Trends and challenges in eHealth Development: a Visualization Study on Stack Overflow

Pedro Almir Oliveira[1], Rossana M. C. Andrade[1], Pedro Santos Neto[2]

[1]Group of Computer Networks, Software Engineering, and Systems (GREat)
Federal University of Ceara (UFC)
Fortaleza, Ceara, Brazil

[2]Laboratory of Software Optimization and Testing (LOST)
Federal University of Piaui (UFPI)
Teresina, Piaui, Brazil

pedromartins@great.ufc.br, rossana@ufc.br, pasn@ufpi.edu.br

Abstract. *Purpose: This meta-paper describes the style to be used in articles and short papers for SBC conferences. Methods: For papers in English, you should add just an abstract while for the papers in Portuguese. Results: For papers in English, you should add just an abstract while for the papers in Portuguese. For papers in English, you should add just an abstract while for the papers in Portuguese. Conclusions: For papers in English, you should add just an abstract while for the papers in Portuguese.*

1. Introduction

The term Electronic Health or just eHealth was defined in 2001 as a research field resulting from the intersection of medical informatics, public health, and business. The health services of this field should be provided through computer networks such as the Internet [EYSENBACH, 2001]. Since then, this research area has been strengthened as a more significant population-share turns their attention to well-being and health issues [BLACK *et al.*, 2011]. Besides, computational paradigms solidification such as Mobile Computing [WHO *et al.*, 2011] and the Internet of Things [ISLAM *et al.*, 2015] also contributed to this strengthening of eHealth.

Due to the eHealth relevance, it is possible to find a large number of published works on this topic. Some researchers even use public databases (social network posts, paper repositories, to name a few) to analyze behaviors and trends of this area from the perspective of health services end-users [CULOTTA, 2010][PAUL and DREDZE, 2014][NGUYEN *et al.*, 2017] or from the researchers' view [CHIARINI *et al.*, 2013][GAGNON *et al.*, 2015][ROBBINS, KEUNG and ARVANITIS, 2018].

However, to the best of our knowledge, we did not find papers aiming to analyze the most discussed eHealth topics from the perspective of technology professionals (developers, analysts, and software engineers). These professionals often use Question & Answer (Q&A) websites to discuss technologies and strategies to solve a particular problem or to discuss macro challenges in a specific area [TREUDE, BARZILAY and STOREY, 2011].

An example of a Q&A website is the Stack Overflow[1] (SO). It was founded in 2008 and had one of the largest online communities of ICT professionals and programmers. Stack Overflow has over 14 million questions and answers, generating 50 million monthly visitors. Also, SO data is public and can be accessed through the Stack Exchange Data Explorer tool[2]. In addition to being a widely used tool by ICT professionals, SO is significantly used for scientific studies [RAGKHITWETSAGUL *et al.*, 2019][WU *et al.*, 2019][CHEN, COOGLE, and DAMEVSKI, 2019].

This paper aims to analyze trends and challenges in the development of eHealth solutions, from the perspective of ICT professionals, and having as data source the Stack Overflow discussions. The questions that guided this study were:

- **RQ1**: What eHealth subjects are being discussed in Stack Overflow?
- **RQ2**: What technologies (tools, frameworks, programming languages, operating systems) are being used to develop solutions in this area?
- **RQ3**: What open challenges concerns eHealth developers?

We highlight as contributions of this work: i) the collection, analysis, and visualization of eHealth development issues discussed in Stack Overflow, exposing trends and open challenges; ii) an automated strategy for continuous extraction of the most relevant topics in this area published using an Observable Notebook.

The remainder of this paper is organized as follows: Section 2 presents a background on the strategies used to conduct this research; Section 3 details our methodology; Section 4 discusses the results; In Section 5, we discuss some validity threats and how they were mitigated; and finally, in Section 6, we present the final considerations and possible future works.

