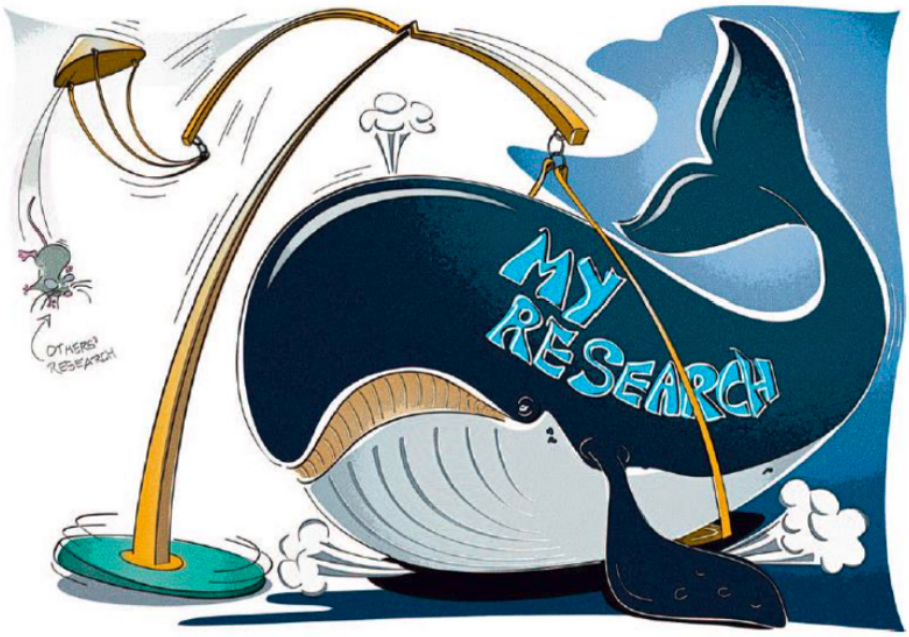
Measuring observable influence and impact of scientific research beyond academia

By CHHAVI VERMA, SHRAY KHANNA, HONGHUI WANG

. . .



INTRODUCTION

The world around us is highly affected by various scientific researches. Health being the priority, a vast number of researches are being carried out in the Healthcare sector. But, nowadays, more and more investors want the researchers to plan and display the impact of their work. By doing this, we will be able to channel resources and efforts in the right direction which will, in turn, benefit human society and the environment. For instance, investors will be able to invest wisely and gain more profits. Also, the researchers will be able to know the impact

of their work and can work on future projects accordingly using our analysis as a guide.

Our approach is to extract references from various downstream documents and match them with the list of academic publications available.

Since 2001, Genome British Columbia has been investing in life science research in BC to address challenges in key sectors such as health, forestry, fisheries, aquaculture, agrifood, energy, mining, and environment. As part of Genome BC's efforts to measure the impact of our funded research and demonstrate accountability to our stakeholders, we would like to identify how academic publications generated by funded research projects, translate to real-world impacts. We have been provided a dataset by Genome BC which contains a list of academic publications that have arisen from Genome BC funded research. The Genome BC's dataset contains Publication Category, Publication Year, PMID, DOI, Authors, Title, Journal, Short Citation associated with the published papers.

The standard unique identifiers for a research article are as follows-

> *PMID—The PubMed Indexing Number, or PMID, is assigned to each article as it is added to the PubMed database. This is a number used by PubMed to index the literature within MEDLINE, the U.S. National Library of Medicine's (NLM) premier bibliographic database*

> *PMCID (PMC ID)—the PubMed Central referencing number, or PMCID, is required in NIH grants proposals, applications, and reports. This number is assigned to an article when is it entered into PubMed Central, a free digital database of full-text scientific literature in biomedical and life sciences.*

> *Manuscript ID—(available only for articles that came in through a manuscript submission system, e.g., NIHMS, Europe PMC, PMC Canada)*

