A preliminary Study of Knowledge Sharing related to Covid-19 Pandemic in Stack Overflow

Konstantinos Georgiou
School of Informatics
Aristotle University of Thessaloniki
Thessaloniki, Greece
kageorgiou@csd.auth.gr

Nikolaos Mittas
Department of Chemistry
International Hellenic University
Kavala, Greece
nmittas@chem.ihu.gr

Lefteris Angelis
School of Informatics
Aristotle University of Thessaloniki
Thessaloniki, Greece
lef@csd.auth.gr

Alexander Chatzigeorgiou
Department of Applied Informatics
University of Macedonia
Thessaloniki, Greece
achat@uom.edu.gr

***Abstract*— The Covid-19 outbreak has changed to an unprecedented extent almost every aspect of human activity. At the same time, the pandemic has stimulated enormous amount of research by scientists across various disciplines, seeking to study the phenomenon itself, its epidemiological characteristics and ways to confront its consequences. Information Technology, and particularly Data Science, drive innovation in all related to Covid-19 biomedical fields. Acknowledging that software developers routinely resort to open 'question & answer' communities like Stack Overflow to seek advice on solving technical issues, we have performed an empirical study to investigate the extent, evolution and characteristics of Covid-19 related posts. Through the study of 464 Stack Overflow questions posted in February and March 2020 and leveraging the power of text mining, we attempt to shed light into the interest of developers in Covid-19 related topics and the most popular problems for which the users seek information. The findings reveal that indeed this global crisis sparked off an intense activity in Stack Overflow with most post topics reflecting a strong interest on the analysis of Covid-19 data, primarily using Python technologies.**

Keywords—Covid-19; Stack Overflow; Data Analytics

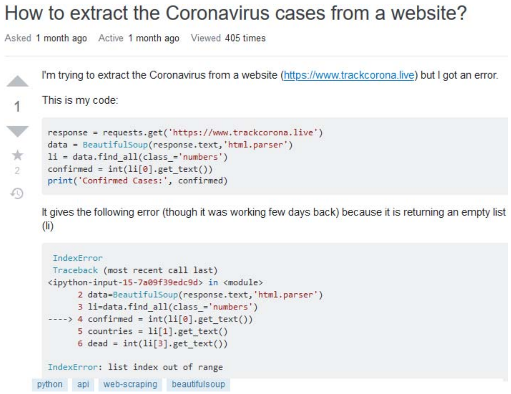
I. Introduction

The outbreak of the Covid-19 disease has caused a global crisis, which besides the profound impact in health care, has affected all other aspects of everyday life. Due to the growing number of infection epicenters, emphasis was placed on tight restrictive measures. Such strategies necessitated distance work, education, entertainment and social activities. These measures soon led to massive demand of technological support, especially regarding communications. Apart from technologies supporting everyday activities during the outbreak, there are growing needs for Information Technologies (IT) that could aid the battle against the disease. Similarly, researchers are investigating the root causes of the phenomenon, its spread patterns and models governing its evolution.

The motivation of our work is based on our perception of the IT activities in demand during the outbreak. Therefore, we decided to investigate whether and how the new intense circumstances are reflected on developers' posts in a "Q&A" portal such as *Stack Overflow* (SO). The goal is to study Covid-19 related posts so as to understand why and for which topics developers are interested and how these topics are associated. Also, our study aims to investigate the evolution of posts over time. We should note that with the term 'developers' we do not necessarily assume that users posting on SO are software professionals. Especially with respect to Covid-19, we consider it possible that scientists from various domains are pursuing research that relies on *scientific software* [1] targeting collection, analysis and visualization of Covid-19 related data.

II. Research Objectives and Research Questions

The main idea behind this study is to investigate, whether the Covid-19 outbreak has triggered developers to post relevant questions in knowledge sharing communities. We focus on posts where the issue being discussed is associated to the use of IT methodologies to investigate, understand and provide solutions to the Covid-19 crisis. To better illustrate our objective, we provide a representative example of a coronavirus-related post (Figure 1).



How to extract the Coronavirus cases from a website?

Asked 1 month ago Active 1 month ago Viewed 405 times

I'm trying to extract the Coronavirus from a website (https://www.trackcorona.live) but I got an error.