2. Background

It is clear that eHealth has the potential to reduce costs and improve the quality of healthcare services. However, there are some technological challenges related to the development of eHealth solutions like high availability, scalability, fault tolerance, data management, interoperability, security and privacy, user experience, among others [ULLAH, FIEDLER and WAC, 2012] [SAHAMA, SIMPSON and LANE, 2013] [FARAHANI *et al.*, 2018]. In this paper, our goal is to use Stack Overflow discussions to understand how developers are dealing with these challenges. We use the same discussion definition proposed by BANDEIRA *et al*. (2019) that is a *"combination of a question and one (or more) answers, where there is at least one answer whose author is not the author of the questions"*.

We decided to use a Topic Detection and Tracking strategy to automate the topic extraction [ALLAN, 2012]. To be more specific, we chose the Latent Dirichlet Allocation (LDA) algorithm [BLEI, NG, and JORDAN, 2003]. This choice was based

[1] Stack Overflow website: https://stackoverflow.com/company
[2] Stack Exchange website: http://data.stackexchange.com

on the need to cluster the most important terms that occur in a *corpus* of documents. Also, several articles refer to LDA as the most used algorithm for this type of study [AIELLO *et al.*, 2013] [SUKHIJA *et al.*, 2016] [DROSATOS, KAVVADIAS and KALDOUDI, 2017]. The LDA uses a Bayesian model to calculate the probability of each topic concerning the documents and the probability of terms (words) to topics (Figure 1). It initializes by creating a random relationship of each word in a document with one of the initial *k* topics, which is parameterized by the user [BLEI, NG, and JORDAN, 2003].

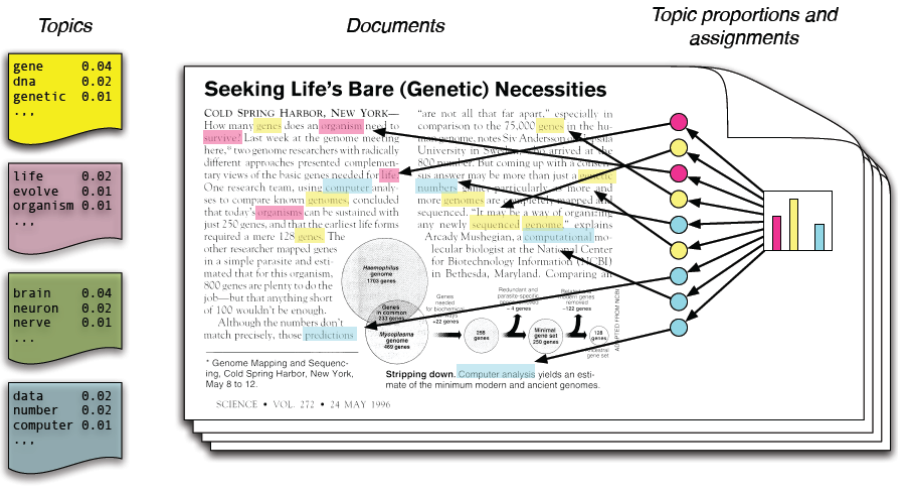

Figure 1. LDA visual representation (BLEI, 2012)

Until now, we described the background about our data source (*Stack Overflow*), the strategy used to select target data (*discussion definition*), and the data mining technique used to discover patterns (*LDA algorithm*). To improve the evaluation of these patterns, we used some data visualization techniques designed to extend human capabilities to observe insights into the data. The combination of human strengths and electronic data processing is defined by KEIM *et al.* (2008) as visual analytics.

The survey proposed by KUCHER and KERREN (2015) presents a taxonomy with some text visualization idioms - approaches for creating and manipulating visual representations. This taxonomy helped us to filter out possible visualization strategies by following criteria like analytic and visualization tasks, domain, data source, and visualization representation. Thus, to answer our questions, we decided to use word clouds [WATTENBERG and VIEGAS, 2008], node-link diagrams [HEER *et al.*, 2010], bubble maps [HEER *et al.*, 2010], and other simple graphics as line, bar, and pie charts. All visualizations were developed using a JavaScript library called D3.js [BOSTOCK, OGIEVETSKY and HEER, 2011]. D3 allows to bind arbitrary data to a Document Object Model (DOM), and then apply data-driven transformations to the document[3].

