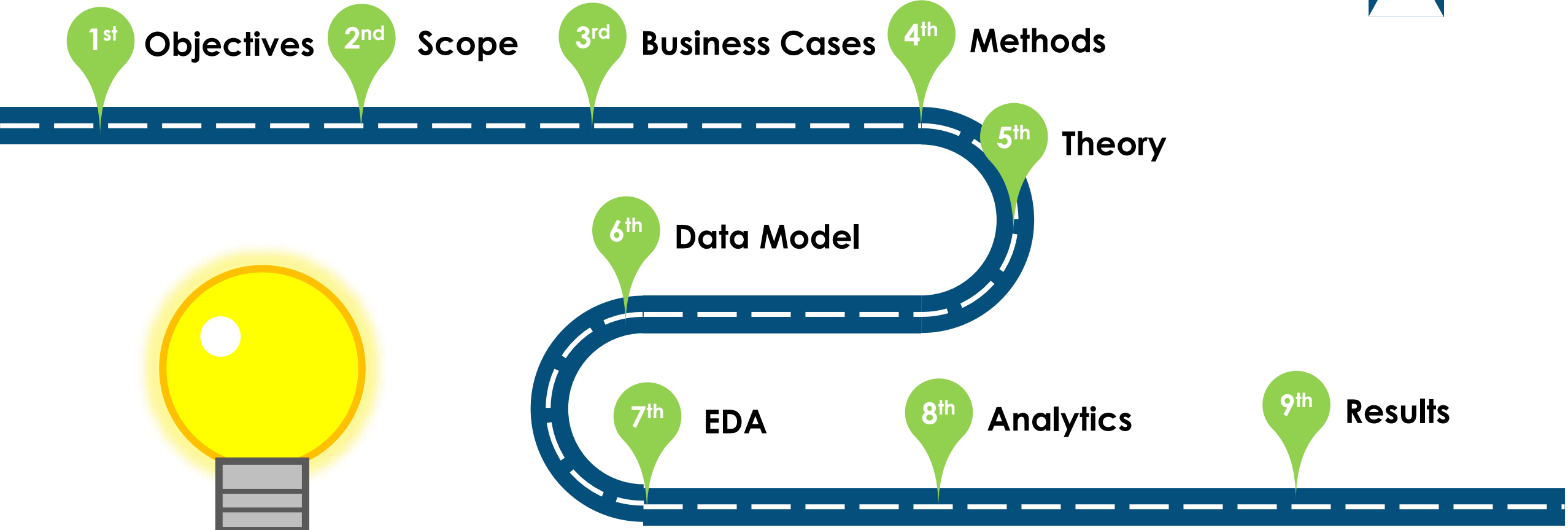
Constructing a Disease-Gene-Drug Knowledge Graph with usage examples.

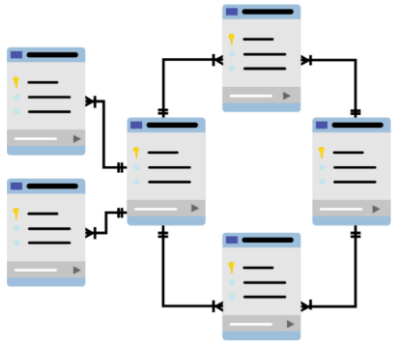
Vipada Siripatanadilok

Index


1st Objectives
2nd Scope
3rd Business Cases
4th Methods
5th Theory
6th Data Model
7th EDA
8th Analytics
9th Results

Motivations

Due to the emerging on tremendous amount of data on drugs and medical field. And the imperfection on inference data, model and on medical data complexation querying.

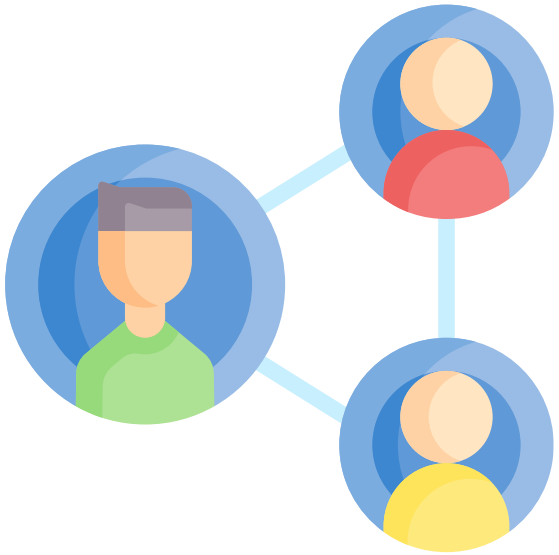


Due to SQL database restrictions on high latency on complex query, cannot store various type of data from text and image, difficulty on scaling the database “Graph database” can solve these problems.

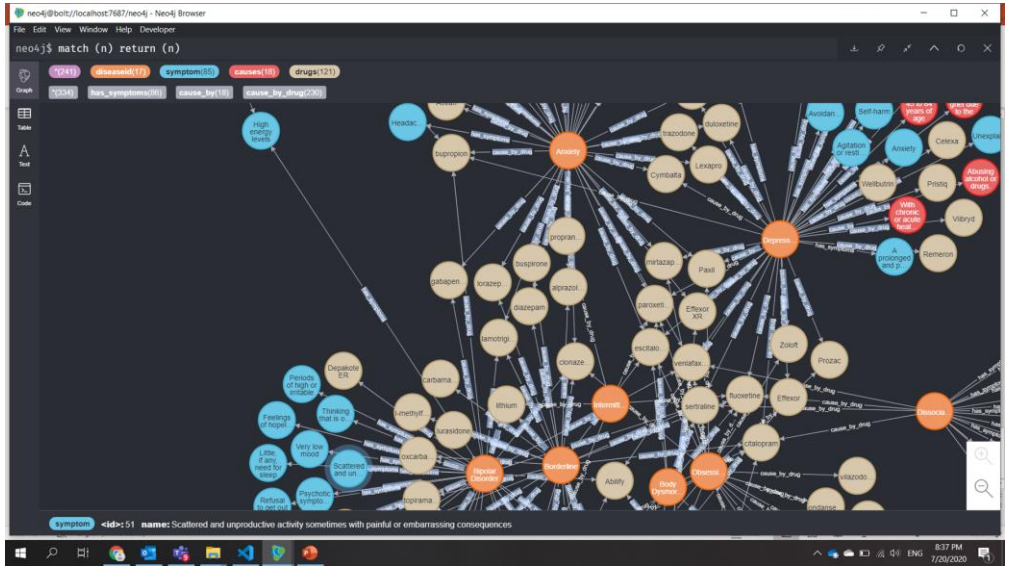
Due to Property of Graph database on interaction between entities that can extract more knowledge/information for analytics.

Knowledge graph* : To helps the potential searching on numerous web content.
Healthcare and medical information emerging tremendously all around the world.
To apply textual drugs and disease’s knowledge in the from of knowledge graph for helping clinicians to recheck drug and symptoms conditions to help correctness decision making.