This is my code:

```
response = requests.get('https://www.trackcorona.live')
data = BeautifulSoup(response.text,'html.parser')
li = data.find_all(class_='numbers')
confirmed = int(li[0].get_text())
print('Confirmed Cases:', confirmed)
```

It gives the following error (though it was working few days back) because it is returning an empty list (li)

```
 IndexError
 Traceback (most recent call last)
<ipython-input-15-7a09f39edc9d> in <module>
      2 data=BeautifulSoup(response.text,'html.parser')
      3 li=data.find_all(class_='numbers')
----> 4 confirmed = int(li[0].get_text())
      5 countries = li[1].get_text()
      6 dead = int(li[3].get_text())

IndexError: list index out of range
```

Fig. 1. Example of a coronavirus-related post

Based on the previous clarifications, we formulate the following research questions:

[RQ1] *When did the interest in coronavirus-related topics in SO arise and which is the evolution trend over the examined period?*

Motivation: Given the severity of the crisis and its detrimental effects to all aspects of our daily lives, in RQ1, we focus on the evolution of developers' interest in Covid-19 related topics over the limited time frame of the study.

[RQ2] *Which are the characteristics of the coronavirus-related posts?*

Motivation: Since Covid-19 outbreak is an emerging, rapidly evolving situation, there is no prior knowledge about

DOI 10.1109/SEAA51224.2020.00086

the sharing of knowledge in SO. To this regard, there is a need to explore and acquire knowledge about the characteristics of the coronavirus-related posts.

[RQ$_{3a}$] *Which are the most popular problem categories expressed by SO tags among coronavirus-related posts?*

[RQ$_{3b}$] *How these problem categories are associated to each other?*

Motivation: In RQ$_{3a}$, we focus on the identification of the most popular technological aspects discussed by developers. Beyer et al. [2] point out that SO posts can be classified into (*i*) *problems* and (*ii*) *questions*. Problems, expressed by SO tags, refer to the topics or technologies that are discussed, whereas questions are associated to the purpose of the post representing "*the kind of information requested in a way that is orthogonal to any particular technology*" [3].

Hence, in RQ$_{3a}$, we wish to investigate the technological content of Covid-19 related topics, and for this reason, we extract and analyse the SO tags associated to each post. In RQ$_{3b}$, our aim is to explore underlying patterns expressed by the co-occurrences of tags assuming that they contain information about interconnected technological aspects.

[RQ$_4$] *What types of topics are developers asking about?*

Motivation: In RQ$_4$, we focus on the discovery of what is being asked in coronavirus-related posts. Although tags can provide meaningful insights of technological content, they do not identify purposes, issues, problems and generally the reasons that lead the users to post a question [2]. To this end, we make use of a topic modelling methodology.

III. Methodology

The methodology of our study is illustrated in Figure 2 and consists of seven-steps: (*i*) data collection, (*ii*) feature extraction, (*iii*) data cleaning and pre-processing, (*iv*) information extraction, (*v*) data analytics, (*vi*) knowledge synthesis and (*vii*) dissemination of the findings. Since our main objective is the investigation of what developers are asking about coronavirus-related topics, the data collection (Step 1) is restricted only to question posts. The employed search string comprised synonyms of coronavirus (*"coronavirus"* OR *"covid*"* OR *"corona-virus"*) to retrieve the set of 601 questions posted until April 1st, 2020.

As far as the feature extraction concerns (Step 2), the basic unit of analysis is the *question post*. Each post P is defined as a tuple of a finite number of elements of the form $P = \langle id, t, tag, d, na \rangle$ where id is the *identification number*, t is the *title*, tag are the associated *tags*, d is the *posting date* and na is the *number of received answers*.

Step 3 involves the application of *Text Mining* (TM) to reduce the size and noise of unstructured data through *data cleaning* and *text pre-processing*. We first filtered out posts that were clearly irrelevant to the objectives and next, we conducted the following steps on the remaining 464 posts: (*i*) removal of stop-words, (*ii*) tokenization, (*iii*) punctuation removal and (*iv*) lemmatization. Initially an exhaustive list of all tags was formed and then it was transformed into a set of indicator variables (FALSE/TRUE) denoting the absence/presence of a specific tag in a post (Step 4).

