Machine Learning of Biomedical Text for Gene Network Knowledge

Published: February 1, 2023

Machine Learning of Biomedical Text for Gene Network Knowledge

A project Report
Presented to
The Faculty of Computer Engineering Department
San Jose State University

In Partial Fulfillment
Of the Requirements for the Degree
Bachelor of Science in Software Engineering

By
Tue Do
Ausin Fong
Thinh Huynh
Mary Markart
02/2023

Authors

- **Mary Markart**
  · 🔘 [marymarkart](#)
  Charles W. Davidson College of Engineering, San Jose State University

- **Austin Fong**
  · 🔘 [austinf01](#)
  Charles W. Davidson College of Engineering, San Jose State University

- **Tue Doe**
  · 🔘 [tuedolm](#)
  Charles W. Davidson College of Engineering, San Jose State University

- **Thinh Huynh** ✉
  · 🔘 [thinhh](#)
  Charles W. Davidson College of Engineering, San Jose State University

✉ — Correspondence possible via [GitHub Issues](#) or email to Thinh Huynh <thinh.huynh@sjsu.edu>.

Abstract

Machine Learning of Biomedical Text for Gene Network Knowledge

By Tue Do, Austin Fong, Thinh Huynh, Mary Markart

Genetic research is the study of DNA to find out what genes and environmental factors contribute to diseases. There are multiple studies conducted on individuals or groups of genes and how they impact health and disease. Thus, there is no lack in the amount of data to draw conclusions from and researchers have the potential to make more connections from within the data.

An issue arises when it takes a long time to find and sift through biomedical research papers to find knowledge useful to researchers. There are already ideas present in literature, but also a lack of finding a connection between already present ideas. Researchers are constantly looking to find these connections; however, there is a limiting factor of time and can be overwhelming without the right tools.

Data from text mining biomedical texts was warehoused through the College of Engineering HPC or AWS DynamoDB and AWS Redshift. Data processing utilized BERT and GNormPlus models to categorize and normalize input. TigerGraph was used to represent the knowledge accumulated to visualize relationships. This project reduced the amount of time needed to search through biomedical research by making it easier to discover knowledge about various biological mechanisms using machine learning.

Chapter 1{.page_break_before}

1.1 Project Goals and Objectives

The goal of this project was to create a usable gene-to-gene knowledge graph to aid in research within the biomedical field. The application takes an input, which is a gene name, and outputs a knowledge graph surrounding the connections of diseases and topics related to the inputted gene. Project objectives consisted of ranking the relevance of the outputs, only giving related results, and giving a result in a reasonable amount of time.

1.2 Problem and Motivation

Researchers must analyze research papers and draw connections to other papers in the process of conducting research. The finding and analyzing of research papers can be a very time-consuming process. Gene research is also very scattered; there can be multiple diseases related to a single gene and finding all these connections can be very cumbersome. Our project solves both of these problems by decreasing the time for research, as well as creating a visualization to help in understanding gene relationships.

1.3 Project Application and Impact

Genetic research studies how individual genes or groups of genes affect health and disease. By improving research times and offering more relevant research surrounding the topic of genes, our project will act as a tool for the creation of new technologies or discoveries that will benefit the well-being of society and the prevention and treatment of diseases.

1.4 Project Results and Deliverables

Our project's results are a knowledge graph over genes and their relationships. The output is a knowledge graph with a label of their connections to a specific gene, which is entered as an input for the tool. Project deliverables consist of eight deliverables.

Deliverable 1: Research problem understanding and preliminary research

The research goal is to find the best correlation papers for an input biomedical gene paper for a map of the most correlated papers and their references. Preliminary research is currently being conducted in machine learning on text mapping and name entity recognition and biomedical research papers to find similar vocabulary for some pre-mapping for training the model later.

Deliverable 2: Collecting data

Biomedical research papers on gene and abstract data will be collected through PubMed and Gene. These abstract data will be mass-collected and analyzed using pre-trained / predefined models like GNormPlus for gene name recognition.

Deliverable 3: Pre-processing data

The data from Pubmed is formatted in either XML or TXT. This data will be stored in the HPC cluster with additional storage power by AWS. Currently, the project will utilize DynamoDB for noSQL data as the data input could become more diverse (Redshift for data warehousing) This data will be stored and automatically processed using python script with some third-party library besides Pandas. The data then needs to be pre-processed to become unified xml or json data.

Deliverable 4: Name entity recognition and relation extraction

After preprocessing data, the data will be processed using BioBERT for text mining training on HPC Cluster to summarize research papers and further create wording correlation mapping on words and weight parameters for graph database construction (maybe the frequency of having similar words between two paper could become the correlation weight scale between two paper). Other parameters like a title / abstract / past collaboration or citations of input paper can be utilized.