Objectives

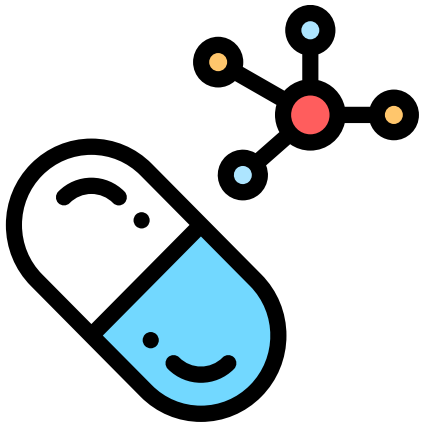


"เพื่อสร้าง Knowledge Graph ทางการแพทย์ที่แสดงความสัมพันธ์ของ โรค, ยีนส์ และ ยารักษาโรค และแสดงตัวอย่างการใช้งาน เช่นความสัมพันธ์ของโรคที่สามารถวิเคราะห์ได้ว่าเป็นกลุ่มเดียวกัน หรือมีโอกาศมีความสัมพันธ์กัน พร้อมการนำ Graph database ไปใช้"



Proposes

“there are three objectives to achieve which state below.”



First Objective

เพื่อสร้าง Knowledge Graph ทางการแพทย์ที่แสดงความสัมพันธ์ของ โรค , ยีนส์ และ ยารักษาโรค

Second Objective

เพื่อหาคุณสมบัติที่สำคัญของกราฟจริงทางการแพทย์และวิเคราะห์ หาคุณสมบัติของ และวิเคราะห์ความสัมพันธ์ด้วยคุณสมบัติของกราฟ

Third Objective

เพื่อแสดงให้เห็นถึงกรณีที่สามารถนำเอา Knowledge Graph ที่สร้างขึ้น มาใช้ประโยชน์โดยการสร้างตัวแบบสำหรับการทำนายว่ายารักษาโรคชนิดใหม่ขึ้นเพื่อรักษาโรคหนึ่งๆ จะสามารถนำเอายานั้นไปรักษาโรคอื่นใดได้บ้างด้วยเทคนิค Link Prediction

Scope

93,710 Drugs

20,498 Diseases

19,033 Genes

MalaCards
HUMAN DISEASE DATABASE

Scraping the data from website: https://www.malacards.org

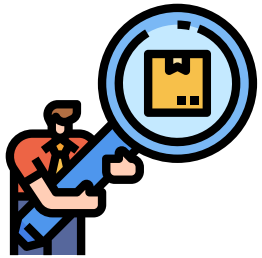
Usages

Case I

Provide some early diagnosis of the diseases (Within same Louvain Community)

Case II

Provide possibility that selected drug can cure disease.(Link Prediction)



Case III

Can query data for find information from the database

Methods

What do I need to get more knowledge on?

1. Data Collection
ทำการรวบรวมข้อมูลจากเวปไซค์โดยการ scraping ข้อมูลตาม scope ที่กำหนด

2. Data Preparation
ทำการจัดการกับข้อมูลที่มีการซ้ำซ้อนและรูปแบบที่ไม่เหมาะกับการใช้งาน เช่น มีการมีหลายข้อมูลในช่องเดียว

3. Create Knowledge Graph
นำข้อมูลที่อยู่ในรูปแบบ CSV File ตามแบบที่ต้องการเข้าเป็น Node และ Edge ใน NEO4J

4. Analysis by Graph Algorithm
-Centrality Detection
-Community Detection
-Similarity Detection

5. กรณีตัวอย่างการใช้งาน:
การหาความสัมพันธ์ของโรคกับยารักษาโรค ด้วยเทคนิค Link Prediction จาก Knowledge Graph ที่สร้างขึ้น

6. กรณีตัวอย่างการใช้งาน:
การสร้างส่วนต่อประสานผู้ใช้เพื่อสอบถามข้อมูลจาก Knowledge Graph ที่สร้างขึ้น

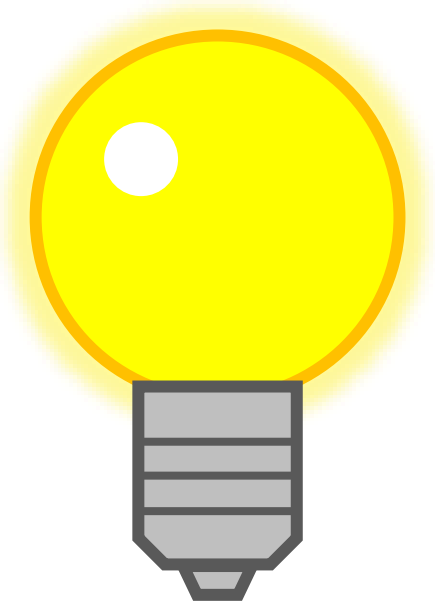
Theory

Centrality Detection

การคำนวณคุณสมบัติความเป็นจุดศูนย์กลาง ทั้งนี้จะใช้ Degree centrality คือการวัดความเป็นศูนย์กลางจาก link ที่มีต่อ node ทั้งหมด

Community Detection

การพิจารณาว่าแต่ละ Node ใน Network ควรจัดกลุ่มของ Node อย่างไร ทั้งนี้ในงานวิจัยนี้จะใช้ Louvain Modularity เป็น Algorithm ในการคำนวณ ซึ่งมีความรวดเร็ว และสามารถจัดกลุ่มแบบ Hierarchy ได้ด้วย



Similarity Detection

การวัดความเหมือนกันของ Node 2 Node ทั้งนี้ จะใช้ ANN Algorithm ซึ่งวัดความเหมือนกันได้รวดเร็ว

Link Prediction/Node2Vec

การแปลงคุณสมบัติของแต่ละ Node จากรูปแบบ Graph Database ให้อยู่ในรูป Vector เพื่อใช้เข้า Machine Learning ต่างๆ

Theory Details

Centrality Detection

PageRank
Betweenness Centrality
Degree Centrality
Closeness Centrality
Eigenvector Centrality

Community Detection

Louvain
Label Propagation
WCC
SCC
Triangular count
Local Clustering Coefficient

Similarity Detection

Node Similarity
K-Nearest Neighbors
Jaccard Similarity
Cosine Similarity
Pearson Similarity
Approximate Nearest Neighbors(ANN)

Link Prediction/Node2Vec

Node2vec
(from node2vec python library)

Tools

Python

- Scarpe data
- Machine Learning Algorithm

NEO4J

- Graph Database
- Community Detection
- Degree Centrality
- Similarity Detection

Flask

- Web framework for interactive query

Diseases Category

Global Category

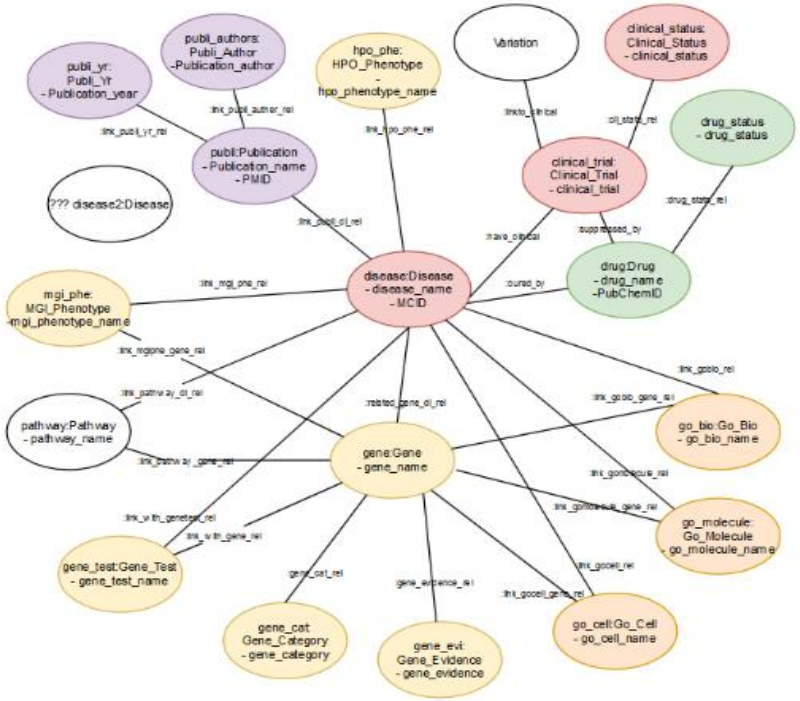
Category	Count
Cancer diseases	3256
Fetal diseases	5254
Genetic diseases	8069
Infectious diseases	678
Metabolic diseases	2512
Rare diseases	14365