3. Study Design

In this work, we propose an investigation of trends and challenges in the development of eHealth solutions considering Stack Overflow discussions. Thus, our study design was

[3] D3.js website - https://d3js.org/

inspired by the KDD process [FAYYAD, PIATETSKY-SHAPIRO, and SMYTH, 1996] and the methodology used by BANDEIRA *et al.* (2019). The following subsections describe each of the steps presented in Figure 2.

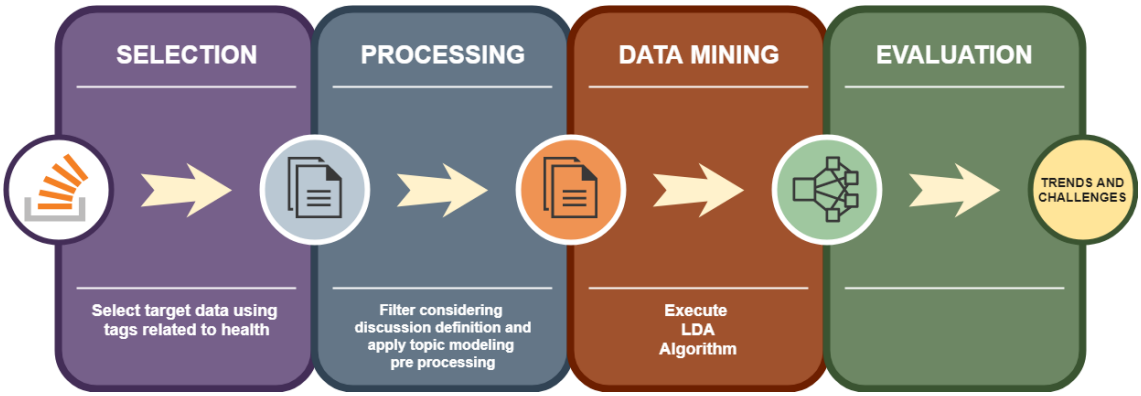

Figure 2. Our study design.

3.1. Selection

Here, we decided to use MeSH[4] controlled vocabulary [LIPSCOMB, 2000] - an instrument of the National Library of Medicine - to reduce researcher bias (detailed in Section 5) in choosing the tags to retrieve data in SO database. This vocabulary associate ehealth to mobile health, mHealth, telehealth, and telemedicine terms. However, this tags not returned any questions. Thus, we choose the tags with some empirical tests on SO tags search tool and using a snowballing approach using as a starting point the term "health". Finally, the selected tags are described in Table 1.

Table 1. Tags description.

Tags	Description (adapted from SO)	Retrieved Posts
dicom pydicom fo-dicom evil-dicom dicomweb	Digital Imaging and Communications in Medicine standard defines formats for storing and communicating medical images.	1411
hl7 hl7-fhir hl7-v2 hl7-v3 hl7-cda	The HL7 messaging standard is a communication standard for exchanging electronic information in the domain of health care.	1341
health-kit	HealthKit is a framework for iOS and watchOS that allows health and fitness services to share their data with the new Health app	848
google-fit	Google Fit is an open platform that lets users control their fitness data.	665
medical	The Medical tag relates to coding solutions in the field of medicine.	286
google-fit-sdk	An open platform that lets users control fitness data. Google Fit lets developers build smarter apps	225
researchkit	ResearchKit is an open source software framework that makes it easy to create apps for medical research or for other research projects.	114
hkhealthstore	The HealthKit store acts as a link to all the data managed by HealthKit.	89
niftynet	NiftyNet is a TensorFlow-based open-source convolutional neural networks (CNNs) platform for research in medical image analysis.	69
withings	Withings is a company that makes hardware and software to track body metrics and health statistics.	41

[4] MeSH: Medical Subject Headings can be accessed by https://www.ncbi.nlm.nih.gov/mesh.

healthvault	Microsoft HealthVault helps you gather, store, use, and share health data	32
heartrate	Questions related to heart signals monitoring.	27
openehr	openEHR is an open standard in health informatics that describes the management and storage, retrieval and exchange of health data	24
carekit	CareKit is Apple's framework includes core modules for developing apps that help people better understand and manage their health.	18
mapmyfitness	A set of Fitness Developer API from mapmyfitness.com	14
google-health	Google-Health is a health record service provided by Google.	9
samsung-health	It provides SDKs to help developers and healthcare providers thrive in an environment that connects apps, devices, and services.	6
intersystems-healthshare	Questions about HL7 (Health Level Seven), CDA (Clinical Document Architecture), Caché or other HIT technologies	2
Total of Questions		5221
Total of Questions without Duplicates		4488
Total of Discussions		3470

