Developing models of the quality of Wikipedia articles in different languages, topics and Wikimedia projects.

Grzegorz Kopaczewski
Wikimedia Poland

dr Włodzimierz Lewoniewski
Poznan University of Economics

Abstract

The main goal of the project is to assess the quality of Wikipedia articles in various language versions and in various topic areas. Compared to existing approaches, the project plans to develop many different models for automatic assessment of the quality of articles in various thematic areas in individual language versions. Research will lead to creating a tool responding live online, giving the user suggestions and information.

Introduction

In some language versions (including Polish), a significant number of articles do not have a quality rating (e.g., in the Polish version there are over 99% of such articles). It is also worth taking into account that a Wikipedia article may change over time, which requires regular verification of its quality rating. Models developed within a given project would allow for automatic evaluation of all articles and their ongoing updating.

An important element of this project is the development of an open-source tool in Python that will allow for the assessment of the quality of articles using various quality scales. Within such a tool, the user will be able to select different criteria (e.g., language version, article topic) and measure the quality of a specific article live through a friendly user interface. Additionally, an API will be available to automatically query the results generated by the tool.

Who will benefit from the research and tools created?

- Wikipedia's editing community and recipients.

What research questions do you ask and answer?

- What measures (parameters) of Wikipedia articles are important from the point of view of assessing the quality of Wikipedia articles?
- What are the differences between quality models developed in different language versions of Wikipedia?
- What are the differences between the quality models developed for Wikipedia articles on different topics (including different WikiProjects)?

What is the problem you want to solve?

The main research problem is improving the process of assessing the quality of articles in Wikipedia. An additional problem - finding

differences between the quality models of Wikipedia articles depending on the article topics and language versions.
As part of the developed models and tools, it will be possible not only to automatically obtain an assessment of the quality of selected articles based on various quality standards (language/topic), but also to obtain information on which elements of the article can be improved in order to improve this assessment (e.g. number of references, sections, etc.)

Date: 1 June 2023 - 31 January 2024

Related work

ListWings (ORES) In some language versions of Wikipedia, it is possible to use models operating within the LiftWings tools (including migration from ORES)
https://wikitech.wikimedia.org/wiki/Machine_Learning/LiftWing
However, these models have some limitations:
Models are only available for 11 languages (no Polish version)
https://analytics.wikimedia.org/published/wmf-ml-models/articlequality/
The models are general in nature - the model within one language version is the same for articles regardless of the content topic.
The model only shows the final result as a numerical value. For example, there is no information about individual components that should be corrected in the article to improve quality.
WikiRank is another example of a publicly available tool:
https://wikirank.net/
https://live.wikirank.net/
WikiRank allows you to assess the quality of Wikipedia articles in various language versions on a scale from 0 to 100. The assessment consists of the results of assessing individual features of the articles, such as: the length of the article's wiki code, the number of footnotes, the number of sections, the number of illustrations with a visible resolution of at least 5000px2, density of footnotes (footnotes/length).
Compared to models from LiftWings, here you can also find information on the evaluation of 5 components that influence the rating. Thus, the creator/editor of a Wikipedia article can learn what needs to be improved to make the article have a higher quality rating.
WikiRank's rating models vary depending on the language version. However, (for now) these models do not depend on the topic area of the content in the articles being assessed.
WikiArticelAuditor
As part of this project, it is planned to develop models of the quality of Wikipedia articles for a larger number of language versions, with particular focus on the thematic division of articles. Focusing on individual groups of articles at the language/topic intersection will allow the models to better match the individual quality standards developed by the Wikipedia user community (within individual language versions and thematic projects). In addition, the model will be able to indicate a larger number of parameters that influence the overall quality assessment result.

Methods

In the first stage, it is planned to use Wikipedia backup copies (https://dumps.wikimedia.org/) in various formats. The research will also use the history of article editing and website visit statistics. The data will also be supplemented by the Wikipedia API, e.g. when there is missing data in backups, in order to collect additional data.

We will use observational, experimental and statistical methods.

Wikimedia Poland staff will be responsible for forming the group of testers (wikimedians volunteers), managing them and gathering feedback. Wikimedia Poland will use its close ties with CEE Hub to gather wikimedians using different languages from CEE region and Let's Connect channel in global scale.

Expected output

Conference presentations and post-conference publications are planned. The extraction results of selected measures and the performance of various quality models will be published openly (e.g. via Figshare).
An important part of this project will be the development of a tool that will enable anyone to assess the quality of Wikipedia articles in different language versions. The Python source code will also be available openly (GitHub).

Risks

The resulting models and the developed tool may not have much interest from the Wikipedia users community. Good communication campaign would decrease the risk.

Community impact plan

As said above, the Wikimedia Poland staff will help to reach the testers from Polish and CEE community. The WMPL Community Support team will reach for volunteers from other regions thought channels like Let's Connect and will ask Wikimania 2024 participants to join. WMPL Communication manager will inform about the research and its results in social media, Diff and other channels.

Evaluation

Automated systems (via API) will be used. Gathering feedback from testers and, then, users by Wikimedia Poland staff (e.g., google forms, discussions in social media channels).

Budget

Total: 37 187 USD

33 722 USD - three researchers salary or stipends
3465 USD - WMPL Institutional overhead: (administration, organizing conference presentation, forming and managing the testers group, gathering feedback from them, communication - including website on Meta)

Prior contributions

Wlodzimierz Lewoniewski has been involved in research related to the analysis of the quality of Wikipedia articles since 2014. The doctoral thesis, which was defended with honors in 2018, concerned the development of a method for comparing and enriching information in multilingual wiki websites (primarily Wikipedia) based on the analysis of their quality. The dissertation is available at: https://www.wbc.poznan.pl/Content/461699/Lewoniewski_Wlodzimierz-rozprawa_doktorska.pdf
Wlodzimierz is Assistant Professor at Department of Information Systems in Poznań University of Economics, the author of over 30 scientific publications on the analysis of the quality of Wikipedia articles and the sources of their information. Some of wiki-related publications were co-created by Krzysztof Węcel and Witold Abramowicz (also researchers working at University of Economics in Poznań, Department of Information Systems).

Wlodzimierz took part in various events related to Wikimedia projects (including Wikimania, Wiki Workshop, Wikimedia Polska conference, etc.)
The link to the webinar recorded for WMPL (in Polish):

https://www.facebook.com/WikimediaPolska/videos/837514828116233/?extid=NS-UNK-UNK-UNK-IOS_GK0T-GK1C&ref=sharing

Włodzimierz Lewoniewski is a co-creator of WikiRank.

References

You can find Wlodzimierz's publications here: https://scholar.google.com/citations?user=rvOakL8AAAAJ

University profiles:
https://kie.ue.poznan.pl/pl/wlodzimierz-lewoniewski/
https://kie.ue.poznan.pl/pl/krzysztof-wecel/
https://kie.ue.poznan.pl/pl/witold-abramowicz/