Anatomical Category

Category	Count
Blood diseases	2417
Bone diseases	3104
Cardiovascular diseases	1870
Ear diseases	1521
Endocrine diseases	1620
Eye diseases	3190
Gastrointestinal diseases	1662
Immune diseases	2016
Liver diseases	869

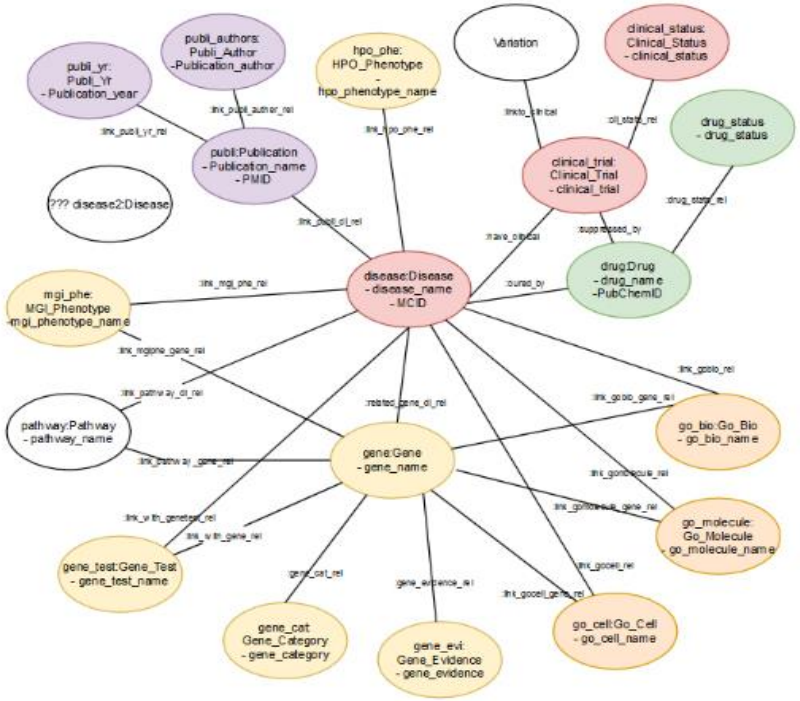
Category	Count
Mental diseases	1804
Muscle diseases	951
Nephrological diseases	1899
Neuronal diseases	7329
Oral diseases	734
Reproductive diseases	1390
Respiratory diseases	1357
Skin diseases	2737
Smell/Taste diseases	287

Data Model and Component Properties



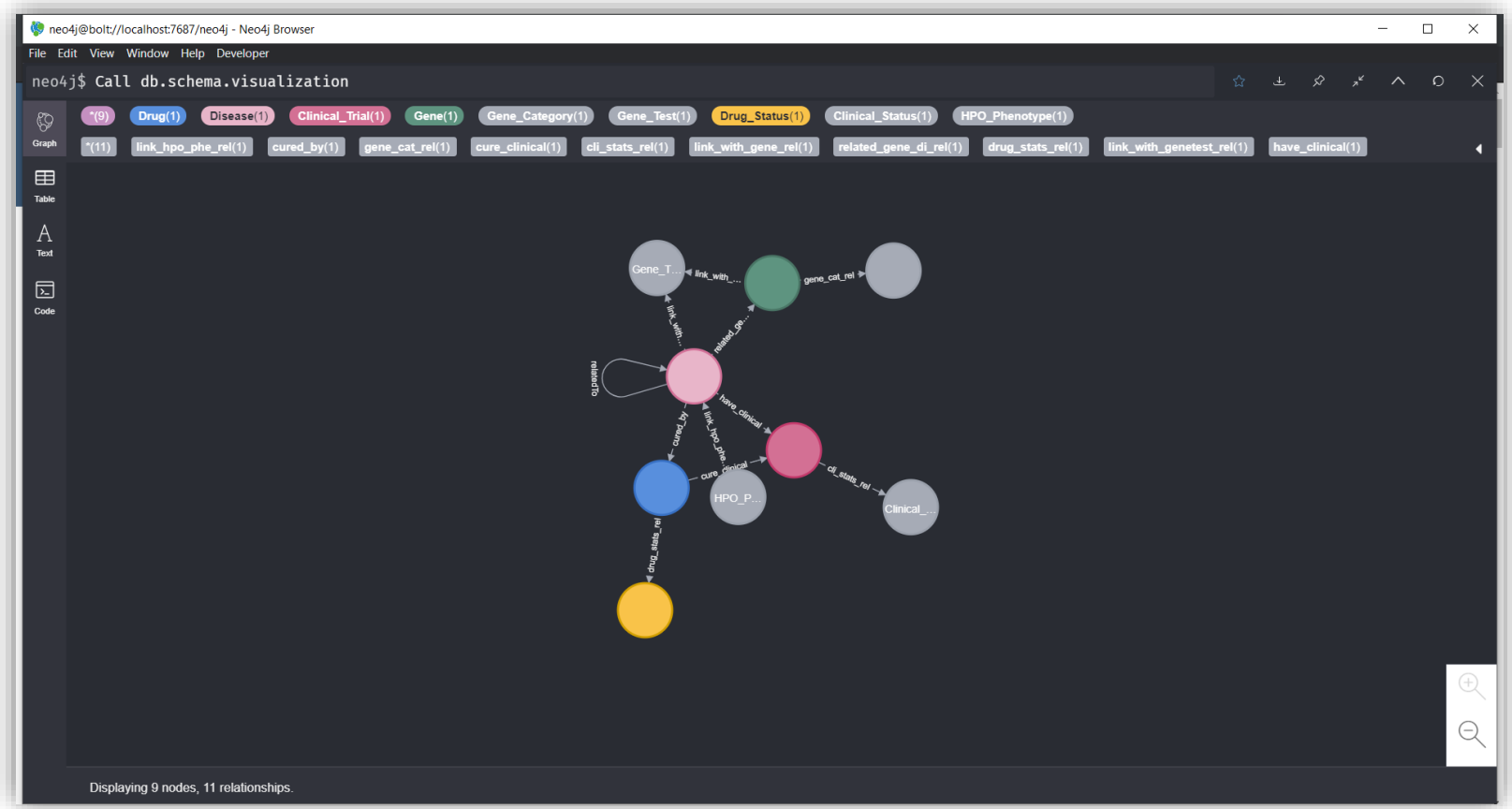
Node Category	Node Property
Drug	drug_name
Disease	community_louv_digene
Disease	disease_name
Disease	category
Disease	MCID
Disease	sub_category
Clinical_Trial	clinical_trial_name
Gene_Category	gene_category
Gene_Test	gene_test_name
Drug_Status	drug_status
Clinical_Status	clinical_status
Gene	gene_name
Gene	community_louv_digene
Gene	gene_synbol
HPO_Phenotype	hpo_phenotype_name