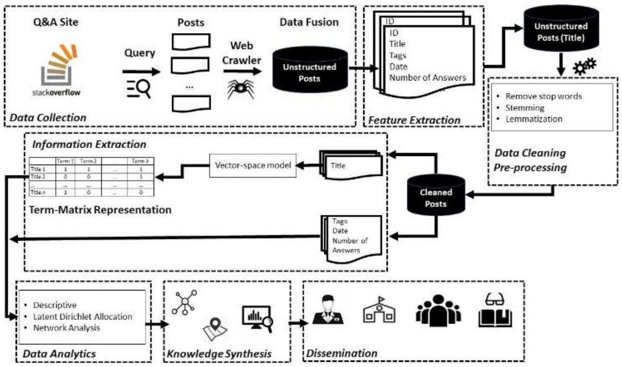


Fig. 2. Methodology of the study

Step 5 involves the application of data analytics on the set of features to derive conclusions about the posed RQs. Concerning RQ$_1$, we studied the distribution of the collected posts regarding their date, whereas for RQ$_2$, we evaluated descriptive statistics on the extracted post features.

Regarding the analysis of tags, we explored the number of tags for each post (RQ$_{3a}$) and then, through graph theory whether there are inter-connections among tags (RQ$_{3b}$). A *tagging action* can be described as a tuple of three elements $tagging = \langle u_i, p_j, tag_k \rangle$, implying that user i tagged post j with a set of tags k. We developed an *Association Rule Graph* (ARG) [4] for tags and their associations representing inter-connected technological aspects.

Three concepts play pivotal role to the ARG approach: The *frequency* ($freq(tag_i)$) of a tag, defined as the number of occurrences of a tag i, the *support* ($supp(tag_i, tag_j)$), defined as the number of co-occurrences of a specific pair of tags $\{tag_i, tag_j\}$ and the *confidence* ($conf(tag_i \rightarrow tag_j)$) representing the conditional probability of the occurrence of tag_j in a post, given that the post is already tagged by tag_i, where $freq(tag_i) < freq(tag_j)$. ARG is then defined as a graph where each tag represents a vertex accompanied by a weight evaluated by the tag frequency. For each pair of tags there is a directed edge weighted by the confidence.

For RQ$_4$, we performed *Latent Dirichlet Allocation* (LDA) analysis [5] on the question titles, since titles provide straightforward information about the problem being asked [6]. The selection of the number of topics K is based on experimentation, since there is no optimum value for all experimental setups [6]. We decided to set $K = 6$, since this value provided a broad collection of topics capturing the patterns hidden in the collection of posts.

IV. Results

[RQ$_1$] The search in SO indicates that "post-zero" is traced back to 26th of January 2020, in which the user wished to print values from an array of Covid-19-related text in PHP. In Figure 3, we present the number of posts per day for two months. The figure shows a significant increase of posts for month March (approximately 95% of all posts) compared to month February (approximately 4% of all posts), when the Covid-19 outbreak presented a peak in several countries.

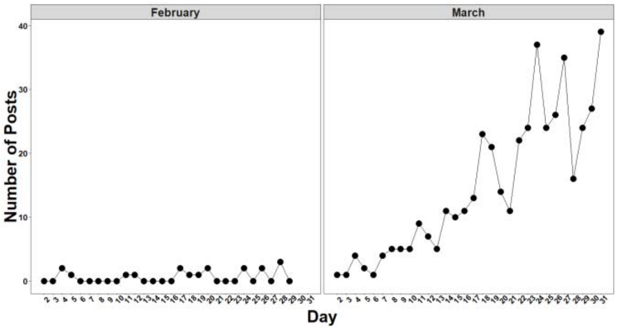


Fig. 3. Number of Covid-19 related posts for each day of February/March

TABLE I. NUMBER OF COVID-19 RELATED POSTS (FEBRUARY-MARCH)

# Answers & Tags			Top 10 Popular Tags			
#	f_i	f_i	Tag	f_i	Tag	f_i
0	38.36%	-	python	28.7%	python-3.x	6%
1	46.26%	9.91%	r	18.1%	html	5.8%
2	14.22%	27.8%	pandas	12.1%	json	5.8%
3	3.88%	28.66%	javascript	12.1%	dataframe	5%
4	0.43%	20.04%	web-scrapping	8%	beautifulsoup	5%
5	0.43%	13.58%				

[RQ$_2$] To gain better insights on the characteristics of the phenomenon, we initially performed statistical analysis on the associated features that can be found in each post.

Table 1 presents the distribution of the number of received answers indicating that more than 60% of the posts has been responded by at least one user (mean value $M = 0.87$, $SD = 0.88$, $Mdn = 1$). Considering that we have tracked Covid-19 related posts for a two-month period, it is rather reasonable that most posts have received a limited number of answers. Due to the observed growing interest, there is a need to examine, whether these questions are posted by different users or whether, there is only a specific proportion of users that are interested in. To this regard, we evaluate the percentage of unique users who posted a question related to a coronavirus issue. The percentage of unique users (# of different users that posted one question divided by the # of posts) was 79.09% signifying a broad awareness of the community.

