 instance. Further minor data transformations and visualisations were performed in javascript on client browser.

Experiments & Evaluation

Comparing efficiency of Nltk vs. custom stopwords list and tokenization algorithm: Nltk library was insufficient due to sub-optimal similarity results. We modified the nltk stopwords list by adding pubmed stopwords along with others and created a new tokenization algorithm, which jointly led to a better set of similarity scores when reviewed by a medical expert.

Comparison of citation network development techniques on our dataset: We compared three broad techniques of developing a citation network: direct citation, co-citation and bibliographic coupling. We determined that the direct citation network was the most intuitive method of display in a network graph. 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.

Comparison of k-means clustering to Louvain algorithm: When testing the Louvain algorithm, we found that the computational time was extensive. 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.

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. 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.

Visualization

Current research databases have limited visualisation. Pubmed, one of the largest databases for medical articles, has only rudimentary methods of displaying information and lacks interactivity. Research Nexus as a unified platform presents interactive visualizations in the form of network graphs, choropleth maps, word clouds, bar & line charts which unfolds various trends and relationships among articles.

Citation Network and co-citation heatmap

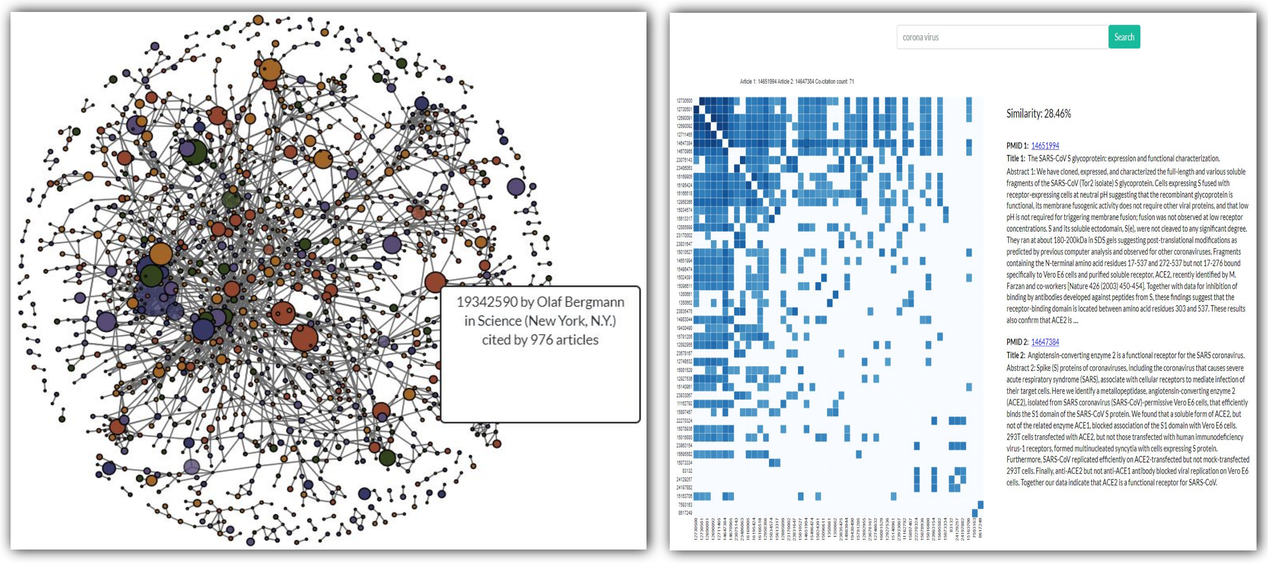

Figure 3: Left) Network graph tracing relationship between top 2000 most cited article within 'cardiovascular' field. Right) Heat Map showing co-citation index within 'coronavirus' meSH term.

