Lucas Derr
Dr. Zamani
CSCI 4022
6 May 2023

Wikipedia Search Engine and Recommendation System

Introduction

Networks are all around us. The internet, social media, and even the economy are often most easily represented as a graph with nodes and edges. One such network that has become almost synonymous with 'Googling' is Wikipedia, a free, open-source internet encyclopedia. Wikipedia has over 6.6 million articles written in English, and each article contains numerous hyperlinks to other articles. In order to make this data useful, it must be navigated effectively and efficiently, filtering out the unimportant.

In this project, I conduct network techniques such as PageRank and Louvain clustering to create a Python search engine and recommendation system. I build a Python application that runs these algorithms. In the application, there is a search bar similar to Google, in which the user can view the output of the PageRank algorithm in real-time. In addition, I will utilize the Wikipedia API for keyword searches and to pull article abstracts when requested by the user. I have also utilized the PageRank algorithm in real time to create a recommendation engine, which will be detailed later in this report. Overall, this is a project that is completed from start to finish and I am happy with the end result.

Data

The data for this project was downloaded from the wikiVitals level 4 dataset. This dataset contains 10,000 'vital' articles. Vital articles are classified by various criteria with the goal of generating a graph that is representative of the entirety of the Wikipedia database. The network structure is derived by each pages' hyperlinks: links are treated as outgoing connections to pages. Each pages' incoming and outgoing page links are different, making the graph directed. The adjacency matrix which represents the entire network is stored in a scipy.sparse matrix, allowing for decreased overall computation time. In addition, the dataset is cleaned and labeled making the project more about applying the PageRank algorithm instead of generating the data.

Real-World Impact

Before Google, the World Wide Web was a mess of 10s of millions of web pages. There was no structure in place to traverse, rank, or make sense of this vast network. This all changed with the 1998 paper, *The Anatomy of a Large-Scale Hypertextual Web Search Engines*, in which Larry Page and Sergey Brin, two Stanford Ph.D. students, detailed Google, a web-crawling search engine. In just two years, Google became the most comprehensive search engine on the market. In just a few more, Google revolutionized the access that the world has to information on the

internet. In this project, I will play the role of Larry Page and Sergey Brin in 1999, and build a simple search engine on the dataset that closely resembles the structure of the internet: Wikipedia.

Methods

Before any search is made in the application, three things happen: PageRanks are calculated for each node in the dataset, Hubs scores are calculated for each node, and labels are generated from the best Louvain clustering model. The outputs of these models are stored in memory and can easily be accessed when necessary. All algorithms come from the Scikit-Network Python library.

$$W_i = (1 - d) + d \sum_{i=1}^{N} \frac{W_i}{n_i}$$

Figure 1: PageRank Equation

The original PageRank formula is shown above. In the formula, W_i is the PageRank of node i, d is the damping factor, N is the total number of nodes, and n_i is the number of incoming links for the current node. The rationale behind the formula is that incoming links are 'votes', and each of these votes is weighted by the PageRank of that incoming node. In addition, the formula includes a damping factor, d, which prevents spider traps and dead ends from aborbing the PageRank score.

In the PageRank model used for this search engine, the damping factor is set to 0.85 and the page ranks are calculated with power iteration, an iterative method for computing dominant eigenvalues.

$$(1)\ Hub(V_i) = \sum_{V_j \in Out(V_i)} e_{i,j} Authority(V_j)$$

$$(2)\ Authority(V_i) = \sum_{V_j \in Out(V_i)} e_{i,j} Hub(V_j)$$

Figure 2: Hyperlink-Induced Topic Search (HITS) Equations

In addition to PageRank, the search engine will calculate the Hubs score for each node to be displayed on the search results screen. The idea behind the algorithm are that in many networks, there are hubs, which have many important outgoing links, and authorities, which have many important incoming links from hubs. In essence, the Hub score is the importance measure of the outgoing links, while the Authority score is the importance measure of the incoming links. The next algorithm which is run before the first query is a Louvain clustering model.

$$Q_c = \frac{\Sigma_{in}}{2m} - \left(\frac{\Sigma_{tot}}{2m}\right)^2$$

Figure 3: Louvain Clustering Modularity Formula

For the equation in *Figure 3*, Q_c defines the modularity of cluster c, Σ_{in} is the sum of edge weights between nodes that are in community c, Σ_{out} is the number of edges within nodes in community c to nodes in other communities, and m is the total number of edges in the network. The Louvain algorithm aims at maximizing this modularity value. A simplified version of the algorithm is shown in *Figure 4*:

1. Assign each node *i* to its own community
2. Repeat until convergence:
    a. For each node *i:*
        i. Calculate the change in modularity for removing *i* from its current community and moving it into a neighboring community (if there is one)
        ii. If applicable, place *i* into the community with the greatest increase in modularity

Figure 4: Louvain Pseudocode

The Louvain algorithm is hierarchical. It begins with each node assigned to its own clusters and combines these clusters until convergence. For this reason, there is no parameter for specifying the number of clusters.