> *DOI—A DOI, or Digital Object Identifier, is a character string used to uniquely and permanently identify an electronic object (like a permanent URL). The DOI has applications that reach far beyond the realm of scientific literature; however, the organization of scholarly material is one primary function.*

The downstream documents used in our project are-

> *1. BC Cancer clinical guidelines website: http://www.bccancer.bc.ca/health-professionals (Clinical Resources, Clinical Trials and Studies)*

> *2. BC Ministry of Health Guidelines: https://www2.gov.bc.ca/gov/content/health/practitioner-professional-resources/bc-guidelines*

> *3. PharmGKB—https://www.pharmgkb.org/*

> *4. CPIC Guidelines: https://cpicpgx.org/guidelines/*

> *5. CADTH pan-Canadian Oncology Drug Review: https://www.cadth.ca/pcodr*

We denote them as bcc, bcm, cadth, cpic and pgkb.

. . .

MOTIVATION

We want to check the impact of research publications in the Healthcare sector which will help the scientists understand the consequence of their work and researches with a good scope will be able

to secure more funds from the investors. We want to make Healthcare research more impactful and beneficial for the welfare of people. We want to help the scientists in finding the true beyond academia impact of their research so that they can find the right direction to continue future work in.

. . .

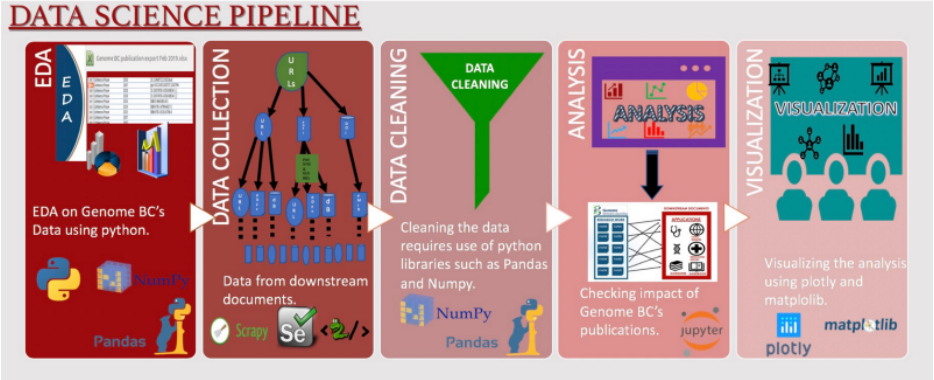
PROBLEM STATEMENT

As part of Genome BC's efforts to measure the impact of their funded research and demonstrate accountability to the stakeholders, we would like to identify how academic publications are generated by funded research projects and translate them to real-world impacts.

This project is also beneficial for the company as any publications made through them which are implemented in the "real life scenarios" makes sure that the company is going in the right direction and from a business point of view their shares and market value will increase if someone uses their research work for implementing.

. . .

DATA SCIENCE PIPELINE



The pipeline used for measuring the impact of academic papers

Firstly, we conduct EDA on both downstream documentations and Genome BC funded paper list. Then we get all the direct and indirect references for every downstream documentation. We represent those reference papers using their PMIDs. Then we clean and integrate the data we got. After that, we count how many times the Genome BC funded papers are cited in the downstream documentations both directly and indirectly. In the end, we visualize the relation between downstream documentations and Genome BC funded papers and show the most influential papers by their cited counts.

. . .

METHODOLOGY

EDA:

We conducted EDA using Numpy, Pandas and Excel on Genome BC's dataset which lists the academic papers funded by them. We explored the downstream documentations using web browser, collecting their citation methods such as by presenting PMID, PMCID, DOI or giving pdf files directly.

Data Collection:

Collection of data is the most challenging part of our project. The downstream documents are variable. By EDA, we find that we can use

PMID to represent the papers they cited as most of them are medicine field papers. (If we want to extend our work to the other fields, we can use DOI instead.)