[RQ$_3$] The analysis on the occurrences of SO tags showed that most users (56.46%) defined from two up to three tags (Table 1). Table I presents also the distribution of the 10 most popular tags from the set of 482 total tags, from which we can observe that the tags "python" and "r", which are two of the most known programming languages for Data Science, dominate in terms of their occurrences.

Due to the multifaceted nature of Data Science, there is a variety of challenges that a practitioner should be able to handle. The remaining most popular tags (except for "python" and "r") are closely related to phases of the whole lifecycle of Data Science. For example, "web-scrapping" and "json" are related to the extraction of information from websites, whereas "beautifulsoup" is a Python package that is used for parsing web documents. In addition, "dataframe" is a well-known data structure for the organization of information and this is also the case for the Python library "pandas". Finally, posts tagged by "html" or "javascript" reveal an interest to seek knowledge about scraping content from HTML documents (usually residing in Javascript-enabled pages), most often through parsers developed in Python or node.js or with the help of test automation tools. Once again the focus is on data collection and developers wish to automate the parsing from online sources and convert the extracted information into data structures of file formats for further processing.

Regarding the topic of discussion, the manual review process through the inspection of the title and body of the posts reveals that the users are mostly interested in collecting, organizing and analyzing data of Covid-19 global cases (i.e. confirmed, deaths, recovered etc.).

After the identification of the most popular tags (RQ$_{3a}$), we focus on the investigation of the interdependencies between technological aspects (RQ$_{3b}$). To this regard, we construct an ARG for the top 10 popular tags. The inspection of Figure 4 indicates that the two tags of languages "python" and "r" are represented by generally large circles (the area of the circle represents the frequency of the tag). Several well-known technologies are associated to the tag "python", including libraries and frameworks for web-parsing ("beautifulsoup", "selenium"), data manipulation ("pandas"), interactive visualization and geospatial data analysis ("geopandas", "geojson").

With respect to another frequent tag, namely "javascript", we observe that almost all associated nodes point to libraries, frameworks and language elements needed for extracting content or posting from/to the Web: the widely used libraries "jQuery" and "d3.js", are related to HTML/XML document manipulation, while module "cheerio", method "get" and API "fetch" facilitate the manipulation of "http" requests and responses. From these tags, it is confirmed that developers in the era of Covid-19 are interested in getting content from sites posting relevant information (presumably virus-related demographics).

[**RQ$_4$**] The results of the application of LDA on the corpus of titles are summarized in Table 2 presenting the most frequent words for each one of the six extracted topics. Moreover, we tried to reify the conceptual meaning of each latent topic by providing a description of topics developers are asking about coronavirus-related posts.

An interesting finding in the list of frequent words is that the most common type of question is "*how to ...*", a type frequently found in SO posts [2,3,6]. Furthermore, the word "*error*" designates that developers frequently face problems related to occurrences of errors in their code.

The most popular technological topic seems to be the collection of data related to Covid-19 cases through web-scraping techniques and content retrieval from websites, whereas the users also post questions related to the data analysis and visualization of reported cases through maps for geospatial analysis. Finally, as far as specific technologies concern, it is evident again that languages Python and R play a predominant role to the community of developers for examining challenges related to Covid-19.

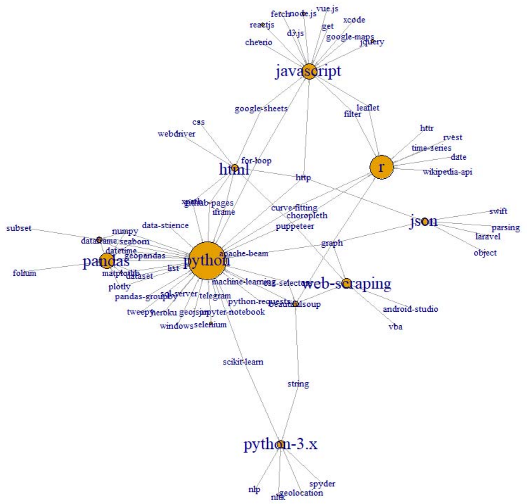


Fig. 4. Association rule graph for the top 10 tags

TABLE II. TOPIC MODELIING OF COVID-19 RELATED POSTS

Topic	**Frequent Words**
Web-scrapping and data manipulation with Python	how, python, pandas, value, column, list, string, web, table, scraping
Error messages in importing data in R	how, data, error, when, r, object, csv, value, file
Importing web content to data structures and files	how, using, data, table, file, get, website, name
Importing web content to data structures and files in R	how, data, error, using, file, column, r, json, url, api
Data manipulation and analysis (time series) in R	how, data, pandas, error, when, using, time, r, value, api
Data visualization (time-series, geospatial) in R	how, using, data, r, date, api, website, map, problem, value