The two hyperparameters which were tuned to maximize the total modularity of the network were: *modularity*, the equation used to measure individual modularity scores, and *resolution*, a tuning parameter used in the modularity calculation. The results of the hyperparameter tuning are depicted in *Figure 5:*

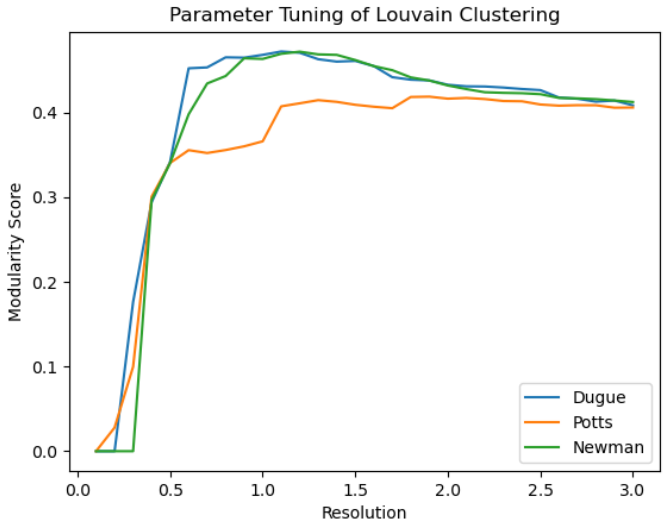

Figure 5: Hyperparameter Tuning for Louvain Clustering

The final model uses Dugue modularity, a resolution of 1.01, achieving a modularity score of 0.472 by dividing the network into 12 clusters.

After a query is entered into the search engine, there are two things that happen: First, A Wikipedia srsearch API call is made to generate candidate articles for the search query (the search profile is set to "empty" so as not to receive assistance from Wikipedia's built-in search engine). Then, the candidates are sorted by their PageRank values and displayed on the screen in descending order.

The recommendation engine is built by running a weighted PageRank using the output of the Louvain model. The process is as follows:

1. *Identify the cluster label of the article to be recommended*
2. *Run a weighted PageRank on the entire network*
    a. *The article to be recommended is weighted as 3.*
    b. *Articles in the same cluster are weighted as* $\frac{1}{n}$*, where n is the number of articles in the cluster.*
    c. *All other articles have a weight of 0*
3. *Subtract the original PageRank score from the weighted PageRank score*
4. *Return the top 20 scores based on this metric*

Figure 6: Detail About Recommendation Process

Interestingly, the output of PageRank does an excellent job as a recommendation algorithm. Initially, my thought process was to use the output of clustering for recommendation. The problem with this, however, is that the clusters are not ranked by importance or in relation to an article. This is where the PageRank is useful. Weighing the article to be recommended as much higher than any other article ensures that articles which are closely related to that article will receive higher PageRank scores. In addition, weighing articles in the same cluster as the article to be recommended will ensure that articles in the same cluster will receive higher rankings. Finally, subtracting the original PageRank from this new PageRank ensures that articles that generally have high PageRanks are not always selected in this process.

Results

For the results section, an examle use case of the Python application will be demonstrated, displaying the output of all functions.

```
Enter your Search Query:France
```