The data is collected from 5 downstream documents entries:

bcm

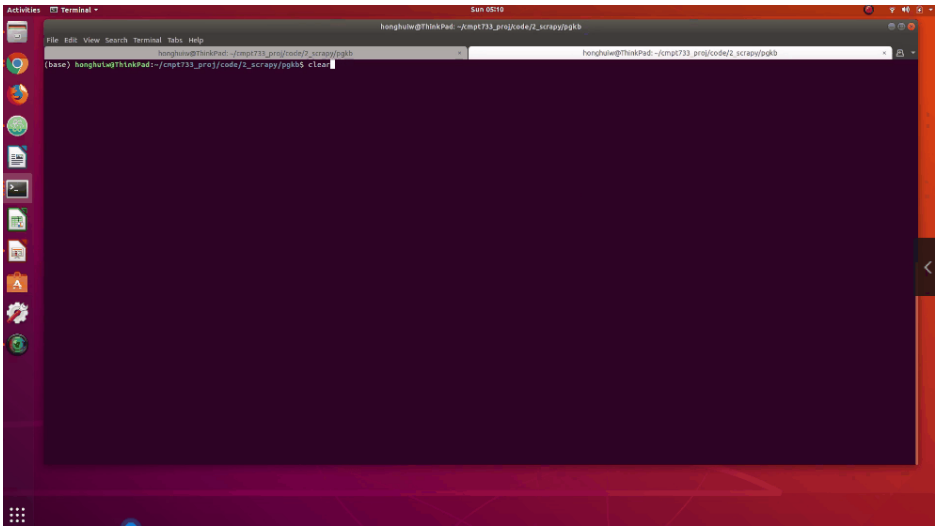
It has a zip file that contains all the information that the website hosts. So we get all the pdf files by downloading the zip file and unzipping it manually.

bcc, cadth and cpic

We use Scrapy to scrape all the pdf files and extract all the DOI, PMCID and PMID references.

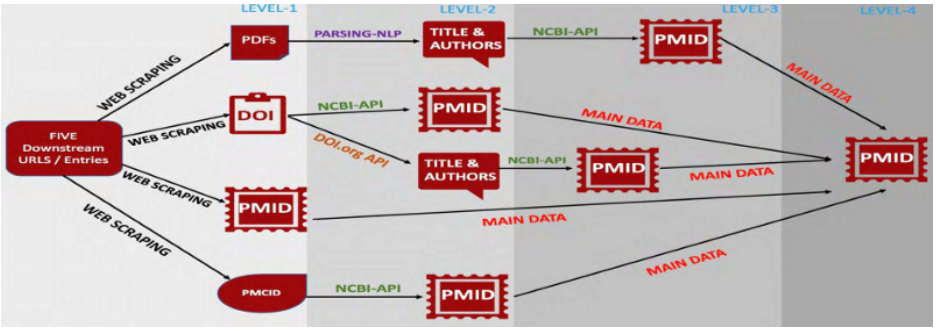
pgkb

pgkb is a javascript-generated dynamic website, so we used Scrapy combined with Selenium to scrape the pdf files and extract all the DOI, PMCID and PMID references.



Web scraping using Scrapy combined with Selenium

Extracting references directly from these documents does not solve the problem as we have to dig in deep by going to each link or reference taken from the previous reference. This task is done up-to a threshold due to the limited computing power and internet availability.



Data Cleaning and Integration:

After collecting the data, for DOI, PMID, and PMCID references, we clean it using python libraries -Numpy and Pandas combined with NCBI APIs which fulfill the null values and get rid of the duplicates.

For the pdf files, at first sight, extracting references from pdf files seems hard. But in fact, as Chenet mentioned [2], reference information is, in fact, structured information. Using regular expression alone can achieve a satisfying result [5].

Still, this problem is not easy, as each website contains many different structures for PDFs and parsing them requires different approaches. We firstly parse pdf files using PDF2XML to get structured data. The lxml library is used to get to reference using XPATH. The implementation of this part becomes more tedious as the type of pdf files increasing.