Data Model and Component Properties



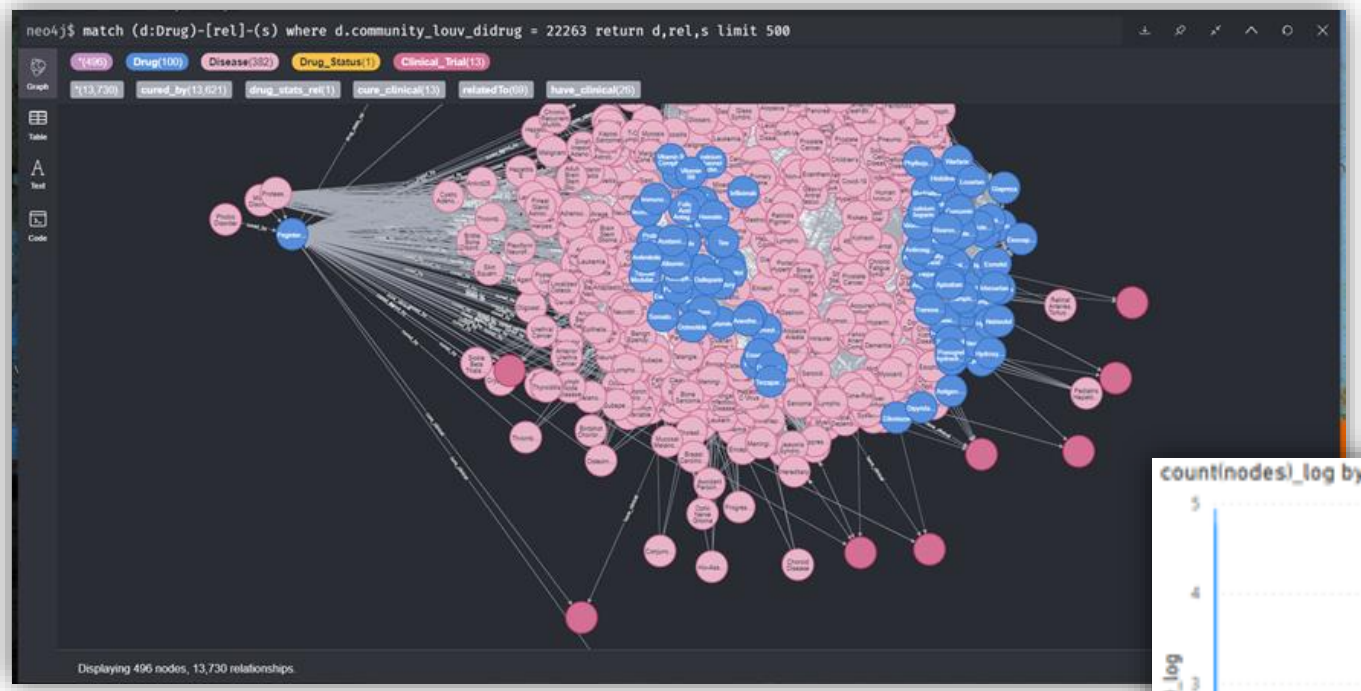
Relationship Type	Start Node	Destination Node
drug_stats_rel	Drug	Drug_Status
cure_clinical	Drug	Clinical_Trial
relatedTo	Disease	Disease
have_clinical	Disease	Clinical_Trial
related_gene_di_rel	Disease	Gene
cured_by	Disease	Drug
gene_cat_rel	Gene	Gene_Category
cli_stats_rel	Clinical_Trial	Clinical_Status

EDA

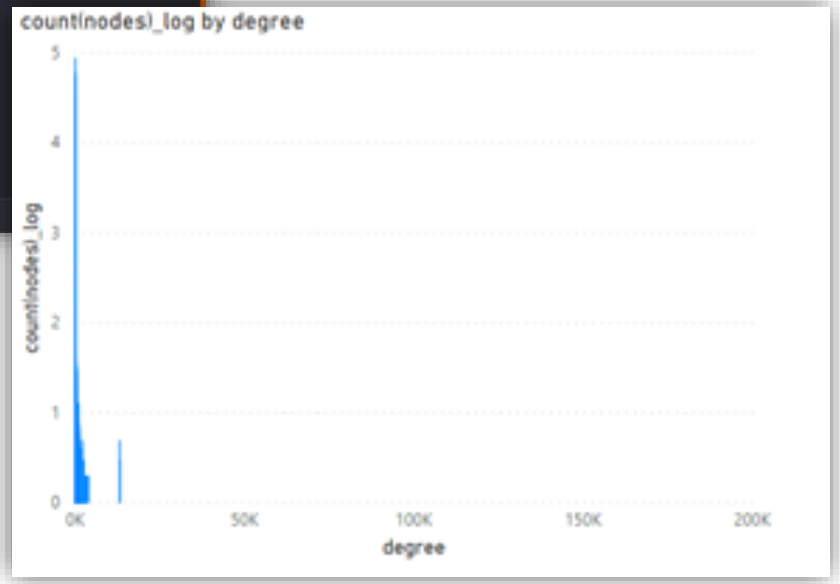


Degree distribution	Min	Max	Avg Interactions	STDEV	Median	P25	P75
Disease-Gene	1	1015	14.848	34.833	7	1	16
Disease-relatedTo (Di)	1	17380	182.09	562.0515	42	8	169
Disease-cured_by (Drug)	1	1692	93.957	172.27	28	8	101
Disease-have_clinical	1	9979	207.68	922.65	9	2	59
Drug- cure_clinical	1	329	1.424	3.054	1	1	1

EDA



Graph density
3.51185532118369e-5



Result Overview

Community Detection (Louvain Modularity)

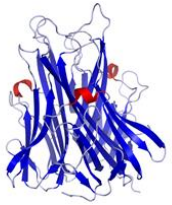
(Gene) กลุ่มโรคที่มียีนส์เกี่ยวข้อง ใกล้เคียงกัน คือกลุ่มโรคใด เพื่อตรวจเช็คความน่าจะเป็นที่
จะเกิดโรคอื่นๆ จากการตรวจพบโรคหนึ่งๆ

Similarity Detection (ANN)

- (Gene) พิจารณาโรคที่เกิดจาก ยีนส์เดียวกัน เพื่อตรวจพบและรักษาโรคที่มีโอกาศเกิดคู่กัน
โดยเร็ว
- (Drug) พิจารณายาที่สามารถช่วยกันรักษาโรคหนึ่งๆ

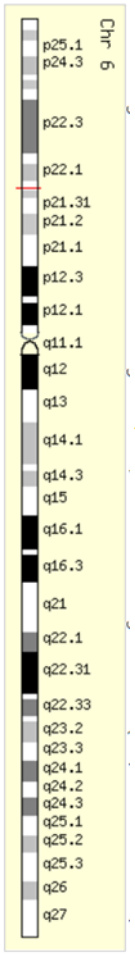
Link Prediction (ML + Node2Vec)

- ทำนายว่า ยา นั้นๆมีโอกาศรักษาโรคอะไรได้บ้าง



Result: Degree Centrality

Degree Centrality: Gene



Gene	No of Edge
TNF	810
IL6	730
TP53	645
CRP	638
ALB	590
CD40LG	569
IL1B	499

Degree Centrality: Drug

Drug	No of Edge
Anti-Infective Agents	2563
Immunologic Factors	2389
Hormones	2302
Anti-Bacterial Agents	2276
Pharmaceutical Solutions	2187
Anti-Inflammatory Agents	2063
Antirheumatic Agents	2060

Anti infective Agent : Substance capable of inhibiting the spread of an infectious organism or by killing the infectious organism outright.