Other tags were tested but did not return posts focused on eHealth. For example, fall, blood, mhealth, telehealth, telemedicine, mysignals, ecg, emg, body, glucose, oximeter, elderly, older, adult, fitness, treatment, caregiver, healthcare, care, hospital, disease, health-monitoring, kubernates-health-check, fiware-health.

3.2. Processing

After data collection, it was necessary to perform some preprocessing activities. These activities represent an essential role in the analysis process [THOMAS, HASSAN, and BLOSTEIN, 2014]. Since we are working with unstructured data, we use natural language processing (NLP) techniques to reduce noise and improve the results obtained with LDA. These include removing code snippets, non-ASCII characters, punctuations, and words with less than three characters. After, we use a set of 565 terms to stop word removal. Finally, the original stream of text was turned in tokens to apply a word stemming process. In this last step, we decided to use the Porter stemming algorithm [THOMAS, HASSAN, and BLOSTEIN, 2014].

3.3. Data Mining

Here, I will try to explain the process of executing the LDA and decisions...

3.4. Evaluation

Here, it is important to explain the process of evaluating the LDA results. Perhaps it would be interesting to assess unanswered questions to identify challenges manually...

4. Results and Discussion

The following are some preliminary views just to exemplify the objects that can be discussed in the article. We still need to better reflect on this.

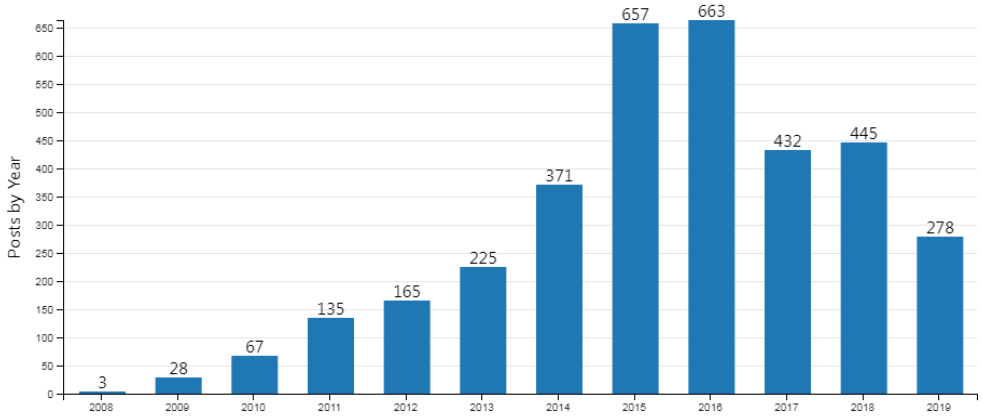

Figure 3. Questions over the years.