Index	Title	Link	Incoming Links	Outgoing Links	PageRank	Hubs
0	France	https://en.wikipedia.org/wiki/France	1390	566	0.00139114	0.0331523
1	French Revolution	https://en.wikipedia.org/wiki/French_Revolution	518	237	0.000530615	0.0128401
2	French colonial empire	https://en.wikipedia.org/wiki/French_colonial_empire	180	215	0.00020312	0.0176827
3	Kingdom of France	https://en.wikipedia.org/wiki/Kingdom_of_France	134	83	0.0001781	0.0222551
4	History of France	https://en.wikipedia.org/wiki/History_of_France	128	251	0.000136648	0.0313439
5	Francis I of France	https://en.wikipedia.org/wiki/Francis_I_of_France	84	67	0.000108773	0.0108316
6	French literature	https://en.wikipedia.org/wiki/French_literature	62	87	9.30368e-05	0.0192388
7	French cuisine	https://en.wikipedia.org/wiki/French_cuisine	79	102	8.83087e-05	0.00939699
8	Henry IV of France	https://en.wikipedia.org/wiki/Henry_IV_of_France	62	58	8.38889e-05	0.0149184
9	Philip II of France	https://en.wikipedia.org/wiki/Philip_II_of_France	45	42	7.15391e-05	0.00850807
10	Louis IX of France	https://en.wikipedia.org/wiki/Louis_IX_of_France	53	54	7.08006e-05	0.0123275
11	Philip IV of France	https://en.wikipedia.org/wiki/Philip_IV_of_France	42	54	6.20579e-05	0.0129898
12	Cinema of France	https://en.wikipedia.org/wiki/Cinema_of_France	41	53	5.84005e-05	0.0123652
13	Tour de France	https://en.wikipedia.org/wiki/Tour_de_France	35	40	5.76352e-05	0.0182061

```
Enter Index to get more information about an article or '-1' to enter another query:
```

Figure 7: Search Query Results

The results of a keyword search for ‘France’ are shown above. The Wikipedia API runs a keyword search to generate all candidates, which are the sorted by their PageRank values and displayed. When each article is initially weighted equally, the number of incoming links is almost always the directly correlated with the PageRank score.

```
Choose Here1
Henry IV (French: Henri IV; 13 December 1553 – 14 May 1610), also known by the epithets Good King Henry or Henry th
e Great, was King of Navarre (as Henry III) from 1572 and King of France from 1589 to 1610. He was the first monarc
h of France from the House of Bourbon, a cadet branch of the Capetian dynasty. He was assassinated in 1610 by Franç
ois Ravaillac, a Catholic zealot, and was succeeded by his son Louis XIII.
Henry was the son of Jeanne III of Navarre and Antoine de Bourbon, Duke of Vendôme. He was baptised as a Catholic b
ut raised in the Protestant faith by his mother. He inherited the throne of Navarre in 1572 on his mother's death.
As a Huguenot, Henry was involved in the French Wars of Religion, barely escaping assassination in the St. Bartholo
mew's Day massacre. He later led Protestant forces against the French royal army.
Henry became king of France in 1589 upon the death of Henry III, his brother-in-law and distant cousin. He was the
first French monarch from the House of Bourbon. Henry initially kept the Protestant faith (the only French king to
do so) and had to fight against the Catholic League, which denied that he could wear France's crown as a Protestant
. After four years of stalemate, he converted to Catholicism to obtain mastery over his kingdom (reportedly saying,
"Paris vaut bien une messe." "Paris is well worth a mass."). As a pragmatic politician (in the parlance of the time
, a politique), he promulgated the Edict of Nantes (1598), which guaranteed religious liberties to Protestants, the
reby effectively ending the French Wars of Religion.
An active ruler, Henry worked to regularise state finance, promote agriculture, eliminate corruption and encourage
education. During his reign, the French colonization of the Americas truly began with the foundation of the colonie
s of Acadia and Canada at Port-Royal and Quebec, respectively. He is celebrated in the popular song "Vive le roi He
nri" (which later became an anthem for the French monarchy during the reigns of his successors) and in Voltaire's H
enriade.
You have selected: Henry IV of France
What would you like to do?
1. Read Article Abstract
2. See Recommended Articles
3. See All Incoming Articles
4. See All Outgoing Articles
5. Enter Another Query
6. Exit Program

Choose Here
```

Figure 8: Fetch Article Abstract for ‘King Henry IV’

Selecting to read the article abstract runs another Wikipedia API call and prints the text onto the screen. In this case, the abstract for ‘King Henry IV’ is shown.