Result: Centrality Detection (Disease)

Algorithm : Page Rank

Name	Page Rank	Interactions
Breast Cancer	69.71511	1015
Retinitis Pigmentosa	62.17727	928
Colorectal Cancer	57.82477	911
Lung Cancer	55.5435	871
Hepatocellular Carcinoma	51.63361	766
Prostate Cancer	48.25554	749
Ovarian Cancer	42.64103	716
Microcephaly	37.23761	469
Alzheimer Disease	35.05159	499

- การศึกษา ตัวอย่างโรคที่เกี่ยวข้องกับ Gene หลายตัวมากๆ จาก Page Rank
- จากงานวิจัย โรค Breast cancer link กับ gene จำนวนมาก
- Cancer : Mutation from tumor suppressor gene and onco gene ซึ่งทำให้การ express ของ gene ลดลง หรือเพิ่มขึ้น จึงเกี่ยวกับ gene เยอะมาก

Result: Centrality Detection (Gene)

Algorithm : Betweenness Centrality

การศึกษา ตัวอย่างโรคที่เกิดจาก Gene หลายตัวมากๆ จาก Betweenness Centrality

Gene	Score
H2AC18	12727105
TNF	11594185
TP53	10841072
IL6	9180900
ALB	8571700
SERPINA3	8286887
CD40LG	7578945

Algorithm : Page Rank

การศึกษา ตัวอย่างโรคที่เกี่ยวข้องกับ Gene หลายตัวมากๆ จาก Page Rank

Gene	Score
TNF	40.98261
IL6	35.70044
TP53	34.98614
CRP	34.32374
CD40LG	31.95646
ALB	31.0629
SERPINA3	26.52023
CD4	26.39531
H2AC18	25.9476
IL1B	24.68423
KRT7	22.93184
IL10	22.29482
VEGFA	22.09135

Result: Community Detection (Gene)

Algorithm : Louvain Modularity

ตัวอย่างของ โรคในกลุ่ม Louvain community ID 2152 (21 members)

Disease in Louvain Community: 2152
Brain Angioma
Cavernous Malformation
Cerebral Angioma
Cerebral Cavernous Malformation, Familial
Cerebral Cavernous Malformations
Cerebral Cavernous Malformations 2
Cerebral Cavernous Malformations 3
Cerebrocostomandibular Syndrome
Cervical Keratinizing Squamous Cell Carcinoma
Cervical Non-Keratinizing Squamous Cell Carcinoma
Encephalopathy, Familial, with Neuroserpin Inclusion Bodies
Glomuvenous Malformations
Gum Cancer
Hemangioma of Liver
Intracranial Cavernous Angioma
Intracranial Structure Hemangioma
Klippel-Trenaunay-Weber Syndrome
Monieziasis
Paralytic Lagophthalmos
Venous Malformations, Multiple Cutaneous and Mucosal
Viral Gastritis

Result: Community Detection (Drug)

Algorithm : Louvain Modularity

Community Code	Amount
22135	2361
22263	742
97579	253
126598	92
12282	1
14994	1
57993	1
62091	1

Community Code	Amount
51852	1
49805	1
55950	1
53903	1
12947	1
37511	1
131733	1
27288	1

“One drug cures only one diseases”

Result: Similarity Detection (Gene)

Algorithm : ANN (Approximate Nearest Neighbors)

From	To	Similarity
Melanoma-Pancreatic Cancer Syndrome	Endometrial Transitional Cell Carcinoma	1
Craniosynostosis 7	20p12.3 Microdeletion Syndrome	0.5
Mandibuloacral Dysplasia with Type a Lipodystrophy	Lmna-Related Dilated Cardiomyopathy	0.5
Coenzyme Q10 Deficiency Disease	Coenzyme Q10 Deficiency, Primary, 3	0.5
Autosomal Recessive Nonsyndromic Deafness 3	Deafness, Autosomal Recessive 7	0.45
Mucinous Intrahepatic Cholangiocarcinoma	Bile Duct Cystadenocarcinoma	0.416667
Cutis Laxa	Autosomal Recessive Type Ib Occipital Horn Syndrome	0.368421

Result: Similarity Detection (Drug)

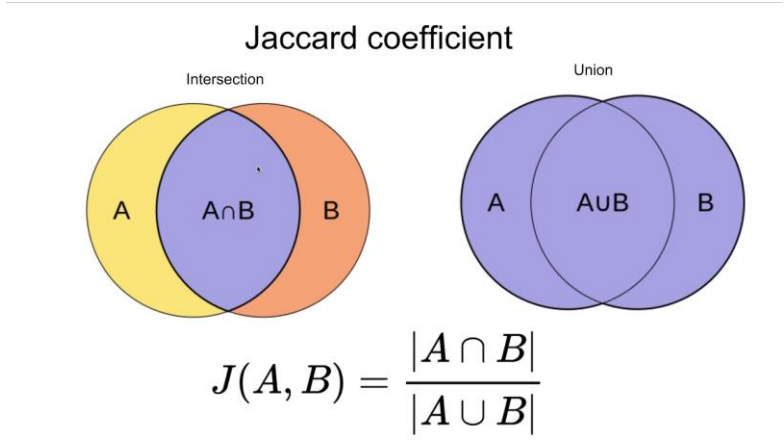
Algorithm : ANN (Approximate Nearest Neighbors)

Form	To	Similarity
Chloramphenicol succinate	Carbenicillin	1
Sulfisoxazole	Dalfopristin	0.555556
Etoposide	Carboplatin	0.516684
Doxorubicin	Cyclophosphamide	0.470356
policosanol	Trandolapril	0.379747
KRN 5500	Diprenorphine	0.375
Cyclosporins	Busulfan	0.372685

