ANALYSIS OF DEPTH-1 HYPERLINK BUBBLES IN SUBREDDITS

AUTHORS
Nico Aebischer & Colin Fingerlin

DATABASES PROJECT HS2021
Universität Basel

A depth-1 hyperlink-bubble of a subreddit is the set of all hyperlinks that are referenced in the comments of a subreddit. Each bubble can be associated with multiple categories, depending on what kind of links are being sent. Categories could be "technology", "business" or "sports". The hyperlinks sent in comments on Reddit are classified into categories. The categorized data is used to make statements about the type of content that is sent within Subreddits and its alignment with the expected categories is measured. We used this framework together with a large dataset of user comments to analyse Subreddits on reddit.com.

OBJECTIVES

The goal of the project was to create categorized data to make statements about the type of content being sent in Subreddits. The Reddit comments of June 2021 are the foundation of our work. The categorization performance of our machine learning tool should be measured and quantified. Unrelevant or misleading results should be filtered out. Indexes of the most frequently sent URLs per subreddit and the corresponding composition of categories should be made easily accessible to users via a web-application.

METHODOLOGY

We detected and extracted hyperlinks from the roughly 200 million Reddit comments of our base dataset. These hyperlinks were scraped for text tokens and subsequently run through a linear SVM classication tool to obtain 265 thousand unique domains with category scores. After integrating our novel dataset with the original comments we created a model for relevancy and combination of scores of particular categories. Our model was then applied to the data with SQL queries and the resulting merged scores for Subreddits make up the results of the project.

RESULTS

Working on the project, we integrated multiple datasets in a PostgreSQL database. In the process of webscraping, a new sizable dataset of website text tokens was created and combined with existing material. A web-application was implemented with the help of which the analysis and visualization per Subreddit is easy and fast. In numbers we summarize:

- 208 million comments in the dataset
- 290 thousand Subreddits
- 10 and a half million comments with at least one link
- Roughly 4 thousand comments lost due to corrupted source data
- 265 thousand domains classified in total.
- 175 thousand domains classified and considered relevant

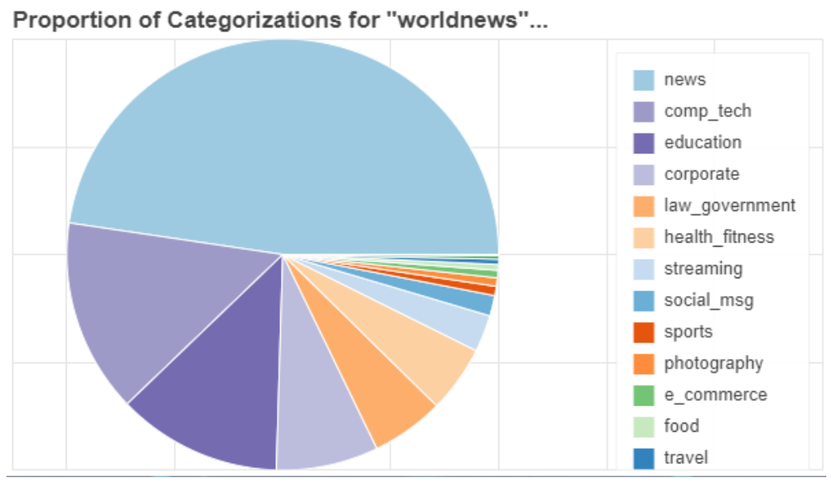

Figure: Unweighted aggregate for categories excluding overrepresented domains.

ANALYSIS

We were able to verify that the expected categories for Subreddits match well with the actual content being sent. This finding was quantified by creating indexes of the predominant categories for Subreddits and then checking whether or not they match the expected categories. An example would be the Subreddit "worldnews", where we would expect a lot of news-related content being shared.

The analysis can be done interactively with the help of our web-application, where one can search for Subreddits and gets an index of the top 11 domains found in the respective Subreddits as well as an indication of how the proportions of categories are distributed. A distiction is made between counting individual domains once versus counting them as many times as they appear in the comments of the Subredit.

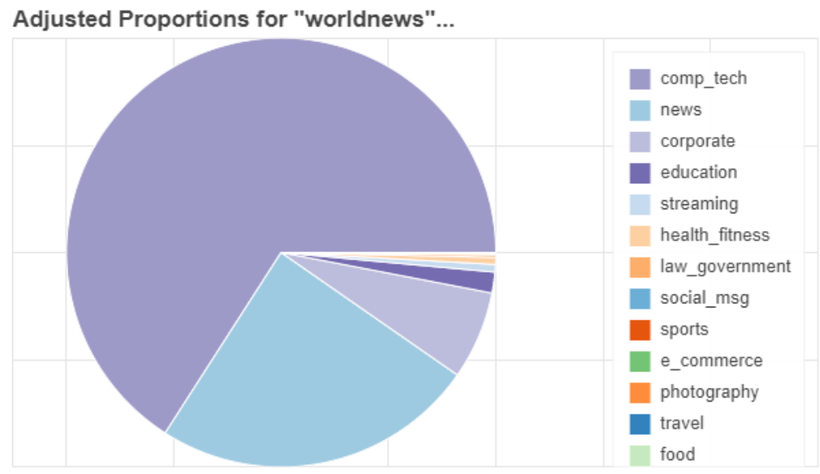

Figure: Weighted aggregate for categories excluding overrepresented domains.