Then we extract reference paper titles using regular expression combined with NCBI APIs. We set the regular expression rule like this: every title is between two periods, and the number of words in the title should greater than 4. Then we verify this title by querying NCBI API. If we can find a paper the title of whom has an edit distance less than 4 to the title, then we can be sure that the title is indeed a valid paper title. In the meantime, we get the corresponding PMID of this paper.

We use edit distance to determine whether the two papers are the same one based on this fact: each paper's title is unique. The similarity method, such as Jaccard Similarity, will fail because papers focused on one specific topic will always have similar titles.

Below is our pdf parsing result, we extract 1246 reference paper titles from 4379 pdf files:

3 get pdf_title_pmid.csv

In [ ]:
```
pdf_title_pmid={'Title':[], 'PMID':[]}
cnt = 0
import time
ss = time.time()

for title in titles:
    pmid = title2pmid(title)
    cnt+=1
    if cnt%100 ==0:
        print(cnt)
        print(time.time()-ss)
    if pmid is not None:
        pdf_title_pmid['Title'].append(title)
        pdf_title_pmid['PMID'].append(pmid)
```

In [78]:
```
print(len(pdf_title_pmid['Title']))
import pandas as pd
pdf_title_pmid_df = pd.DataFrame(pdf_title_pmid)
pdf_title_pmid_df.to_csv('pdf_title_pmid.csv',index=False)
```

```
1246
```

In [79]:
```
pdf_title_pmid_df.head(20)
```

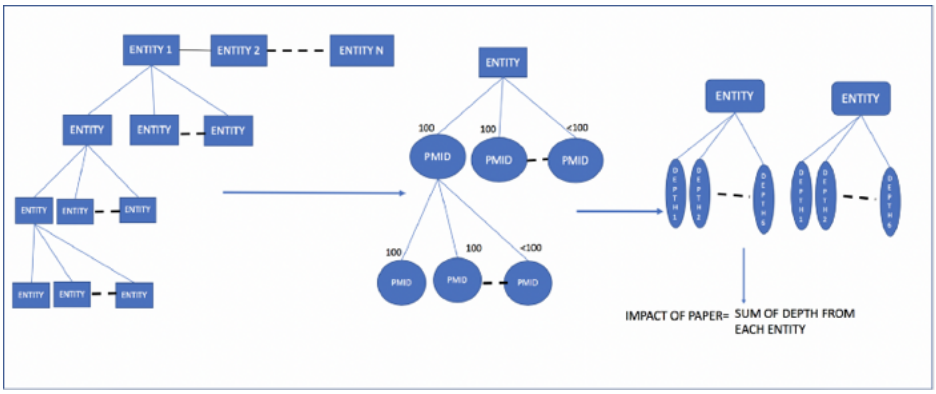
Out[79]:

	Title	PMID
0	Lapatinib plus capecitabine in women with HER-...	20736298
1	Lapatinib plus capecitabine for HER2-positive ...	17192538
2	Clinical Cardiac Tolerability of Trastuzumab	14722042
3	Bleeding risk with trastuzumab (Herceptin) tre...	10612314
4	Clinical cardiac tolerability of trastuzumab	14722042
5	High rate of febrile neutropenia in patients w...	19652050
6	Impact of colony-stimulating factors to reduce...	20232087
7	Randomized trial of dose-dense versus conventi...	12668651
8	Benefit of a high-dose epirubicin regimen in a...	11157009
9	Clinical cardiac tolerability of trastuzumab	14722042
10	High rate of febrile neutropenia in patients w...	19652050
11	Impact of colony-stimulating factors to reduce...	20232087
12	Clinical Cardiac Tolerability of Trastuzumab	14722042
13	Phase II study of weekly docetaxel and trastuz...	11919237
14	docetaxel combined with trastuzumab is an acti...	15020608
15	Clinical Cardiac Tolerability of Trastuzumab	14722042

Analysis:

We measure the impact of a Genome BC funded paper in real life by counting how many times they are cited in the downstream documentations both directly and indirectly.

