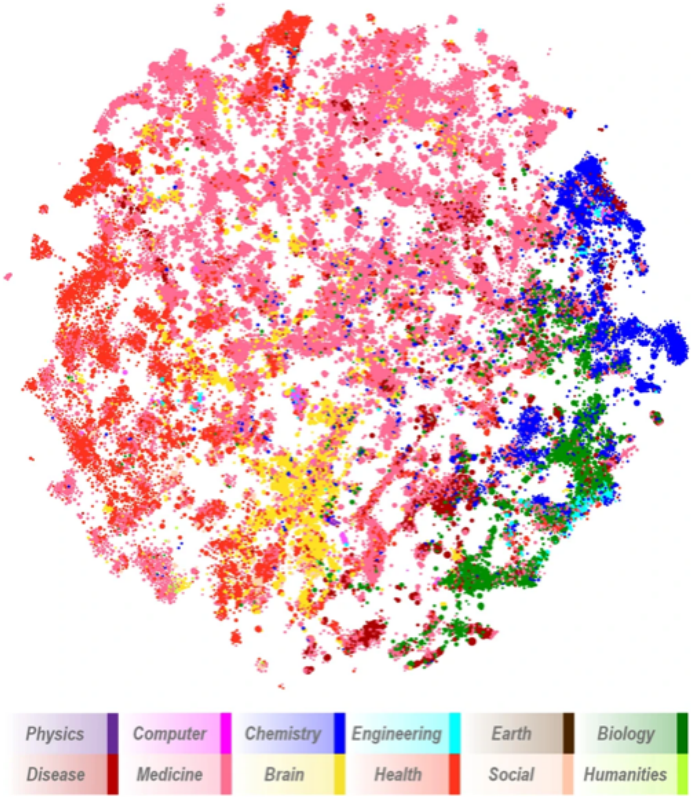
Clustering Medline/ Pubmed Baseline Data

F2021: CS543 Group 4 Project Proposal, Supervised by Prof. J. Abello



Visual map of the PubMed model showing 28,743 clusters. Each cluster is colored according to its dominant field (see legend).

Introduction and Problem

PubMed comprises citations for biomedical literature from MEDLINE, life science journals, and online books. The citation information includes title, authors, journal, publication date, abstracts of articles and also links to full text content from PubMed Central and publisher web sites.

With such vast amounts of data, very interesting insights could be gained with various techniques like clustering, analysis and visualization.

Objective

This project aims to cluster the citations from the Medline/ PubMed baseline based on the authors, journals, country of origin, and/or the keywords/ topics associated with the citations. This will also allow us to visualize and analyze meaningful patterns in the data.

This data is huge in terms of size (approx. 250 GB+) and has around 40-45 Million data points (32 million+ & 10M+ Update points) as given at https://ftp.ncbi.nlm.nih.gov/pubmed/baseline/ and https://ftp.ncbi.nlm.nih.gov/pubmed/updatefiles/ respectively. So, definitely this requires multi-cluster processing as it would not fit in RAM of a personal computer.

Data Collecting and Processing, Tasks to Complete

Since the data is huge as mentioned above, we plan to explore options from using Amazon cloud resources like AWS EMR or AWS Glue, Google cloud resources like Google Cloud Dataproc, which provides multi-cluster nodes for processing data using Big Data technologies like Apache Spark/ Hadoop or Cluster on multiple local machines.

Apart from the infrastructure, we would need to clean, transform and then process the data, as the data is across several xml files (1500+) and there is important information as well as irrelevant tags. Once we do data collection and processing, we need to identify specific keywords or categories, use a relevant clustering algorithm, and query, display & visualize data.

The data appears to be updated every day, so we will also need to set up a stream that identifies a new file, processes it and updates the existing cluster.

Potential Challenges and Timeline

There are some potential challenges we might face like the costs of computing if we plan to use cloud resources, even though there are free tier available in some of them. If we are planning to set up clusters on personal machines, we might need at least 10-12 machines, if we plan to use in-memory based processing (like in spark) as the data is too huge to be processed. So we will probably need to explore ways to reduce this data or do some testing on small data and ramp up for a large or complete dataset so that we end up using less resources than expected.

After sorting out the above challenge, we will need to identify a good clustering algorithm for our data (most likely CURE algorithm considering a large dataset) and then finally the visualization of our result data.

References

https://spark.apache.org/docs/latest/ml-clustering.html
https://www.nature.com/articles/s41597-020-00749-y
https://www.leydesdorff.net/pubmed/
https://www.nature.com/articles/nbt.4267
https://www.nlm.nih.gov/bsd/licensee/2021_stats/2021_LO.html
https://www.nlm.nih.gov/databases/download/pubmed_medline.html
https://aws.amazon.com/glue/
https://spark.apache.org/docs/latest/cluster-overview.html
https://cloud.google.com/dataproc
https://aws.amazon.com/emr/