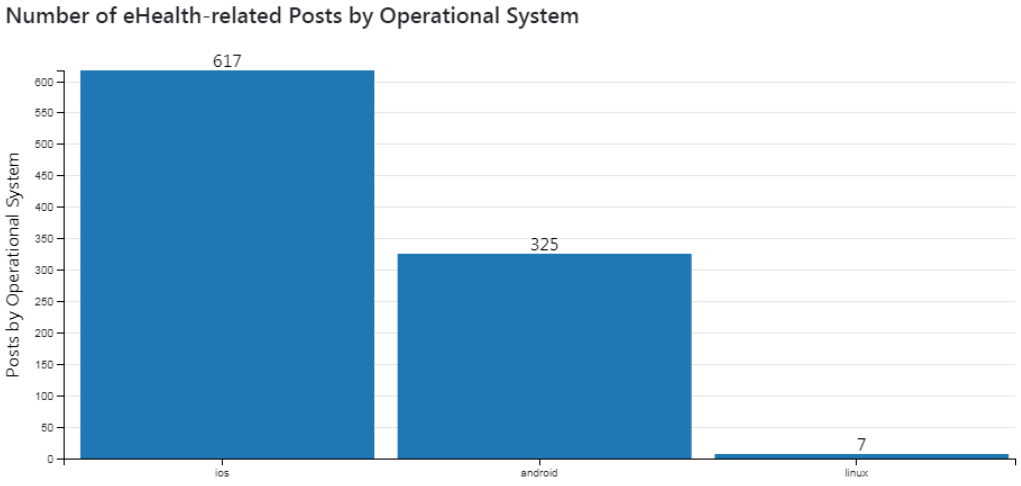

Figure 4. Questions by Operational System.

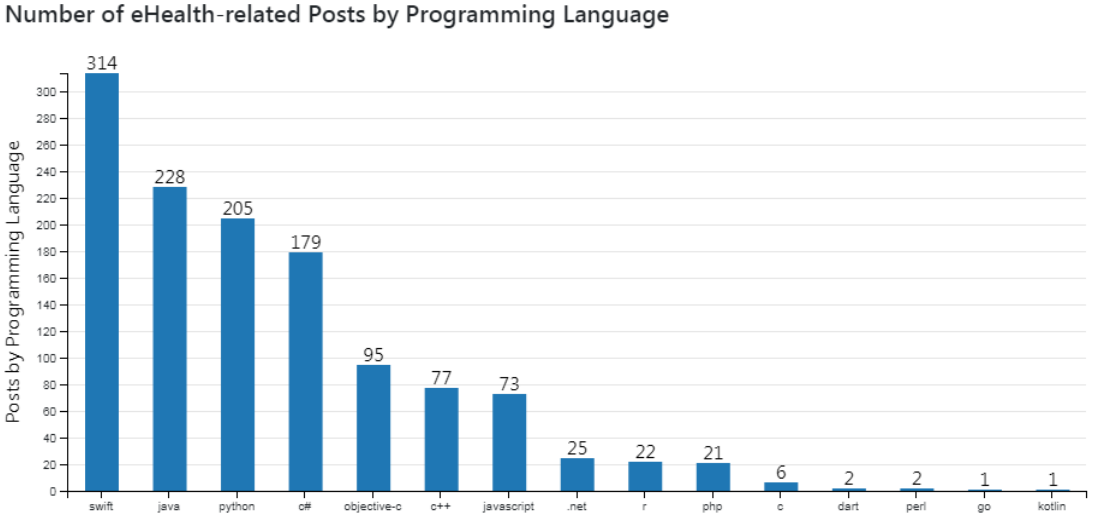

Figure 5. Questions by Programming Language..