We get all the direct and indirect cited papers for every downstream documentation and represent it as a tree.



Accelerate the generation of the reference graph

At first glance, it seems a trivial task—just perform BFS to get the reference tree by query NCBI database. But as 1000 downstream entities will refer to around 40 million papers given a depth constraint of 6, plus that the NCBI API only allows querying 10 times every second, this becomes an impossible mission. We come up with an accelerated solution: as we only concern about the depth but not the actual structure, we can merge nodes at the same depth.

As NCBI API allows a 100 id list query every time, we merge the nodes into a package of 100 nodes and do the query. This method decreases the number of nodes from 40 million to around 400 thousand and reduces the running time to 25 hours given a depth constraint of 6.

After the query, we merge all the nodes of the same depth into one node, which further reduce the running time for later analysis.

Finally, we will go and measure the real-life impact of Genome BC funded papers. It is somehow straightforward: go through the reference map using Breadth-First-Search (BFS) to do the counting for our Genome BC funded papers.

Finally, we find the actual impact of research papers in the real world.

Visualization:

In order to show the relation between downstream documentations and the Genome BC funded papers, we use igraph to construct a map with coordinates associated with every paper and downstream

documentation entity. Then, we plot the map using plotly. Plotting this helps both the researchers and the company to move into the right direction. It provides motivation to researchers and at the same time increases revenue for the company. These figures give a better picture to the company and are understandable to a greater audience.

. . .

Tool Selection

The list of tools and technologies used in this project are as follows-

> *Scrapping Tools : Scrapy, Selenium*

> *Python libraries : Numpy, Pandas, rdflib*

> *Analytics Tools : pdf2xml, pdfssa4me, pdfextract, lxml, Jupyter notebook, Microsoft Excel*

> *Data Visualization Tools : python-igraph, plotly*

The various APIs used are listed below-

> *https://www.ncbi.nlm.nih.gov/home/develop/api/*

> *https://support.datacite.org/docs/api*

> *https://www.doi.org*

Seeking Guidance

We have contacted our mentor Jessica Lu, Corporate Analyst, Genome British Columbia via Email at regular times for closely understanding the problem and working out the best way to solve the problem.

Challenges

Data collection is the biggest challenge faced by us. We had to go through various data sources, let alone the different depths for getting references from the downstream documents we have to search. For doing so, we have used web scrapping, PDF parsing, and APIs. At the last stage, we have the list of PMIDs for all the papers.

. . .

EVALUATION

We have divided the product into 3 parts. The first is the scraping module which gets the PDFs, PMIDs, DOIs and PMCIDs. We used Selenium and Scrapy for carrying out this job and were successful enough to extract all the downstream entities. We were also successful in extracting all the references through PMIDs, PMCIDs and DOIs and this helped us to go to next level or depth to extract the linked reference. Our script ran for a longer duration due to less computing power and still achieved a high success in extraction of the resources. We were successful in extracting PMIDs from DOIs and PMCIDs by using NCBI API. This API allows 1 extraction per search but can be changed to 100 as bulk. NCBI API was provided to us by Genome BC and we used it for extraction of PMIDs for every level.

The PDFs are parsed using Natural Language Processing so that we can extract references from every such website. The parsing is done using Semantic Analysis where we've used regular expression to get to reference and extract them. A single website has 1000s of PDFs and each website has at least 30 structures for PDFs. Extracting these was a hard task for us as every new structure has a different approach and we cannot scroll through 5000 PDFs. Our approach cannot extract 100% but is able to extract at least 80% from the PDFs. Considering that PDFs

are unstructured data and extraction with so many different structures requires more time and more computing; our script can still get good results. To get these results we used a tool named PDF2XML which made the pdfs to XML trees very efficiently and we could extract references from this tree. The csv made from these extracted references still contain redundant values, null values and other sentences which are not part of the reference. We cleaned this csv by further validate the titles by querying NCBI APIs.