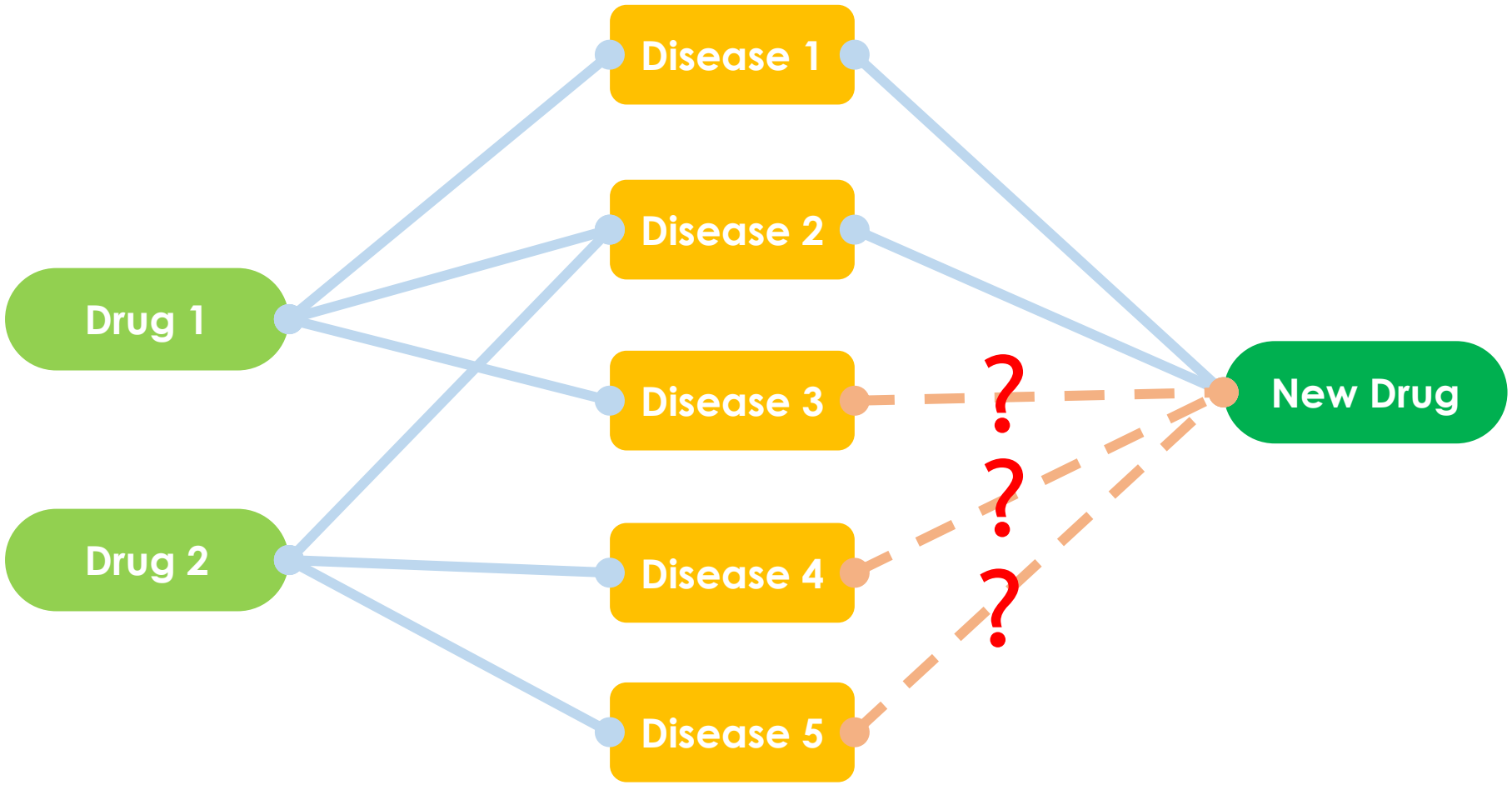
Result: Similarity Detection (Gene2)

Algorithm : Jaccard

Disease	Disease2	Similarity
Ovary Epithelial Cancer	Malignant Ovarian Surface Epithelial-Stromal Neoplasm	1
Peripheral Nervous System Benign Neoplasm	Autonomic Nervous System Benign Neoplasm	1
Cerebrum Cancer	Cerebral Ventricle Cancer	1
Cervix Uteri Carcinoma in Situ	Uterus Carcinoma in Situ	1
Chronic Inflammation of Lacrimal Passage	Dacryocystocele	1
Cerebral Hemisphere Lipoma	Corpus Callosum Lipoma	1
Autonomic Nervous System Benign Neoplasm	Peripheral Nervous System Benign Neoplasm	1
Dacryocystocele	Chronic Inflammation of Lacrimal Passage	1
Corpus Callosum Lipoma	Cerebral Hemisphere Lipoma	1
Cerebral Ventricle Cancer	Cerebrum Cancer	1

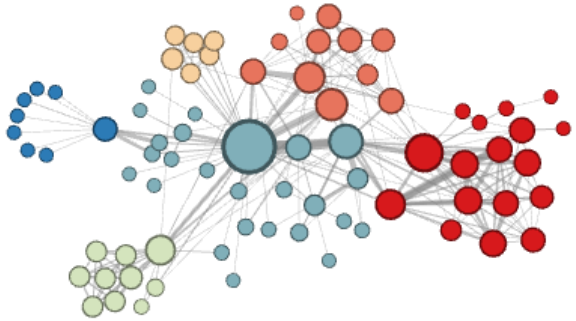


Link Prediction Overview



http://education.abcom.com/link-prediction-using-node2vec/

Link Prediction Overview

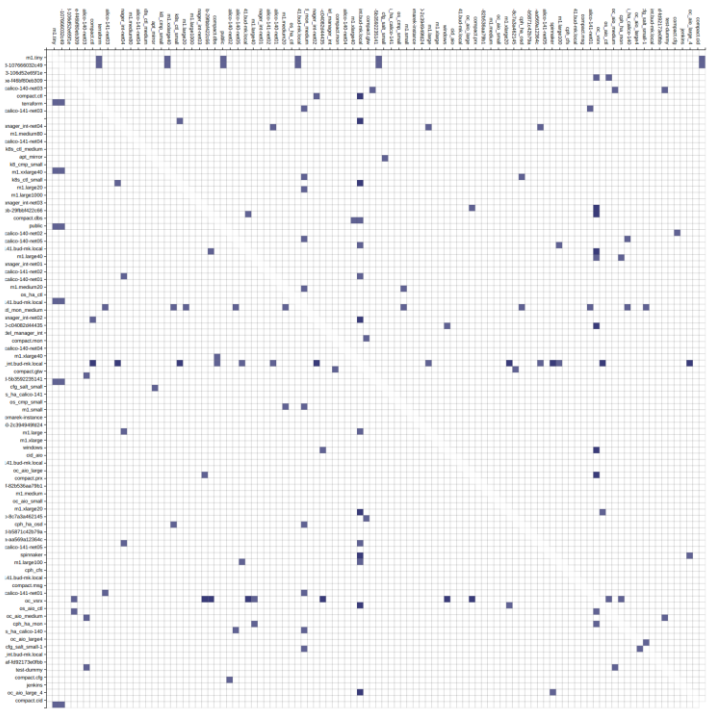


node2vec

xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	...	xxx
1	2	3	4	5	6	7	8	...	100

Link Prediction Overview

Adjacency Matrix



Label

Link Prediction Overview

Random Forest

k-nearest neighbors

XG Boost

Ada Boost

Logistic Regression

Decision Tree

Light GBM

Naive Bayes

https://medium.com/@hjhuney/implementing-a-random-forest-classification-model-in-python-583891c99652

Link Prediction Result (1)