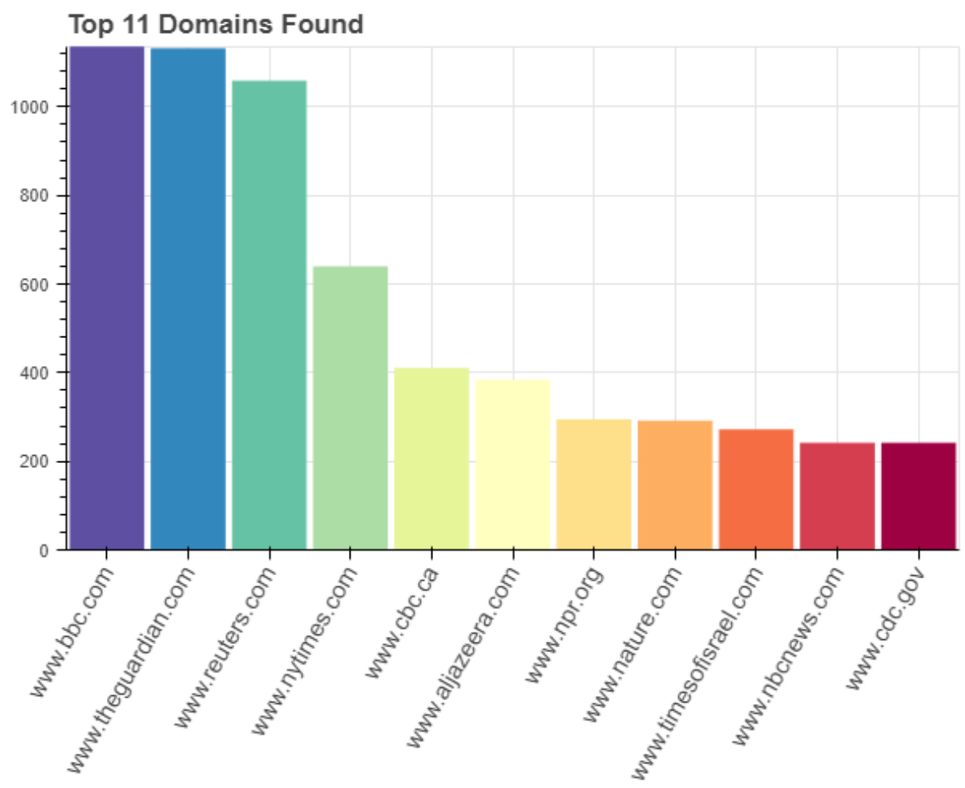

Figure: Top 11 domains posted in r/news excluding overrepresented domains.

We display both results, weighted, if we account for duplicity and unweighted, if we disregard it in our application. We further discarded over-represented domains, we found to be "staples" which are abundant cross Subreddit such as self-references to reddit.com or references to certain image and video hosting platforms like imgur.com and youtube.com as they would otherwise overshadow nuances.

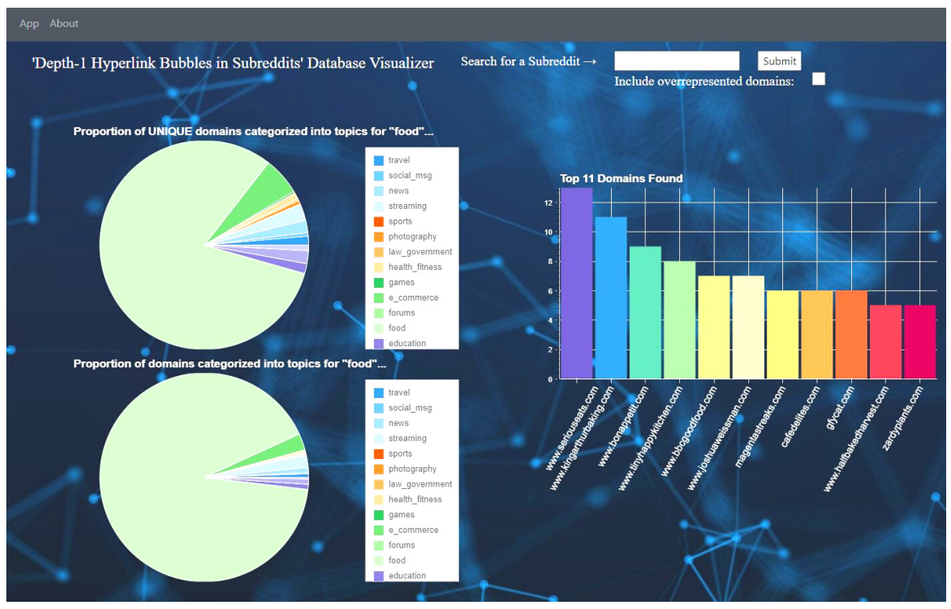

Figure: Example Query from the web-application displaying the results for r/food.