After we get all the references, we analyze the result. To ensure better visualizations so as to get better results we have used igraph to plot the links with Published papers and downstream documents. Igraph helped us to get a better efficiency as putting these PMIDs as nodes to get information on connectivity with the downstream documents gave us a figure to help and find the most impactful papers from the given list of published papers. We first went till depth 3 which states the linkages between the downstream documents in order to get to the PMID of published papers. After analyzing the shallow depth of main bar plot we went to depth 6 which makes a very complex graph but the results of this graph gave efficient insights as the bar plot for depth 6 showed the impact of downstream documents more clearly.

. . .

DATA PRODUCT

Our data product is the report below showing the insights of the Genome BC funded papers.

The graph below shows the interconnectedness of Genome BC's papers. Each crowd indicates a topic. These topics are the current topics that are being worked on and thus, it also shows that the company is focused on a particular subject more. The loosely connected nodes are the research areas or the topics that are not very popular or the company does not want to invest in it.

This **graph below** is a 2D representation of the above graph. To make it more convenient for the user to understand. We can see the title of the paper by hovering over both the graphs. This shows famous or common research areas or topics the company is interested in clarity.

Connectedness between Genome BC's published papers and downstream documents with a DEPTH constraint of 3:

The red circle shows the Genome BC's paper while the blue ones are the downstream documents. The number between every line is the reference depth the script had to go to, to get the reference link.

When you hover over the red circle you see the titles and a number in parenthesis. This number demonstrates the impact or the references it got from the downstream documents. This graph shows the relation between downstream documentations and Genome BC funded paper with a reference depth of 3.

Below shows the top 10 most influential papers:

It showcases the published papers of Genome BC that are the most influential and has the most impact on the procedures or guidelines. The ones with yellow means they are referred more directly, it may infer that they are new papers or more practical papers.

Digging in more deeply to get better insights (DEPTH 6):

With a reference depth constraint of 6, we can get more insights and have a better understanding. The graph made from this helps us to understand more on the impact of Genome BC's papers (red circles) and how they are connected to the real world.

The top 20 most influential papers with a reference DEPTH of 6:

This graph showcases the most influential papers in Genome BC's dataset but for a depth 6. The depth 6 indicates that it is linked through many references to the downstream document. Thus, this paper has more impact on the market and are more fundamental. This analysis also helps companies to realize the area they should work on more to increase their revenue at the same time. But this also occludes the new papers as they will not have so much reference, we can unselect the deep depth to show the newer ones.