10-Fold Cross Validation Score

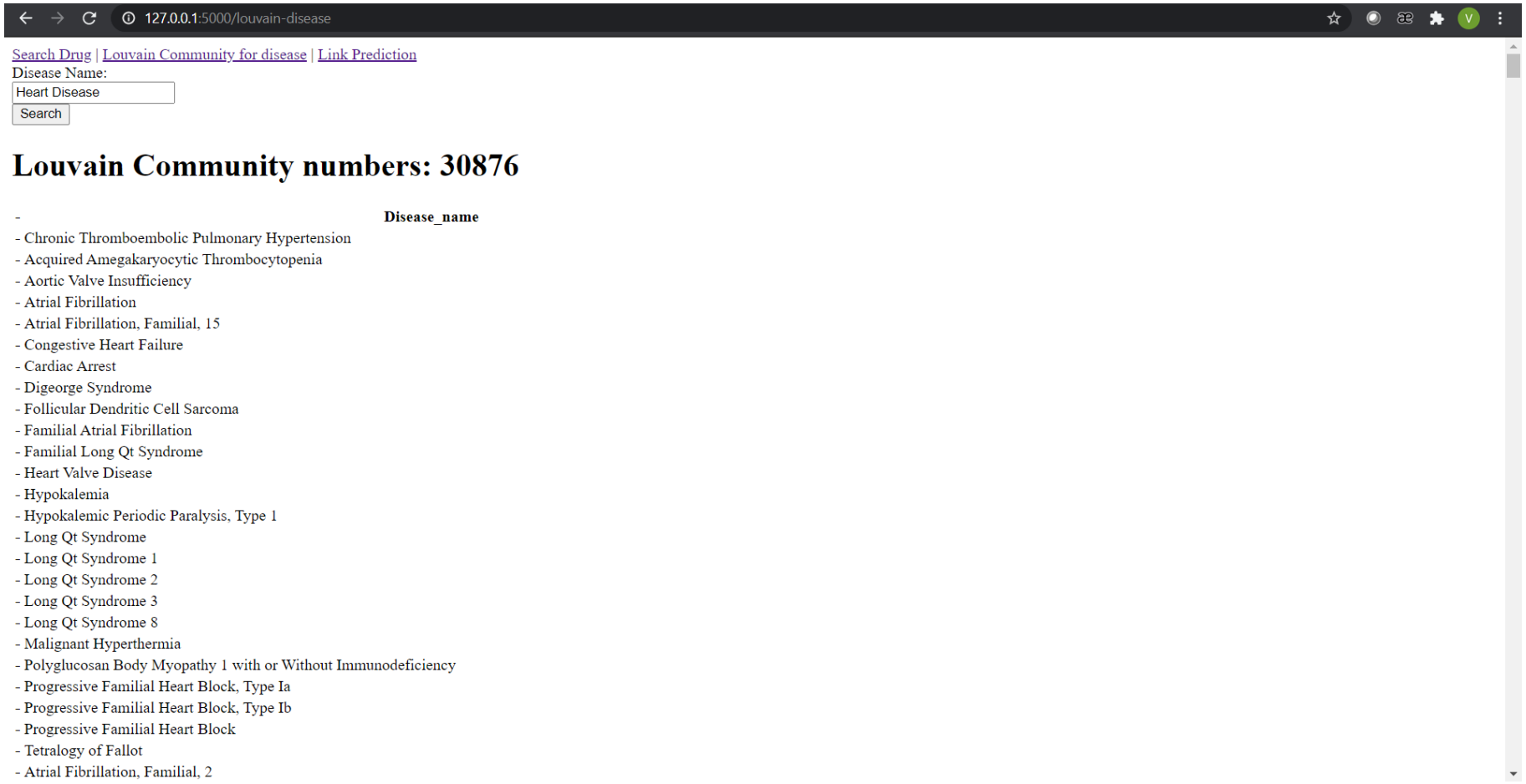
MODEL	Accuracy		Average Precision		Average Recall		Average F1	
	Average	SD	Class 0	Class 1	Class 0	Class 1	Class 0	Class 1
k-NN	0.9944	0.0003	0.9970	0.6333	0.9974	0.5943	0.9972	0.6132
Naive Bayes	0.8987	0.0011	0.9968	0.0443	0.9008	0.6145	0.9464	0.0826
Logistic Regression	0.9914	0.0001	0.9926	0.0436	0.9988	0.0076	0.9957	0.0129
LightGBM	0.9940	0.0001	0.9956	0.6552	0.9984	0.4056	0.9970	0.5011
XGBoost	0.9955	0.0001	0.9961	0.8524	0.9994	0.4755	0.9977	0.6105
Random Forest	0.9963	0.0001	0.9964	0.9796	0.9999	0.5185	0.9982	0.6781
Ada Boost	0.9908	0.0002	0.9930	0.1590	0.9978	0.0568	0.9954	0.0837
Decision Tree	0.9950	0.0002	0.9970	0.6842	0.9980	0.5944	0.9975	0.6362

Link Prediction Result (2)