Index	Title	Link	Incoming Links	Outgoing Links	PageRank	Hubs
0	Napoleon	https://en.wikipedia.org/wiki/Napoleon	339	171	0.000375011	0.0250736
1	Philip V of Spain	https://en.wikipedia.org/wiki/Philip_V_of_Spain	52	37	6.69685e-05	0.0122528
2	Franks	https://en.wikipedia.org/wiki/Franks	160	107	0.000195445	0.0239944
3	Thirty Years' War	https://en.wikipedia.org/wiki/Thirty_Years'_War	263	197	0.000304685	0.0213022
4	Philip IV of France	https://en.wikipedia.org/wiki/Philip_IV_of_France	42	54	6.20579e-05	0.0129898
5	Maximilian I, Holy Roman Emperor	https://en.wikipedia.org/wiki/Maximilian_I,_Holy_Roman_Emperor	70	45	8.74419e-05	0.00640693
6	Philip II of Spain	https://en.wikipedia.org/wiki/Philip_II_of_Spain	141	119	0.000179176	0.0111436
7	Francia	https://en.wikipedia.org/wiki/Francia	130	203	0.000176701	0.0118352
8	Philip II of France	https://en.wikipedia.org/wiki/Philip_II_of_France	45	42	7.15391e-05	0.00850807
9	Rome	https://en.wikipedia.org/wiki/Rome	529	292	0.00055201	0.0260497
10	Charlemagne	https://en.wikipedia.org/wiki/Charlemagne	224	179	0.000240153	0.0128955
11	Roman Empire	https://en.wikipedia.org/wiki/Roman_Empire	813	298	0.000865319	0.0307861
12	Charles the Fat	https://en.wikipedia.org/wiki/Charles_the_Fat	25	57	4.12391e-05	0.00981635
13	Papal States	https://en.wikipedia.org/wiki/Papal_States	245	144	0.00028794	0.0120906
14	Capetian dynasty	https://en.wikipedia.org/wiki/Capetian_dynasty	23	36	4.24665e-05	0.0200845
15	Hundred Years' War	https://en.wikipedia.org/wiki/Hundred_Years'_War	115	92	0.000142514	0.00800322
16	Louis XVI	https://en.wikipedia.org/wiki/Louis_XVI	64	72	8.02408e-05	0.022744
17	Merovingian dynasty	https://en.wikipedia.org/wiki/Merovingian_dynasty	43	51	6.4924e-05	0.0162285
18	Rhine	https://en.wikipedia.org/wiki/Rhine	149	82	0.000187907	0.007573
19	Ferdinand II of Aragon	https://en.wikipedia.org/wiki/Ferdinand_II_of_Aragon	54	28	7.34365e-05	0.00895923

Figure 9: Article Recommendations for ‘King Henry IV’

The output of the recommendation algorithm for ‘King Henry IV’ is shown above. The results appear to be logical as they all relate to King Henry IV in some way.

Index	Title	Link	Incoming Links	Outgoing Links	PageRank	Hubs
0	France	https://en.wikipedia.org/wiki/France	1390	566	0.00139114	0.0331523
1	Catholic Church	https://en.wikipedia.org/wiki/Catholic_Church	728	369	0.000752	0.0214694
2	Paris	https://en.wikipedia.org/wiki/Paris	673	396	0.000663328	0.0227327
3	Protestantism	https://en.wikipedia.org/wiki/Protestantism	431	315	0.000467854	0.0167763
4	Napoleon	https://en.wikipedia.org/wiki/Napoleon	339	171	0.000375011	0.0250736
5	Florence	https://en.wikipedia.org/wiki/Florence	269	156	0.00030725	0.011767
6	Voltaire	https://en.wikipedia.org/wiki/Voltaire	293	247	0.0002803	0.00367012
7	Reformation	https://en.wikipedia.org/wiki/Reformation	246	233	0.00027053	0.0131572
8	Julius Caesar	https://en.wikipedia.org/wiki/Julius_Caesar	221	114	0.000261765	0.0164198
9	Charlemagne	https://en.wikipedia.org/wiki/Charlemagne	224	179	0.000240153	0.0128955
10	James VI and I	https://en.wikipedia.org/wiki/James_VI_and_I	218	165	0.000230373	0.0137491
11	Andorra	https://en.wikipedia.org/wiki/Andorra	156	192	0.000213539	0.00903437
12	French colonial empire	https://en.wikipedia.org/wiki/French_colonial_empire	180	215	0.00020312	0.0176827
13	Counter-Reformation	https://en.wikipedia.org/wiki/Counter-Reformation	183	185	0.00020299	0.0226043
14	Franks	https://en.wikipedia.org/wiki/Franks	160	107	0.000195445	0.0239944
15	Philip II of Spain	https://en.wikipedia.org/wiki/Philip_II_of_Spain	141	119	0.000179176	0.0111436
16	Kingdom of France	https://en.wikipedia.org/wiki/Kingdom_of_France	134	83	0.0001781	0.0222551
17	Carolingian Empire	https://en.wikipedia.org/wiki/Carolingian_Empire	155	203	0.000176561	0.0190058
18	Hugo Grotius	https://en.wikipedia.org/wiki/Hugo_Grotius	169	202	0.000171296	0.0140834
19	Thomas Müntzer	https://en.wikipedia.org/wiki/Thomas_Müntzer	154	184	0.000149534	0.00708167
20	Michelangelo	https://en.wikipedia.org/wiki/Michelangelo	117	47	0.000137562	0.0118177
21	History of France	https://en.wikipedia.org/wiki/History_of_France	128	251	0.000136648	0.0313439
22	Palace of Versailles	https://en.wikipedia.org/wiki/Palace_of_Versailles	94	64	0.000118324	0.0149336

