Slovene Wikipedia Network Analysis

Andrej Kronovšek[a,1], **Tim Vučina**[a,2], **and Šimen Ravnik**[a,3]

[a]University of Ljubljana, Faculty of Computer and Information Science, Večna pot 113, SI-1000 Ljubljana, Slovenia

The manuscript was compiled on April 29, 2022

Wikipedia is the largest free encyclopedia in the world, created by hundreds of thousands of individuals from all over the world. Wikipedia was originally created in English language, but was then quickly translated into other languages. In the year 2002, Slovene Wikipedia was created and since then around 176 000 articles were published. The Wikipedia's core idea is based on hyperlinks. Hyperlinks are the essence of the internet since they provide quick and efficient information transmission. Our goal is to construct a knowledge network from articles published on Slovene Wikipedia which will serve as a ground truth for our project. Moreover, we will construct two networks; network where links represent *wikilinks* (internal links) between articles and network where links represent the lexical similarity between them. Upon those two networks we will apply different network science approaches (community detection, link prediction, clustering, etc.)

Problem definition, motivation and background. The motivation for our project consist of several problems that we would like to tackle. First we would focus on **exploratory data analysis**, where we would search for different patterns that exist in the constructed networks. We would try to examine how the network constructed from the hyperlinks **differs** from the network constructed with lexical similarity. In both of these networks we would try to find the **communities** and compare them with the actual categories to see if the categories really reflect the communities or not. Next, we would focus on different **modularity measures** to find which articles are the most important or dominant. We would also focus a lot on **visualization** which we think is an essential part in any data science related work. If we can present Wikipedia articles in a meaningful presentation, we would be able to quickly distinguish between the network from hyperlinks and network from lexical similarity.

Finally, when thousands of volunteers are writing their articles to be published on Wikipedia, they are dealing with the problem where they might **forgot to reference** an important wiki page that should be included in their article. We want to have an encyclopedia that is as precise and as informational as possible and we achieve that with referencing the right articles. We would tackle this problem with **link prediction** methods, where we would recommend which *wikilinks* an author might include in the article.

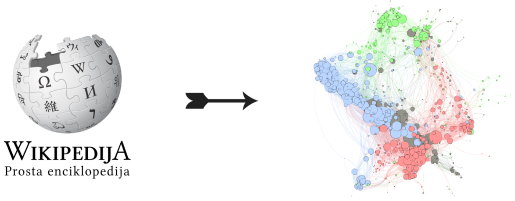

Fig. 1. Illustrative representation of our project. [source: Wikipedia and Medium]

Related work

Some network science work has been done on Wikipedia articles, but their goals were different than ours. Authors of article (1) have tackled collective behaviour analysis on Wikipedia, meaning which pages are most viewed in certain time frame. Next, some research has been done in measuring article quality using collaboration network (2). The challenge of organizing a set of domain-specific terms into a hierarchical structure was solved using so-called DF-Miner (3). Authors of article (4) were interesting in the role of diversity in Wikipedia, more specifically, the correlation of teams and the quality of article.

Project proposal

Our project will consist of the full network science pipeline, from retrieving the data to exploring the networks. The first task would be to extract data from the **Slovenian Wikipedia** page. Since the total number of pages in Slovenian Wikipedia is slightly more than **176 000**, we would try to scrape all of the pages and their content to construct the two mentioned networks. If retrieving all articles would for some reason not be possible, we would use some of the **sampling methods** that we addressed in the course (Random Sampling and BFS – we might try them anyway). Firstly we would of course start with the core network science metrics such as average degree, clustering coefficient, degree distribution, etc., apply them to both networks and compare the results. Next, as already mentioned, we would use different methods for **community detection** (Louvain, Infomap, Label Propagation) and examine their results by comparing them to the actual categories that the articles are included in. Then we would try to find articles that are most important for the Wikipedia using different **modularity** approaches (Betweennes centrality, Closeness centrality, PageRank, etc.). And finally we would tackle the problem of **link prediction**, using different prediction methods, based on the structural properties of our networks. We will construct a framework for **recommending** which articles should an author reference and will be based both on already included *wikilinks* and the textual similarity of the articles.

1. V. Miz, K. Benzi, B. Ricaud, and P. Vandergheynst. Wikipedia graph mining: dynamic structure of collective memory. *arXiv:1710.00398 [cs]*, February 2018. URL http://arxiv.org/abs/1710.00398. arXiv: 1710.00398.
2. B. de La Robertie, Y. Pitarch, and O. Teste. Measuring Article Quality in Wikipedia using the Collaboration Network. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, pages 464–471, Paris France, August 2015. ACM. ISBN 978-1-4503-3854-7. . URL https://dl.acm.org/doi/10.1145/2808797.2808895.
3. B. Wei, J. Liu, Q. Zheng, W. Zhang, C. Wang, and B. Wu. DF-Miner: Domain-specific facet mining by leveraging the hyperlink structure of Wikipedia. *Knowledge-Based Systems*, 77: 80–91, March 2015. ISSN 0950-7051. . URL https://www.sciencedirect.com/science/article/pii/S0950705115000088.
4. K. Baraniak, M. Sydow, J. Szejda, and D. Czerniawska. Studying the Role of Diversity in Open Collaboration Network: Experiments on Wikipedia. In A. Wierzbicki, U. Brandes, F. Schweitzer, and D. Pedreschi, editors, *Advances in Network Science*, Lecture Notes in Computer Science, pages 97–110, Cham, 2016. Springer International Publishing. ISBN 978-3-319-28361-6. .

[1]ak1601@student.uni-lj.si. [2]tv3843@student.uni-lj.si. [3]sr8905@student.uni-lj.si.