Topic analysis performed on data from Reddit

Maria Kędzierska, Marcelina Kurek, Mikołaj Spytek

Motivation

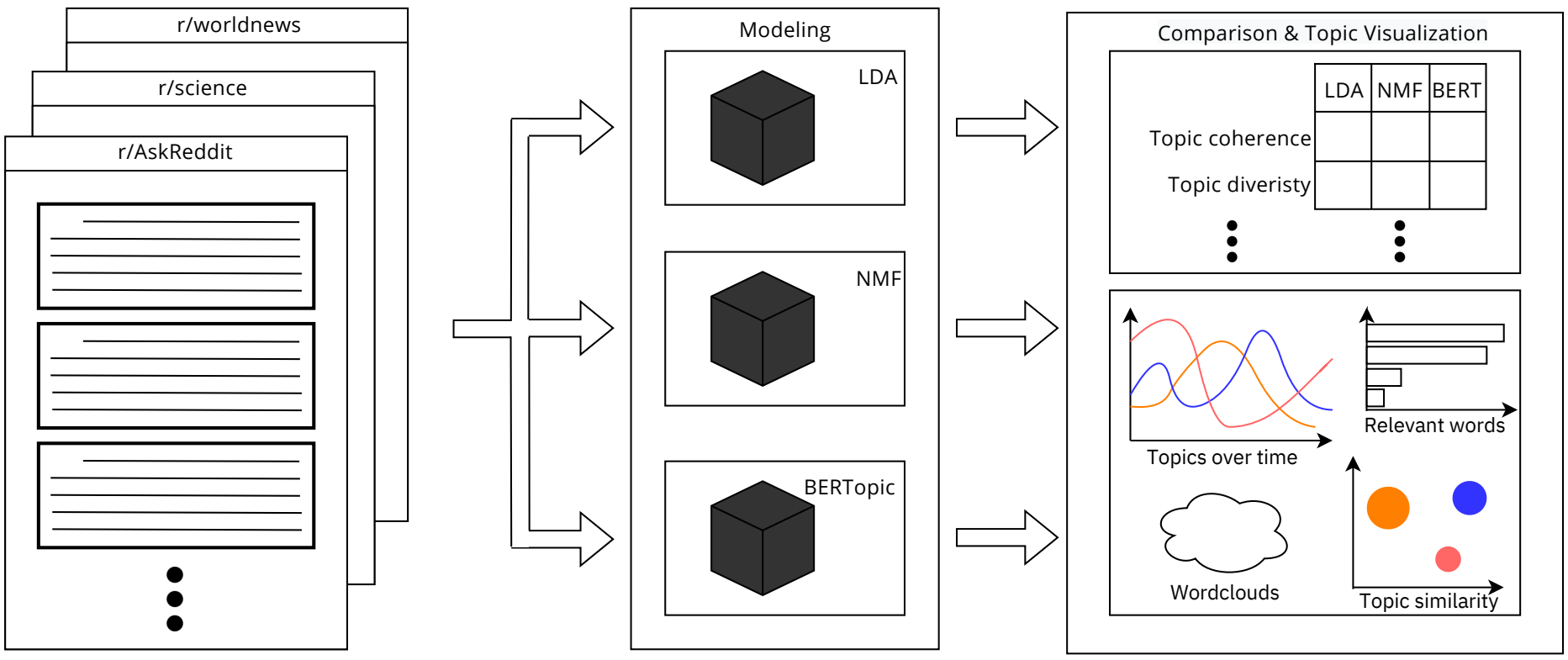
Reddit is a valuable source of textual information in both commercial and scientific applications. However, the amount of posts is too large to study manually, without limiting the dataset to a small number of documents.

Topic modeling techniques, are tools for capturing main ideas contained in documents and summarizing results in short lists of topic keywords that are interpretable by people. However, because of the different mathematical backgrounds of these models, they often yield varying results.

Proposed solution

We propose an open-source, end-to-end application for the exploration of topics present on Reddit using different modeling techniques. It allows for automatic downloading, processing, modeling of posts, as well as comparing and visualizing the extracted topics.

The application collects different topic models and evaluation techniques from various Python packages. Additionally it is able to download posts. It presents the results as an interactive dashboard implemented in *Dash Plotly*.

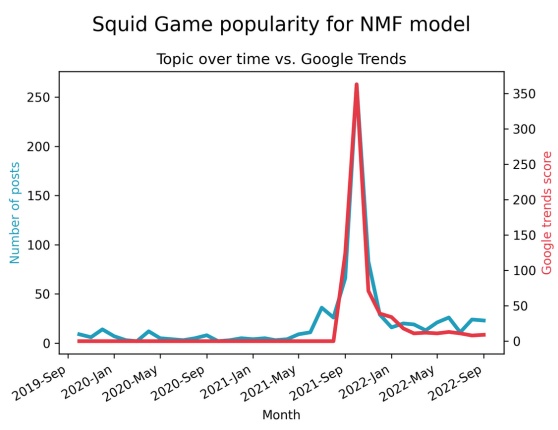


A diagram illustrating the proposed solution

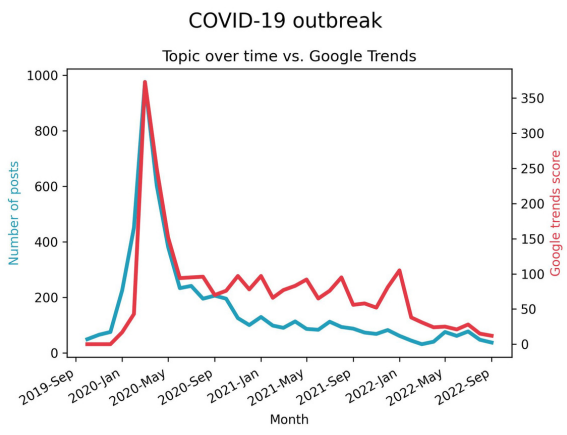
Experiments

The application was thoroughly tested to find out if the extracted topics are insightful. We compared the results with another tool for gauging topic popularity on the Internet - Google Trends.

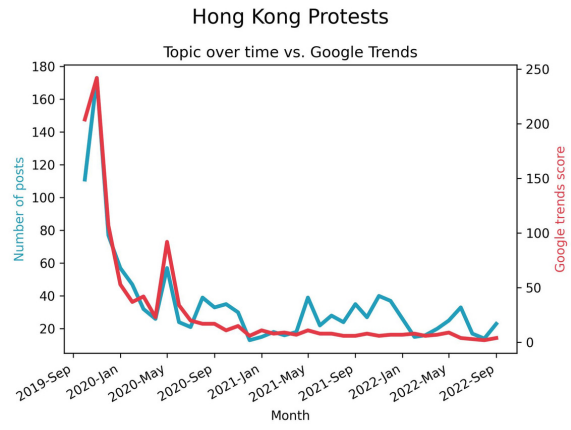
Given the topic keywords from the output of the model, which are ordered according to the keyword importance, we tried to summarize them in a way it could be searched for on Google.



Topic keywords: game, squid, make, video, korean, adult, news, borderland, alice, play



Topic keywords: coronavirus, case, confirm, test, first, outbreak, deaths, death, toll, health



Topic keywords: hong, kong, protest, protesters, china, police, law, chinese, tiananmen, prodemocracy

Faculty of Mathematics and Information Science
WARSAW UNIVERSITY OF TECHNOLOGY

https://github.com/kaluskam/topic-models-comparison