V. DISCUSSION

In this section, we discuss key findings organized per research question and the implications to stakeholders.

RQ$_1$: *Developers interest on Covid-19-related topics in SO*

From the number of Covid-19 related posts per day, it becomes evident that the developers' increasing interest over time follows closely the global trends for the number of cases, fatalities and even the popularity of Covid-19 terms in search engines. This is reasonable but also proves the quick responsiveness of the scientific software community against global phenomena and emerging topics.

RQ$_2$: *Characteristics of the coronavirus-related posts*

Despite the rapidly increasing trend in the number of Covid-19 related posts, it seems that the pandemic has taken the world by storm. All identified Covid-19 posts have received a rather limited number of answers and at the same time most users posting Covid-19 related questions are unique. This is of course a consequence of the limited period in which the analysis of SO posts has been carried out, but it could also imply that the rapid developments did not allow the forming of mature research communities.

RQ$_3$: *Popular problem categories expressed by SO tags*

The analysis of SO tags implies a strong interest on seeking knowledge about data collection, processing and visualization. The developers' goal is in most of the cases to extract information from web pages and the primary languages for this purpose appear to be Python and R. Considering that all posts are related to Covid-19, we can postulate that developers of scientific software focus on Data Science techniques to analyze the dimensions of the coronavirus outbreak. Knowledge sharing communities such as SO prove useful not only for solving everyday technical challenges but also as a means of collaboration in cases of emergencies. It should also be praised that developers are willing to share their knowledge to assist fellow researchers underlining the benefits of open source communities.

RQ$_4$: *Topics of interest*

The application of LDA confirms the overall picture obtained by RQ$_3$. The extracted topics reveal that during the pandemic developers are applying data science techniques to a typical problem, that of analyzing the day-by-day changing virus-related figures. Developers experiment with collecting data mostly through web-scraping and in most of the cases build python, r, or JavaScript applications to visualize aspects of the phenomenon. The most frequent words associated with each topic reflect the early stages of programming in a new environment where "how-to" questions come first and fixing errors right after.

All findings provide evidence that the outbreak of a global health crisis triggers increased interest for data collection and processing. Based on the evidence, we can assume as main motives the need for understanding the nature of the crisis and for forecasting its evolvement.

VI. CONCLUSIONS AND FUTURE WORK

Covid-19 is considered by many the worst health crisis of a generation. In this study, we have focused on the interest of developers on Covid-19 related topics based on SO posts, to shed light into their concerns, goals and means. The results indicate an increasing interest that matches the evolution of the pandemic. The findings clearly indicate the ongoing effort to collect, store, analyze, visualize and interpret Covid-19 related data. Collection is mostly performed through Web scraping, with Python, R and JavaScript clearly being the languages of choice.

REFERENCES

[1] Heaton, D., & Carver, J. (2015). Claims about the use of software engineering practices in science: A systematic literature review. Information and Software Technology, 67, 207-219.

[2] Beyer, S., Macho, C., Di Penta, M., & Pinzger, M. (2019). What kind of questions do developers ask on Stack Overflow? A comparison of automated approaches to classify posts into question categories. Empirical Software Engineering, 1-44.

[3] Allamanis, M., & Sutton, C. (2013, May). Why, when, and what: analyzing stack overflow questions by topic, type, and code. In Proc. of the IEEE 10th Working Conference on Mining Software Repositories (MSR) (pp. 53-56).

[4] Cui, B., Yao, J., Cong, G., & Huang, Y. (2010, December). Evolutionary taxonomy construction from dynamic tag space. In International Conference on Web Information Systems Engineering (pp. 105-119). Springer, Berlin, Heidelberg.

[5] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

[6] Rosen, C., & Shihab, E. (2016). What are mobile developers asking about? a large-scale study using stack overflow. Empirical Software Engineering, 21(3), 1192-1223.