Deliverable 5: Knowledge Representation

Once the weight correlation is determined for the connections between papers, TigerGraph will be used to initiate a graph database. This will be implemented for visualizing recommendation research papers with a weight scale to reduce preliminary research time and increase research productivity.

Deliverable 6: Model Evaluation

The model base will be evaluated and compare the model output statistics to some model / service out there in the market on research paper recommendation.

Deliverable 7: Model Deployment

After doing quality assurance, the model was deployed on HPC.

Deliverable 8: Driving insights and generating Tableau dashboard

Tableau dashboards can be created to compare metrics between the model produced and other models on the market for performance comparison as well as the current model and previous performance to make decisions based on insights to change the machine learning model parameters for improvements.

This manuscript is a template (aka "rootstock") for [Manubot](https://...), a tool for writing scholarly manuscripts. Use this template as a starting point for your manuscript.

The rest of this document is a full list of formatting elements/features supported by Manubot. Compare the input (`.md` files in the `/content` directory) to the output you see below.

Basic formatting

Bold text

Semi-bold text

<div align="center">Centered text</div>

<div align="right">Right-aligned text</div>

Italic text

Combined *italics and **bold***

~~Strikethrough~~

 1. Ordered list item

2. Ordered list item
    a. Sub-item
    b. Sub-item
        i. Sub-sub-item
3. Ordered list item
    a. Sub-item

- List item
- List item
- List item

subscript: H_2O is a liquid

superscript: 2^{10} is 1024.

[unicode superscripts](#)^{0123456789}

[unicode subscripts](#)$_{0123456789}$

A long paragraph of text. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Putting each sentence on its own line has numerous benefits with regard to [editing](#) and [version control](#).

Line break without starting a new paragraph by putting
two spaces at end of line.

Document organization

Document section headings:

Heading 1

Heading 2

Heading 3

Heading 4

Heading 5

Heading 6

A heading centered on its own printed page

Horizontal rule:

`Heading 1`'s are recommended to be reserved for the title of the manuscript.

`Heading 2`'s are recommended for broad sections such as *Abstract*, *Methods*, *Conclusion*, etc.

`Heading 3`'s and `Heading 4`'s are recommended for sub-sections.

Links

Bare URL link: https://manubot.org

[Long link with lots of words and stuff and junk and bleep and blah and stuff and other stuff and more stuff yeah](#)

[Link with text](#)

[Link with hover text](#)

[Link by reference](#)

Citations

Citation by DOI [1].

Citation by PubMed Central ID [2].

Citation by PubMed ID [3].

Citation by Wikidata ID [4].

Citation by ISBN [5].

Citation by URL [6].

Citation by alias [7].

Multiple citations can be put inside the same set of brackets [1,5,7]. Manubot plugins provide easier, more convenient visualization of and navigation between citations [2,3,7,8].

Citation tags (i.e. aliases) can be defined in their own paragraphs using Markdown's reference link syntax:

Referencing figures, tables, equations

Figure 1

Figure 2

Quotes and code

> Quoted text

> Quoted block of text
>
> Two roads diverged in a wood, and I—
> I took the one less traveled by,
> And that has made all the difference.

Code `in the middle` of normal text, aka `inline code`.

Code block with Python syntax highlighting:

```python
from manubot.cite.doi import expand_short_doi

def test_expand_short_doi():
    doi = expand_short_doi("10/c3bp")
    # a string too long to fit within page:
    assert doi == "10.25313/2524-2695-2018-3-vliyanie-enhansera-copia-i-
        insulyatora-gypsy-na-sintez-ernk-modifikatsii-hromatina-i-
        svyazyvanie-insulyatornyh-belkov-vtransfetsirovannyh-geneticheskih-
        konstruktsiyah"
```

Code block with no syntax highlighting:

```
Exporting HTML manuscript
Exporting DOCX manuscript
Exporting PDF manuscript
```

Figures

Figure 1: A square image at actual size and with a bottom caption. Loaded from the latest version of image on GitHub.

Figure 2: An image too wide to fit within page at full size. Loaded from a specific (hashed) version of the image on GitHub.

Figure 3: A tall image with a specified height. Loaded from a specific (hashed) version of the image on GitHub.

Figure 4: A vector `.svg` image loaded from GitHub. The parameter `sanitize=true` is necessary to properly load SVGs hosted via GitHub URLs. White background specified to serve as a backdrop for transparent sections of the image. Note that if you want to export to Word (`.docx`), you need to download the image and reference it locally (e.g. `content/images/vector.svg`) instead of using a URL.

Tables

Table 1: A table with a top caption and specified relative column widths.

Bowling Scores	Jane	John	Alice	Bob
Game 1	150	187	210	105
Game 2	98	202	197	102
Game 3	123	180	238	134

Table 2: A table too wide to fit within page.

