Assignment 1

Yuan Zhou

Abstract

COVID-19 has been a catalyst for a new technology trend because of how it reshapes our lives. In this research, we experimented with building a focused web crawler to generate a corpus from news articles related to COVID-19 reported by tech media, intending to discover the trending new technologies post-pandemic era.

1. Introduction

COVID-19 has fundamentally shifted our perspective of the surrounding world and reshaped our ways of working, entertaining, and socializing. To some extent, it is a catalyst for many new kinds of blooming technology: telemedicine, robot-deliveries, and 3D-printing, to name a few. This research intends to discover the most prevalent technologies covered by the news articles reporting COVID-19 to enlighten us on future technology trends.

To achieve that, we built a focused web crawler, following the steps initially introduced by Chakrabarti, S., Van Den Berg, M., et al. [1], with our adjustments to the theory and practice.

2. Literature Reviews

The idea of focused web crawling becomes increasingly important due to the rapid growth of web information. Many researchers have contributed to the development and improvement of this field. In his profound article[1], S. Chakrabarti introduced a system utilizing a classifier and a distiller to guide the crawler. After that, Aggarwal et al. [2] presented a significant improvement to the focused crawling

approach called intelligent crawling. K. Stamatakis adopted a different approach to focused web crawling named CROSSMARC.[3]

The success of focused crawling made the Web a goldmine of valuable language data, and it has been exploited to build training and test corpus for a variety of NLP tasks.[4][5] Hence, web-based NLP research becomes prevalent. A piece of evidence is that Clément de Groc developed a focused crawler Babouk for corpus compilation.[6]

3. Methodology

3.1 Topic/Taxonomy creation and refinement

COVID-19 is a suitable name for the topic because its free of ambiguity and its prevalence in usage. Additionally, we expanded the candidate topic keywords to a group of synonyms to get more crawling results relevant to our research. Next, we relied on the internal taxonomy mechanisms provided by publicly available news search APIs like https://newsapi.org/ to collect the first set of example documents to evaluate the effectiveness of such keywords to crawl the results we cared most. After a few rounds of search and evaluations, we decided the final keywords used to retrieve relevant news URLs are: COVID-19, coronavirus, SARS-CoV-2, 2019-ncov.

3.2 Resource Discovery and Distiller Algorithms

Our focus is on the trending technologies, so the sources reporting technology news are our ideal hub for information. In practice, we leveraged the News API's source taxonomy to narrow down our crawling only to 'technology' category sources.

In theory, we could prioritize web pages resources based on a distiller. As the crawler progresses, it could maintain a graph of references, where vertices are unique websites, and edges are URL links between them, with a direction from the referrer to

referee. The graph expands as new websites are reached. Periodically, the crawler calculates each website's eigenvector centrality to prioritize websites and select the information hubs based on the ranking.

3.3 Classifier Training

We could use the Deep Learning CNN to build our classification model that stacks with a word-embedding layer, multiple 1D-CNN/MaxPooling, and DNN layers. To start, we could run the focused crawler without any classifier integrated to gather the unclassified documents and manually label them as relevant or irrelevant to construct our initial training/test set. Then we train the classification model and evaluate the relevancy of crawled documents going forward. The classified docs could continuously be fed to the model to improve its accuracy.

3.4 Building the crawler in practice

We built a web crawler on top of the Scrapy framework, focusing only on English articles. Once executed, the crawler first searches for a URLs list by keywords defined in 3.1 from News API and then scrapes each referred website's HTML content. Each website's URL, title, and text contents are extracted to construct a JSON object with a unique ID. Some cleanups are also applied: the crawler cleans the text body by removing HTML tags, non-English words, and punctuations, and strips off all text after the first 500 words in the doc. The crawler then stores all crawled data into a JSON-lines file.

4. Results & Evaluation

In the crawler's last execution, 402 technology articles of topic COVID-19 from tech media are converted to a JSON-lines file named items.jl. We also saved each web

page's HTML source under a directory named news-html. We have evaluated the quality of the generated corpus: no empty values in any fields, all articles in English, and the corpus are relevant to COVID-19 as we read it.

There are some key learnings from designing and building this focused crawler:

- Iterative development of the crawler with a feedback loop played a significant role in crawling high-quality corpus. In the beginning, we restricted our corpus to 10 lines to debug the crawler and validate each scraped values. Due to heterogeneous HTML elements' layouts on different websites, some bugs and missing contents could only be caught after manually reviewing the contents generated.
- Scrapy could scrape dynamically rendered web pages when integrated with Splash. We encountered a problem that crawling Google search results always return an empty list because Scrapy could not execute embedded javascript scripts to make service requests and rendering. If integrated with Splash, Scrapy crawler could dynamically render the web pages before the scraping step; this way, we could the same results as presented in web browsers.

5. Conclusion

In this research, we built part of a focused web crawler on the Scrapy framework. The corpus generated provides a fundamental dataset for further study. For example, one possible next step is developing a model that extracts the frequently occurred keywords from the corpus or a clustering model to define the document classes and eventually generates the most frequent technology terms referred by tech articles related to COVID-19.

References

[1] Chakrabarti, Soumen, Martin Van den Berg, and Byron Dom. "Focused crawling: a new approach to topic-specific Web resource discovery." Computer networks 31, no. 11-16 (1999): 1623-1640.

[2] Aggarwal, Charu C., Fatima Al-Garawi, and Philip S. Yu. "Intelligent crawling on the World Wide Web with arbitrary predicates." In Proceedings of the 10th international conference on World Wide Web, pp. 96-105. 2001.

[3] Stamatakis, Konstantinos, Vangelis Karkaletsis, Georgios Paliouras, James Horlock, Claire Grover, James R. Curran, and Shipra Dingare. "Domain-specific Web site identification: the CROSSMARC focused Web crawler." In International Workshop on Web Document Analysis (WDA2003). pp, vol. 75, p. 78. 2003.

[4] Nakov, P., & Hearst, M. A. (2005, June). Search engine statistics beyond the n-gram: Application to noun compound bracketing. In Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005) (pp. 17-24).

[5] Banko, Michele, Eric Brill, Susan Dumais, and Jimmy Lin. "Askmsr: Question answering using the worldwide web." In Proceedings of 2002 AAAI Spring Symposium on Mining Answers from Texts and Knowledge Bases, pp. 7-9. 2002.

[6] De Groc, Clément. "Babouk: Focused web crawling for corpus compilation and automatic terminology extraction." In 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, vol. 1, pp. 497-498. IEEE, 2011.