LEVERAGING AI TO IDENTIFY THE KOLs OF THE FUTURE

Introduction

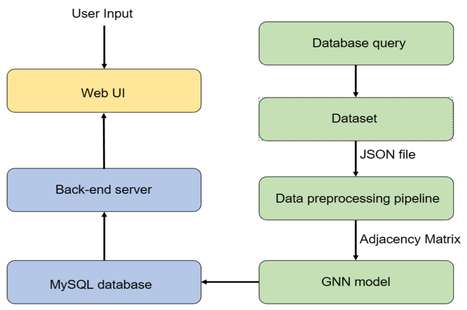
- Accurately identifying Key Opinion Leaders (KOLs) in the pharmaceutical industry can make a huge difference in the successful introduction of new therapeutic treatments to markets.
- The system consists of a back-end machine learning pipeline based on GNN method to process data and a front-end application interface allowing users to search by a field of medicine and returning a list of KOLs.
- Through this system, Genmab will be able to effectively identify and target the most influential individuals in their selected areas of interest.

The Definition of a KOL

- In the healthcare domain, KOLs can be members of the academic community who have advanced knowledge or have contributed in important ways for specific health issues.
- In this project, we narrow down our application scenario to identify the KOLs in the Immunology and the Oncology domain for certain disease.
- We consider several factors to decide whether a person is a KOL: number of papers, number of co-authorships, participation in clinical trials, and years of experience.

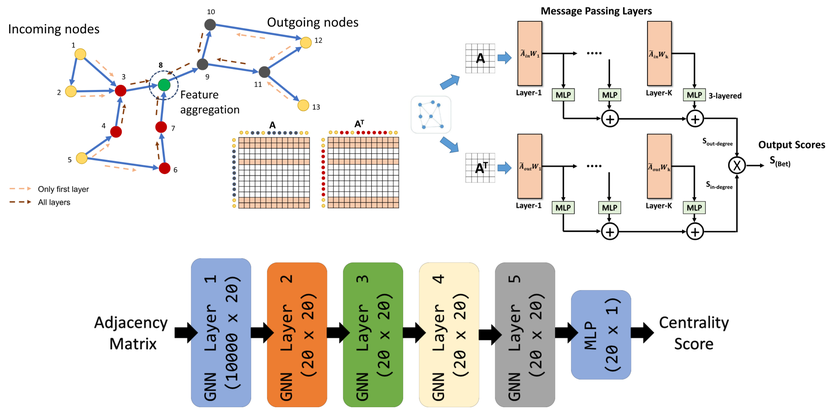
System Overview

- In the Data Query module, we utilize APIs provided by the database and write Python scripts to automatically retrieve the required data instances and fields.
- In the Data Preprocessing module, we use the NetworkX and Networkit lib to convert the dataset into a Graph structure and compute the adjacency matrix and Centrality measures.
- In the Machine Learning Algorithm module, we build a GNN model based on PyTorch framework to fit the Centrality measures of the nodes in the dataset. Finally, we compute the score for each node with the Centrality measures and Years of Experience and rank KOLs with the score.
- In the Web development module, we store the data using MySQL database and build the back-end server and front-end web page based on Django framework, in which user can search KOLs with several key words.



ML Technique: Graph Neural Network

- Researchers have proposed centrality measures for node ranking and these measures provide us a quantitative view of how nodes play a certain role in the graph.
- In this project, we use Betweenness Centrality (BC), Closeness Centrality (CC), and Degree Centrality (DC) to rank the nodes in the citation graph.
- A high value of BC for a node means this node lies on many shortest paths between other nodes.
- Higher CC means that the node can spread information easier to other nodes.
- A high value of DC means that the number of connected nodes is high.
- GNN plays an important part in accelerating the computation speed to get those centrality scores compared to the traditional computing method.
- Our final ranking score calculate with this formula: $SCORE = DC + BC + CC + 0.2 * YOE$



- For the model training step, the model uses some randomly generated Scale-free Graph dataset as the training set. The code generates graphs with the number of nodes varying from 5,000 to 10,000. 40 graphs are used for training and 10 graphs for testing.
- We evaluate the result using KT score which measures the similarity between two vectors, with a higher KT score indicating greater similarity, as it approaches 1.
- The following table shows the extent to which the fitted values of the GNN model for the centrality metric are similar to the true values for different datasets.

