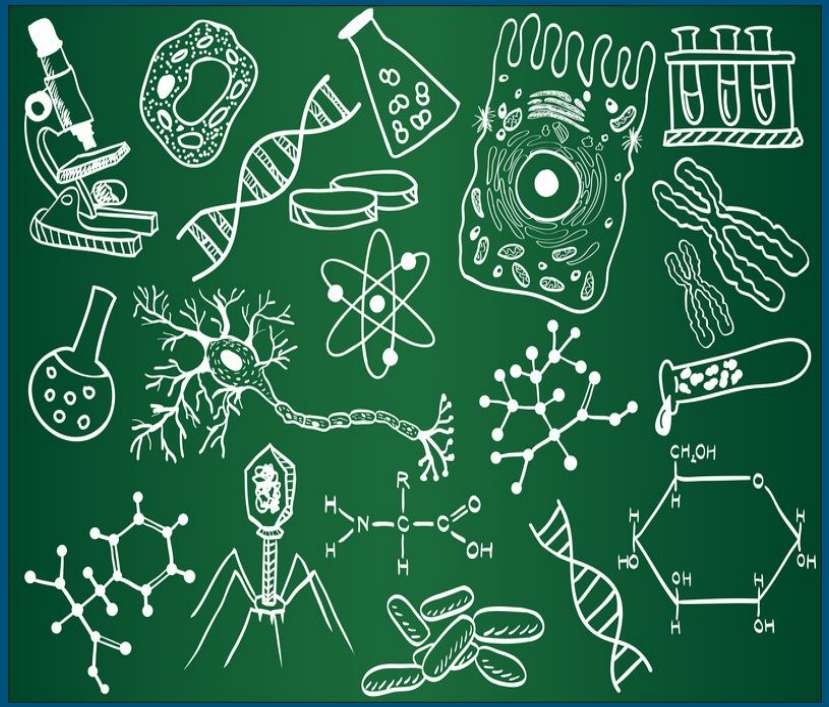
Classification of Biological Data - Common Crawl

Big Data Analytics

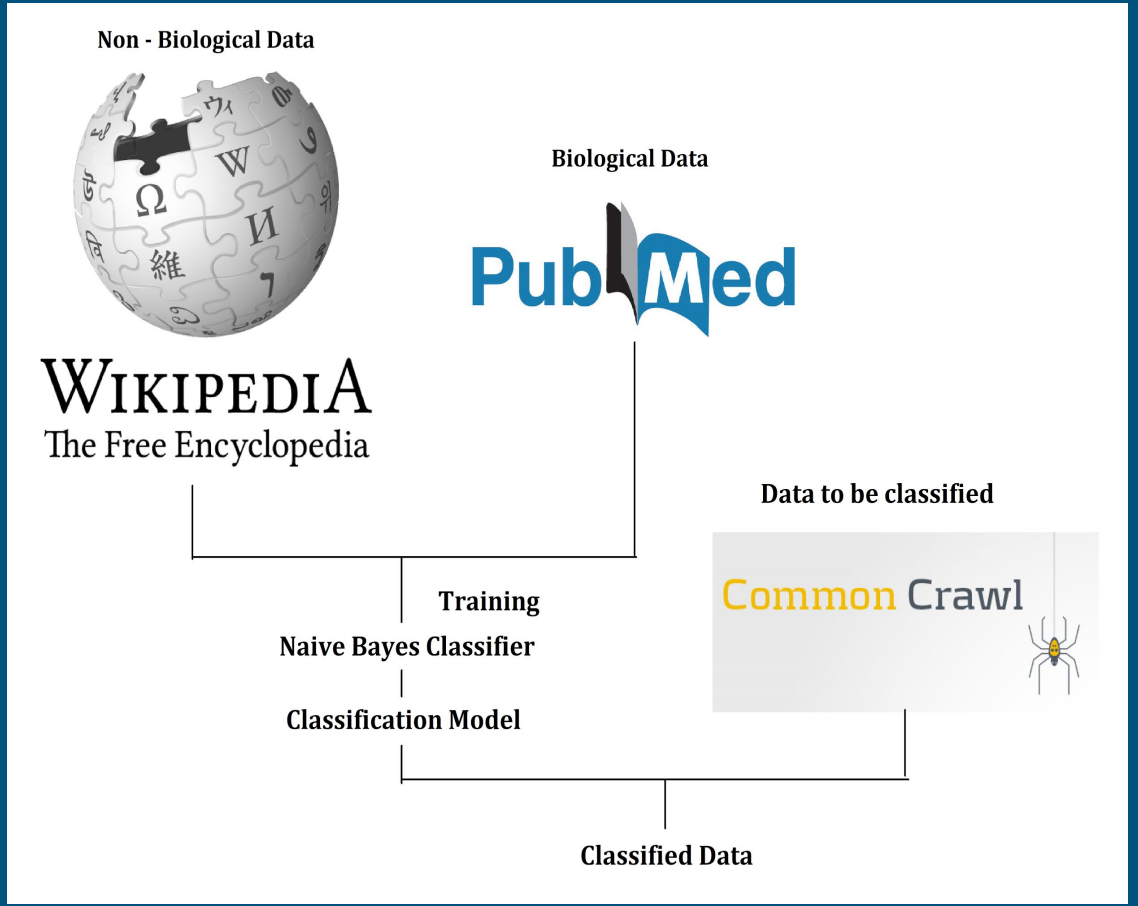
Problem we are looking to solve

- The Common Crawl is an **open repository of web crawl data** collected over last 7 years which contains several petabytes of data.
- This project aims to separate **Biological data** from other types which are not related to biology. It further aims at **classifying the Biological data into different classes** such as: human microbiome, carcinoma, infectious diseases, genetic diseases, etc.
- The topics in biology are quite diverse and thus the documents may fall into one or more classes. Thus the classification is not disjoint.