Journal and MeSH related analytics

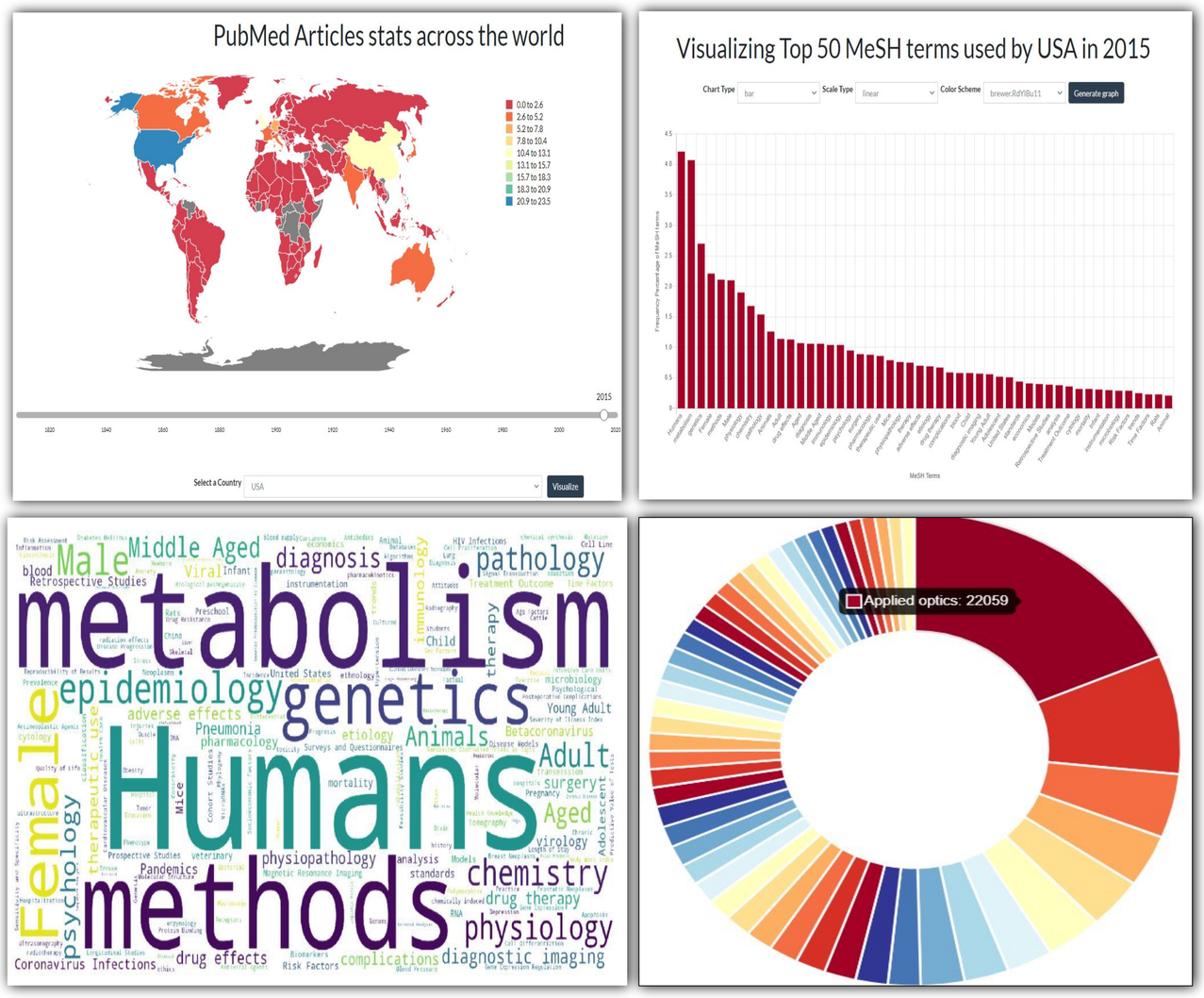

Figure 4: Top left) Choropleth map of research contribution by nations around the world in 2015. Top Right) Bar chart of top 50 meSH terms selected country (USA) in 2015 Bottom Left) Word cloud of MeSH terms used in 2020 Bottom Right) Doughnut chart showing top 50 journals by the selected country (USA) in 2015.

RESEARCH POSTER PRESENTATION DESIGN © 2019
www.PosterPresentations.com