Input

```
"ArticleTitle": "Risk of pancreatic ductal
"AuthorList": [
    "Jennifer L McGarry",
    "Ben Creavin",
    "Michael E Kelly",
    "Tom K Gallagher"
],
"AffiliationList": [
    "Department of Hepatobiliary Surgery,
    "Department of Hepatobiliary Surgery,
    "Department of Hepatobiliary Surgery,
    "Department of Hepatobiliary Surgery,
],
"Years": [
    {
        "Year": "2022",
        "Month": "06",
        "Day": "30"
    }
]
```

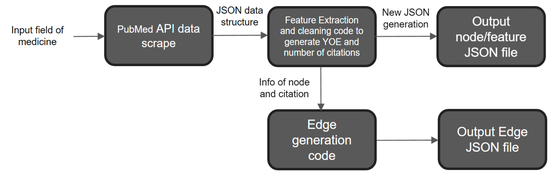
Output

```
Result:
top 1: Rolando Herrero
top 2: Allan Hildesheim
top 3: Joan L Walker
top 4: Annarosa Del Mistro
top 5: Pekka Nieminen
top 6: Phillip E Castle
top 7: Mahboobeh Safaeian
top 8: Paolo Dalla Palma
top 9: Carolina Porras
top 10: Anna Gillio-Tos
top 11: Francesco Fanfani
top 12: Matti Lehtinen
top 13: Anna R Giuliano
top 14: Francesca Carozzi
top 15: Mark H Einstein
top 16: Peter G Rose
top 17: Massimo Confortini
top 18: Mark Schiffman
top 19: Silvia Franceschi
top 20: Christhardt Köhler
top 21: Paolo Giorgi-Rossi
top 22: Satoshi Yamaguchi
top 23: Diane Solomon
top 24: Taro Shibata
top 25: Laura De Marco
top 26: Laura A Koutsky
```

	Betweenness Centrality	Closeness Centrality
ER	0.8981	0.9571
SF	0.9748	0.8941
GRP	0.8924	0.9468

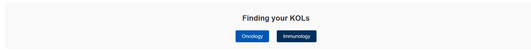
Data Processing Pipeline

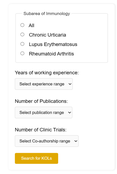
- For co-authorships, we could extract and rank this feature by calculating the Centrality score from the citation network.
- We construct a citation graph and a co-trials graph using NetworkX lib.



Web Application

- We built the web application based on Python's Django framework and a MySQL database. Users can select different filters, e.g., number of papers, years of experience, and the web page will query the database and return a list of the corresponding KOLs, which will eventually be displayed on the page.
- Table design:
  - Table kol_info : stores the KOL details.
  - Table pub_list : stores the paper information of each KOL respectively.
- API:
  - The list query API will return the result to the front-end through conditional query according to the filter selected by the user.
  - The info query API will get the KOL name field in the GET request and return the details of the KOL and the list of papers.





Future Work, References, and Acknowledgments

- Further improvements to GNN structure and hyperparameters.
- Add more online database as data source to consider more factors of deciding a KOL.

Faculty: Jeremy Meehan, John Raiti
Graduate Students: ChungWei Wu, Manyu Chen, Yizhou Li

[1] Chen Z, Li X, Bruna J. 2021. Graph Neural Networks for Fast Node Ranking Approximation. ACM Trans Knowl Discov Data 15:78.
[2] Gotecha MR, Patwardhan MS. 2016. Identification of key opinion leaders in healthcare domain using weighted Social Network Analysis. Proceedings of the 2016 International Conference on Computing Communication Control and automation (ICCUBEA), Pune, India, pp 1–6.

ELECTRICAL & COMPUTER ENGINEERING
UNIVERSITY of WASHINGTON

ADVISERS: JEREMY MEEHAN, JOHN RAITI, MINGFEI CHEN

SPONSOR: GENMAB