Below shows the 50 most influential Genome BC funded papers along with their cited counts:

```
The 50 most influential Genome BC funded papers and their cited counts are:

Count   Title
210     Comparative genomics of the eukaryotes.
209     The ENCODE (ENCyclopedia Of DNA Elements) Project.
207     Genome sequence of the Brown Norway rat yields insights into mammalian evolution.
199     A physical map of the human genome.
199     Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences.
164     A tiling resolution DNA microarray with complete coverage of the human genome.
161     A comprehensive analysis of common copy-number variations in the human genome.
154     The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MG
C).
152     Evolutionary and biomedical insights from the rhesus macaque genome.
139     Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome.
135     ChIPSeq: Mapping of Protein-DNA Interactions.
132     A physical map of the mouse genome.
131     Functional genomic analysis of cell division in C. elegans using RNAi of genes on chromosome III.
130     A set of BAC clones spanning the human genome.
127     Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal an
d transformed human cells.
126     Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel s
equencing.
120     Systematic sequencing of cDNA clones using the transposon Tn5.
117     cisRED: a database system for genome-scale computational discovery of regulatory elements.
117     Oligonucleotide microarray analysis of genomic imbalance in children with mental retardation.
116     Integrating copy number polymorphisms into array CGH analysis using a robust HMM.
115     SeeGH--a software tool for visualization of whole genome array comparative genomic hybridization data.
115     Resolving the resolution of array CGH.
114     An efficient strategy for large-scale high-throughput transposon-mediated sequencing of cDNA clones.
112     Mapping segmental and sequence variations among laboratory mice using BAC array CGH.
111     High resolution analysis of non-small cell lung cancer cell lines by whole genome tiling path array CGH.
110     Comprehensive copy number profiles of breast cancer cell model genomes.
109     Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences.
108     Identification by full-coverage array CGH of human DNA copy number increases relative to chimpanzee and go
rilla.
108     Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencin
g.
107     A stepwise framework for the normalization of array CGH data.
107     ORegAnno: an open access database and curation system for literature-derived promoters, transcription fact
or binding sites and regulatory variation.
103     Sequence biases in large scale gene expression profiling data.
101     Assessment and integration of publicly available SAGE, cDNA microarray, and oligonucleotide microarray exp
ression data for global coexpression analyses.
99      Array CGH technologies and their applications to cancer genomes.
95      A global profile of germline gene expression in C. elegans.
95      Sockeye: a 3D environment for comparative genomics.
95      A mammalian organelle map by protein correlation profiling.
93      Systematic recovery and analysis of full-ORF human cDNA clones.
93      Personalized copy number and segmental duplication maps using next-generation sequencing.
92      The new paradigm of flow cell sequencing.
90      Full-genome RNAi profiling of early embryogenesis in Caenorhabditis elegans.
89      Cerebral: a Cytoscape plugin for layout of and interaction with biological networks using subcellular loca
lization annotation.
85      Purification and unique properties of mammary epithelial stem cells.
85      When good drugs go bad.
85      ORegAnno: an open-access community-driven resource for regulatory annotation.
```

. . .

LESSONS LEARNT

- Learned about research papers and their different unique identifiers.
- Got to know about various APIs such as NCBI APIs, crossref APIs, and doi.org APIs which help to get the metadata of a research paper.
- Marsted Scrapy and learned how to combine it with Selenium.
- Working with Python libraries such as rdflib.
- We have learnt NLP techniques for getting references from PDFs.
- Learned how to parse pdf files and extract the references.

- Learned python-igraph which helps understand connections among nodes better.
- Encounter and solve many real problems such as unstable internet and limited API usage when data set becomes huge.
- Used plot.ly online mode, and find it really convenient and integrate the figures with Medium.
- Learned how to make a video.

. . .

SUMMARY

In this project, we have observed the impact of Genome BC's academic publications on the real world by observing the references in downstream documents. We have provided the insights by showing the relation graphs generated by igraph and plotly bar charts which help us understand the connections between Genome BC publications and the downstream documents. The more impactful the publication is, the more connections it will have and the higher the reference depth is, the more powerful and fundamental the academic publication is. We have also found the top influencers of Genome BC publications for different depths.

. . .

References

[1] Ravenscroft J, Liakata M, Clare A, Duma D. Measuring scientific impact beyond academia: An assessment of existing impact metrics and proposed improvements. PloS one. 2017 Mar 9;12(3):e0173152.

[2] Chenet M. *Identify and extract entities from bibliography references in a free text* (Master's thesis, University of Twente).

[3] Connan J, Omlin CW. Bibliography extraction with hidden markov models.

[4] Tkaczyk D, Szostek P, Fedoryszak M, Dendek PJ, Bolikowski Ł. CERMINE: automatic extraction of structured metadata from scientific literature. International Journal on Document Analysis and Recognition (IJDAR). 2015 Dec 1;18(4):317–35.

[5] Tang X, Zeng Q, Cui T, Wu Z. Regular expression-based reference metadata extraction from the web. In2010 IEEE 2nd Symposium on Web Society 2010 Aug 16 (pp. 346–350). IEEE.

. . .

> 😃 **THANK YOU** 😃