Architecture

- Train the data with Non - Biological data from Wikipedia and Biological Data from PubMed using Naive Bayes.
- Process archived common crawl files in parallel: unpack, process, delete unpacked version.
- Classify Common Crawl data based on the classification model.
- Use different categories in PubMed to classify data into different biological categories using parallel pipelines.

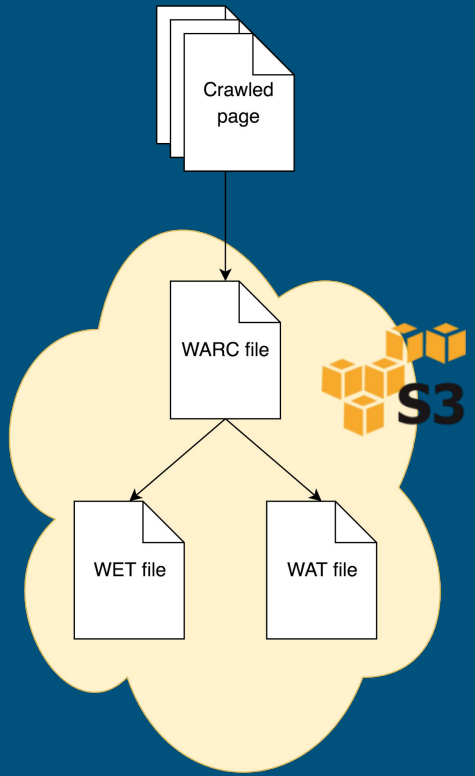


Technologies

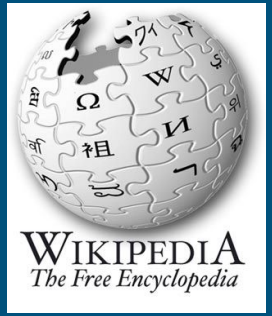
- Common Crawl data: stored on **Amazon S3**. We use **Globus** for data transfers.
- ML pipeline: **PySpark**.
- Storage: **MongoDB** (We wish to use this but we haven't. For now we have stored them in folders on AWS s3.)

DataSets: Common Crawl

- Stored as gzipped archive files, with many pages in each file
- 3 main file formats: **WARC**, **WAT** and **WET**.
- WET contains the extracted page text, could be useful for classification.
- Extraction algorithm is available online, but probably hard to implement in a Spark pipeline.
- Other possibility: strip HTML ourselves. (Get rid of XML like structure)



DataSets: Wikipedia for Non-Biology



- We have used the English documents from the Wikipedia Dataset used by the LORELEI classification project for training of Non - biology documents.
  - *Enwiki-latest-pages-articles-multistream.xml*
  - *extracted by* WikiExtractor.py
- There are 5297852 articles in this dataset.
- We have not separated the biology articles from the non - biology ones. (The separation must be done in order to achieve better results. This is a part of the future work as we don't currently have pure non-biology data set to train the classifier .)

DataSets: PubMed for Biology

- We have used the PubMed articles to train our classification model for Biology related documents.
- These PubMed articles can be further categorized into classes specific to different topics in biology such as: Microbiome, Cancer, Infectious diseases, Genetic Diseases, Gene Therapy, etc.
- The articles were extracted from ftp://ftp.ncbi.nlm.nih.gov/pub/pmc/oa_bulk/

Pipeline

- Data retrieval and preprocessing using custom Bash+Python scripts
- Actual classification implemented in pySpark

Preprocessed Articles → Punctuation Stripper → Tokenizer → N-Gram Features → CountVectorizer → NaiveBayes

- Part A : Classification of Non-biological data from Biological data
  - N - gram model used : Unigram
- Part B: Classification of Biological Data into other Bio related categories (Microbial, Cancer and others)
  - N - gram model used : Trigram and Four-gram
- The microbiome subset of data was too small for training, hence we have used microbe data from PubMed.

Result - Small Test Data Set (Confusion Matrix)

	bio	other	total
bio	86193	2100	88293
other	2712	65833	68545

	microbial	cancer	other	total
microbial	69	3	47	119
cancer	0	63	64	127
other	6	4	282	292

Result - Full Data Set (Confusion Matrix)

	bio	other	total
bio	9402063	650493	7659957
other	746637	6913320	10052556

	microbial	cancer	other	total
microbial	222007	16391	282559	520957
cancer	17743	256881	282782	557406
other	47922	42102	504066	594090

Result - Performance (Bio and Non-Bio)

	Small Data Set	Full Data Set
Test Error	0.0306813399814	0.0788781354741
Runtime	Less than 10 minutes on: master(m4.large), 2 core(m4.xlarge)	Less than 20 minutes on: master(r3.2xlarge), 10 core(r3.2xlarge)

Result - Performance (Microbes - Cancer - Other)

	Small Data Set	Full Data Set
Test Error	0.230483271375	0.412268087653
Runtime	Less than 10 minutes on: master(m4.large), 2 core(m4.xlarge)	Less than 20 minutes on: master(r3.2xlarge), 10 core(r3.2xlarge)

References

Common Crawl Data Format

PubMed FTP Service

The Wikipedia dataset

Thank You!