	Digits 1-33	Digits 34-66	Digits 67-99	Ref.
pi	3.14159265358979323 846264338327950	28841971693993751 0582097494459230	78164062862089986 2803482534211706	`piday.org`
e	2.71828182845904523 536028747135266	24977572470936999 5957496696762772	40766303535475945 7138217852516642	`nasa.gov`

Table 3: A table with merged cells using the `attributes` plugin.

Size	Colors	
	Text Color	Background Color
big	blue	orange
small	black	white

Equations

A LaTeX equation:

$$\int_0^\infty e^{-x^2}dx = \frac{\sqrt{\pi}}{2}$$

(1)

An equation too long to fit within page:

$$x = a + b + c + d + e + f + g + h + i + j + k + l + m + n + o + p + q + r + s + t$$
$$+ u + v + w + x + y + z + 1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 + 9$$

(2)

Special

⚠ **WARNING** *The following features are only supported and intended for* `.html` *and* `.pdf` *exports. Journals are not likely to support them, and they may not display correctly when converted to other formats such as* `.docx` *.*

LINK STYLED AS A BUTTON

Adding arbitrary HTML attributes to an element using Pandoc's attribute syntax:

> Manubot Manubot Manubot Manubot Manubot. Manubot Manubot Manubot Manubot. Manubot Manubot Manubot. Manubot Manubot. Manubot.

Adding arbitrary HTML attributes to an element with the Manubot `attributes` plugin (more flexible than Pandoc's method in terms of which elements you can add attributes to):

> Manubot Manubot Manubot Manubot Manubot. Manubot Manubot Manubot Manubot. Manubot Manubot Manubot. Manubot Manubot. Manubot.

Available background colors for text, images, code, banners, etc:

`white` `lightgrey` `grey` `darkgrey` `black` `lightred` `lightyellow` `lightgreen` `lightblue` `lightpurple` `red` `orange` `yellow` `green` `blue` `purple`

Using the Font Awesome icon set:

✔ ? ★ 🔔 ✖ ⋯

📜 Light Grey Banner
useful for *general information* - <u>manubot.org</u>

ⓘ Blue Banner
useful for *important information* - <u>manubot.org</u>

🚫 Light Red Banner
useful for *warnings* - <u>manubot.org</u>

References

1. **Sci-Hub provides access to nearly all scholarly literature**
   Daniel S Himmelstein, Ariel Rodriguez Romero, Jacob G Levernier, Thomas Anthony Munro, Stephen Reid McLaughlin, Bastian Greshake Tzovaras, Casey S Greene
   eLife (2018-03-01) https://doi.org/ckcj
   DOI: 10.7554/elife.32822 · PMID: 29424689 · PMCID: PMC5832410

2. **Reproducibility of computational workflows is automated using continuous analysis**
   Brett K Beaulieu-Jones, Casey S Greene
   Nature biotechnology (2017-04) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6103790/
   DOI: 10.1038/nbt.3780 · PMID: 28288103 · PMCID: PMC6103790

3. **Bitcoin for the biological literature.**
   Douglas Heaven
   Nature (2019-02) https://www.ncbi.nlm.nih.gov/pubmed/30718888
   DOI: 10.1038/d41586-019-00447-9 · PMID: 30718888

4. **Plan S: Accelerating the transition to full and immediate Open Access to scientific publications**
   cOAlition S
   (2018-09-04) https://www.wikidata.org/wiki/Q56458321

5. **Open access**
   Peter Suber
   MIT Press (2012)
   ISBN: 9780262517638

6. **Open collaborative writing with Manubot**
   Daniel S Himmelstein, Vincent Rubinetti, David R Slochower, Dongbo Hu, Venkat S Malladi, Casey S Greene, Anthony Gitter
   Manubot (2020-05-25) https://greenelab.github.io/meta-review/

7. **Opportunities and obstacles for deep learning in biology and medicine**
   Travers Ching, Daniel S Himmelstein, Brett K Beaulieu-Jones, Alexandr A Kalinin, Brian T Do, Gregory P Way, Enrico Ferrero, Paul-Michael Agapow, Michael Zietz, Michael M Hoffman, … Casey S Greene
   Journal of The Royal Society Interface (2018-04) https://doi.org/gddkhn
   DOI: 10.1098/rsif.2017.0387 · PMID: 29618526 · PMCID: PMC5938574

8. **Open collaborative writing with Manubot**
   Daniel S Himmelstein, Vincent Rubinetti, David R Slochower, Dongbo Hu, Venkat S Malladi, Casey S Greene, Anthony Gitter
   PLOS Computational Biology (2019-06-24) https://doi.org/c7np
   DOI: 10.1371/journal.pcbi.1007128 · PMID: 31233491 · PMCID: PMC6611653