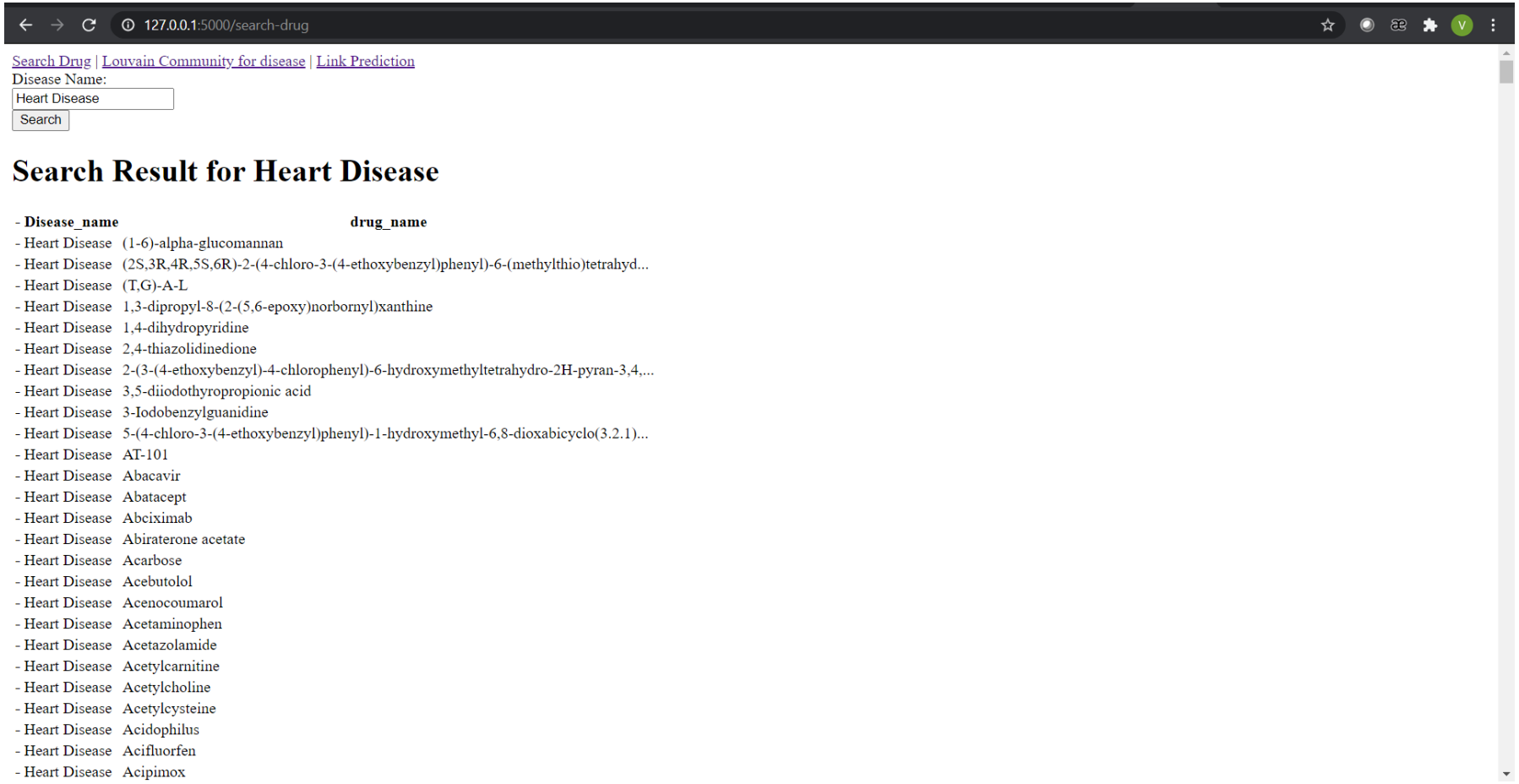
Test Set Score

Model	Accuracy	Precision		Recall		F1	
		Class 0	Class 1	Class 0	Class 1	Class 0	Class 1
k-NN	0.9944	0.9969	0.6318	0.9974	0.5885	0.9972	0.6094
Naive Bayes	0.8982	0.9967	0.0434	0.9004	0.6046	0.9461	0.0810
Logistic Regression	0.9915	0.9926	0.0381	0.9988	0.0062	0.9957	0.0106
LightGBM	0.9940	0.9955	0.6600	0.9985	0.3965	0.9970	0.4954
XGBoost	0.9956	0.9962	0.8624	0.9994	0.4846	0.9978	0.6205
Random Forest	0.9964	0.9965	0.9851	0.9999	0.5275	0.9982	0.6871
Ada Boost	0.9907	0.9929	0.1304	0.9978	0.0442	0.9953	0.0660
Decision Tree	0.9953	0.9970	0.7236	0.9983	0.5936	0.9976	0.6522

DEMO


127.0.0.1:5000/louvain-disease
Search Drug | Louvain Community for disease | Link Prediction
Disease Name:
Heart Disease
Search
Louvain Community numbers: 30876
-
Disease_name
- Chronic Thromboembolic Pulmonary Hypertension
- Acquired Amegakaryocytic Thrombocytopenia
- Aortic Valve Insufficiency
- Atrial Fibrillation
- Atrial Fibrillation, Familial, 15
- Congestive Heart Failure
- Cardiac Arrest
- Digeorge Syndrome
- Follicular Dendritic Cell Sarcoma
- Familial Atrial Fibrillation
- Familial Long Qt Syndrome
- Heart Valve Disease
- Hypokalemia
- Hypokalemic Periodic Paralysis, Type 1
- Long Qt Syndrome
- Long Qt Syndrome 1
- Long Qt Syndrome 2
- Long Qt Syndrome 3
- Long Qt Syndrome 8
- Malignant Hyperthermia
- Polyglucosan Body Myopathy 1 with or Without Immunodeficiency
- Progressive Familial Heart Block, Type Ia
- Progressive Familial Heart Block, Type Ib
- Progressive Familial Heart Block
- Tetralogy of Fallot
- Atrial Fibrillation, Familial, 2

DEMO


127.0.0.1:5000/search-drug
Search Drug | Louvain Community for disease | Link Prediction
Disease Name:
Heart Disease
Search
Search Result for Heart Disease
- Disease_name drug_name
- Heart Disease (1-6)-alpha-glucomannan
- Heart Disease (2S,3R,4R,5S,6R)-2-(4-chloro-3-(4-ethoxybenzyl)phenyl)-6-(methylthio)tetrahyd...
- Heart Disease (T,G)-A-L
- Heart Disease 1,3-dipropyl-8-(2-(5,6-epoxy)norbornyl)xanthine
- Heart Disease 1,4-dihydropyridine
- Heart Disease 2,4-thiazolidinedione
- Heart Disease 2-(3-(4-ethoxybenzyl)-4-chlorophenyl)-6-hydroxymethyltetrahydro-2H-pyran-3,4,...
- Heart Disease 3,5-diiodothyropropionic acid
- Heart Disease 3-Iodobenzylguanidine
- Heart Disease 5-(4-chloro-3-(4-ethoxybenzyl)phenyl)-1-hydroxymethyl-6,8-dioxabicyclo(3.2.1)...
- Heart Disease AT-101
- Heart Disease Abacavir
- Heart Disease Abatacept
- Heart Disease Abciximab
- Heart Disease Abiraterone acetate
- Heart Disease Acarbose
- Heart Disease Acebutolol
- Heart Disease Acenocoumarol
- Heart Disease Acetaminophen
- Heart Disease Acetazolamide
- Heart Disease Acetylcarnitine
- Heart Disease Acetylcholine
- Heart Disease Acetylcysteine
- Heart Disease Acidophilus
- Heart Disease Acifluorfen
- Heart Disease Acipimox

References

[1] David N. Nicholson, Casey S. Greene, "Constructing knowledge graphs and their biomedical applications" , Computational and Structural Biotechnology Journal 18 (2020).

[2] Yong Zhang, Ming Sheng, Ruizhong, etc., "HKGB: An Inclusive, Extensible, Itelligent, Semi-auto-constructed Knowledge Graph Framework for Healthcare with Clinicians's expertise Incorporate"

[3] Lisa Ehrlinger, Wolfram Wor, "Towards a Definition of Knowledge Graphs"

[4] Li Tian, Weinan Zhang, Haofen Wang, etc., "MeDetect: Domain Entity Annotation in Biomedical References Using Linked Open Data" , ISWC 2012

[5] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, etc.,(2019) "BioBERT: a pre-trained biomedical language representation model for biomedical text mining" , Bioinformatics OXFORD

[6] Ruijie Wang, Yuchen Yan, Jialu Wang, etc.,"AceKG: A Large-scale Knowledge Graph for Academic Data Mining", arXiv, 2018

[7] Longxiang Shi, Shijian Li, Xiaoran Yang, etc., "Semantic Health Knowledge Graph: Semantic Integration of Heterogenous Medical Knowledge Services." , Hindawi BioMed Research Interational.

[8] Maya Rotmensch, Yoni Halperm, Abdulhakim Tlimat, etc., "Learning a Health Knowledge Graph from Electronic Medical Records.", scientific reports.

[9] "How to use knowledge graph for precision medicine" – Syed Irtaza Raza

[10] "Retail Graph – Walmart's Product Knowledge Graph" – Karthik Deivasigamani

[11] DoctorFinder! By GraphGist

[12] Georgios A Pavlopoulos, Maria Secrier, Charalampos N Moschopoulos, etc.. "Using graph theory to analyze biological networks", BioData Mining, 2011

[13] Esra Gundogan, Buket Kaya ., "A Link Prediction Appproach for Drug Recommendation in Disease-Drug Bipartite Network", IEEE, 2017

Thank You