Figure 10: Incoming Articles for King Henry IV, sorted by PageRank

View the incoming articles for ‘King Henry IV’.

Index	Title	Link	Incoming Links	Outgoing Links	PageRank	Hubs
0	France	https://en.wikipedia.org/wiki/France	1390	566	0.00139114	0.0331523
1	French language	https://en.wikipedia.org/wiki/French_language	697	138	0.0008667	0.0197914
2	Ottoman Empire	https://en.wikipedia.org/wiki/Ottoman_Empire	727	319	0.00071973	0.0242688
3	Paris	https://en.wikipedia.org/wiki/Paris	673	396	0.000663328	0.0227327
4	French Revolution	https://en.wikipedia.org/wiki/French_Revolution	518	237	0.000530615	0.0128401
5	Holy Roman Empire	https://en.wikipedia.org/wiki/Holy_Roman_Empire	418	231	0.000470681	0.0283214
6	Protestantism	https://en.wikipedia.org/wiki/Protestantism	431	315	0.000467854	0.0167763
7	Tunisia	https://en.wikipedia.org/wiki/Tunisia	379	231	0.000408253	0.0187022
8	Constantinople	https://en.wikipedia.org/wiki/Constantinople	341	86	0.000380754	0.0210072
9	Napoleon	https://en.wikipedia.org/wiki/Napoleon	339	171	0.000375011	0.0250736
10	Maldives	https://en.wikipedia.org/wiki/Maldives	314	233	0.000334618	0.0137638
11	Thirty Years' War	https://en.wikipedia.org/wiki/Thirty_Years'_War	263	197	0.000304685	0.0213022
12	Voltaire	https://en.wikipedia.org/wiki/Voltaire	293	247	0.0002803	0.00367012
13	Constitutional monarchy	https://en.wikipedia.org/wiki/Constitutional_monarchy	195	129	0.00026895	0.0106441
14	Charlemagne	https://en.wikipedia.org/wiki/Charlemagne	224	179	0.000240153	0.0128955
15	Charles V, Holy Roman Emperor	https://en.wikipedia.org/wiki/Charles_V,_Holy_Roman_Emperor	192	128	0.000217748	0.0316937
16	Euro	https://en.wikipedia.org/wiki/Euro	162	119	0.00020096	0.0118589
17	Bronze	https://en.wikipedia.org/wiki/Bronze	193	112	0.000195444	0.00524502
18	Philip II of Spain	https://en.wikipedia.org/wiki/Philip_II_of_Spain	141	119	0.000179176	0.0111436
19	Kingdom of France	https://en.wikipedia.org/wiki/Kingdom_of_France	134	83	0.0001781	0.0222551
20	Francia	https://en.wikipedia.org/wiki/Francia	130	203	0.000176701	0.0118352
21	Sumatra	https://en.wikipedia.org/wiki/Sumatra	161	70	0.000168466	0.01145
22	House of Habsburg	https://en.wikipedia.org/wiki/House_of_Habsburg	120	80	0.000152704	0.0119041

Figure 11: Outgoing Articles, sorted by PageRank

View the outgoing articles for ‘King Henry IV’.

Conclusion

Search engines and recommendation engines are immensely useful when users need a way to filter results from large databases, such as the internet. This project served as a way to understand the basic framework of a search engine and dive into the mathematical theory behind Google’s PageRank. In addition, using PageRank as a recommendation algorithm illustrates its utility as a ranking system in general.

This project was successful in its initial goal: To create a search engine. The recommendation algorithm was a stretch goal which stemmed from curiosity and reading various internet sources for use cases of PageRank.

If given more time to work on this project, I would expand the dataset and scale to a larger network. Using a prebuilt dataset made my life very easy and made the implementation of each algorithm a single line of code. Expanding the project to use the web crawlers in a similar way as in Google would make the project more realistic and applicable to a real-world scenario.