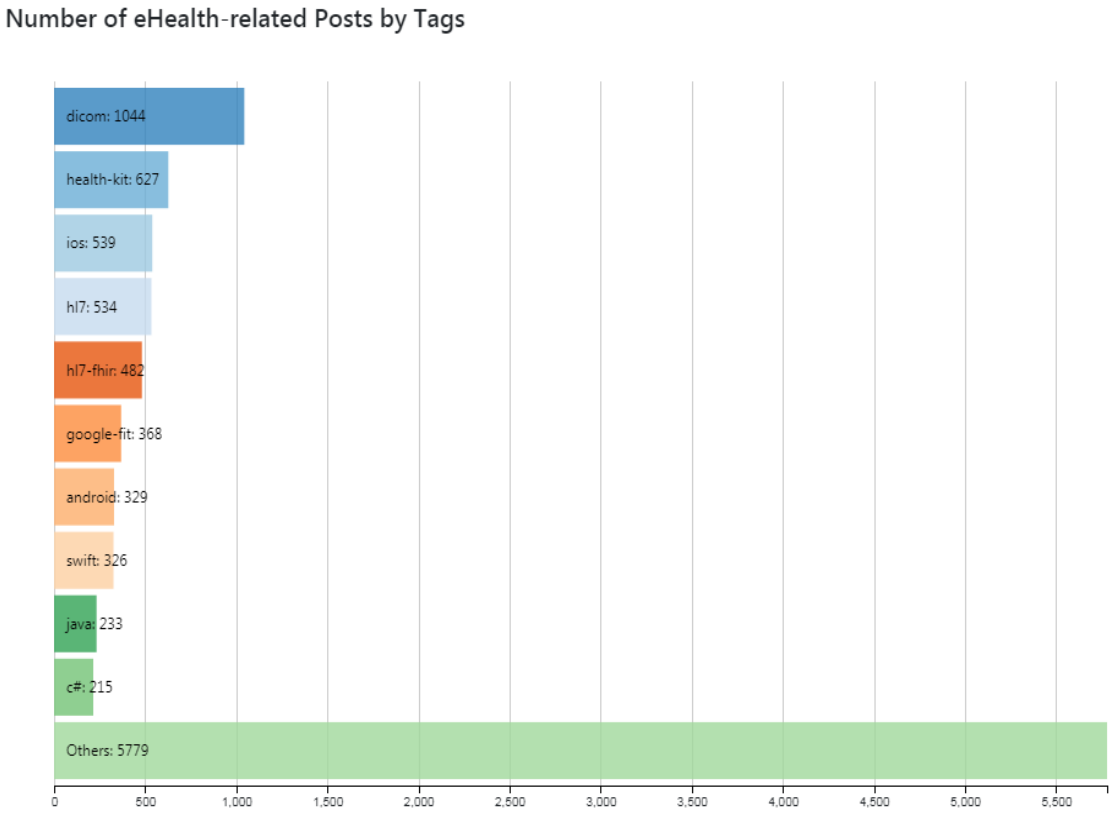

Figure 6. Questions by Tags..

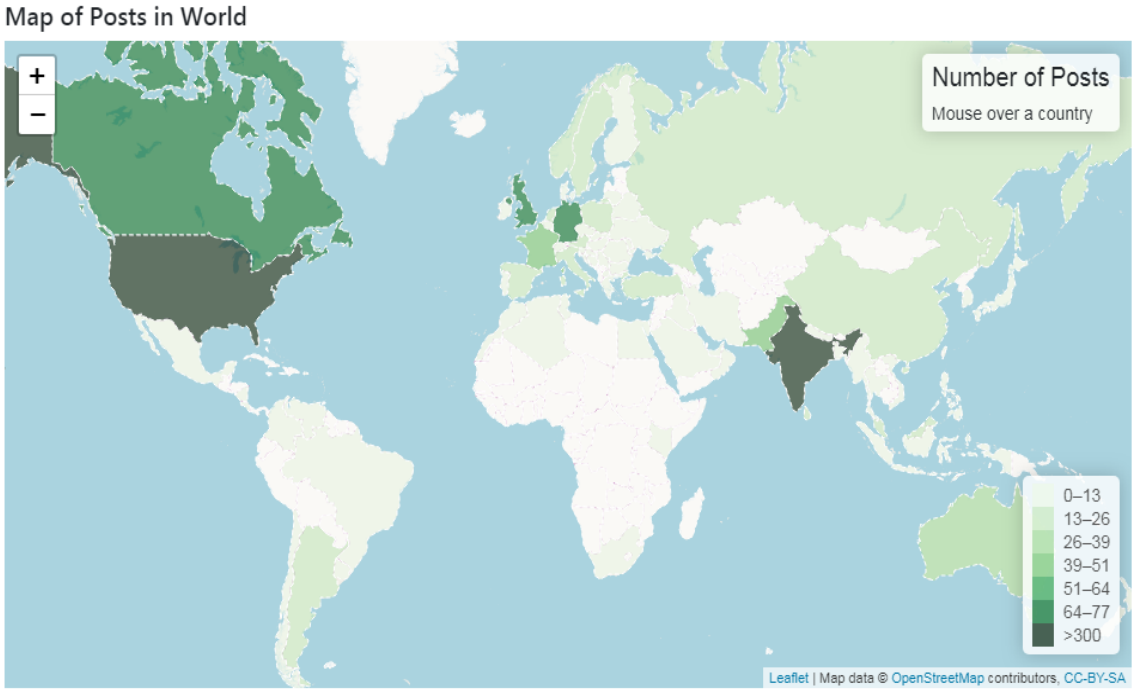

Figure 7. Distribution of Questions in World.

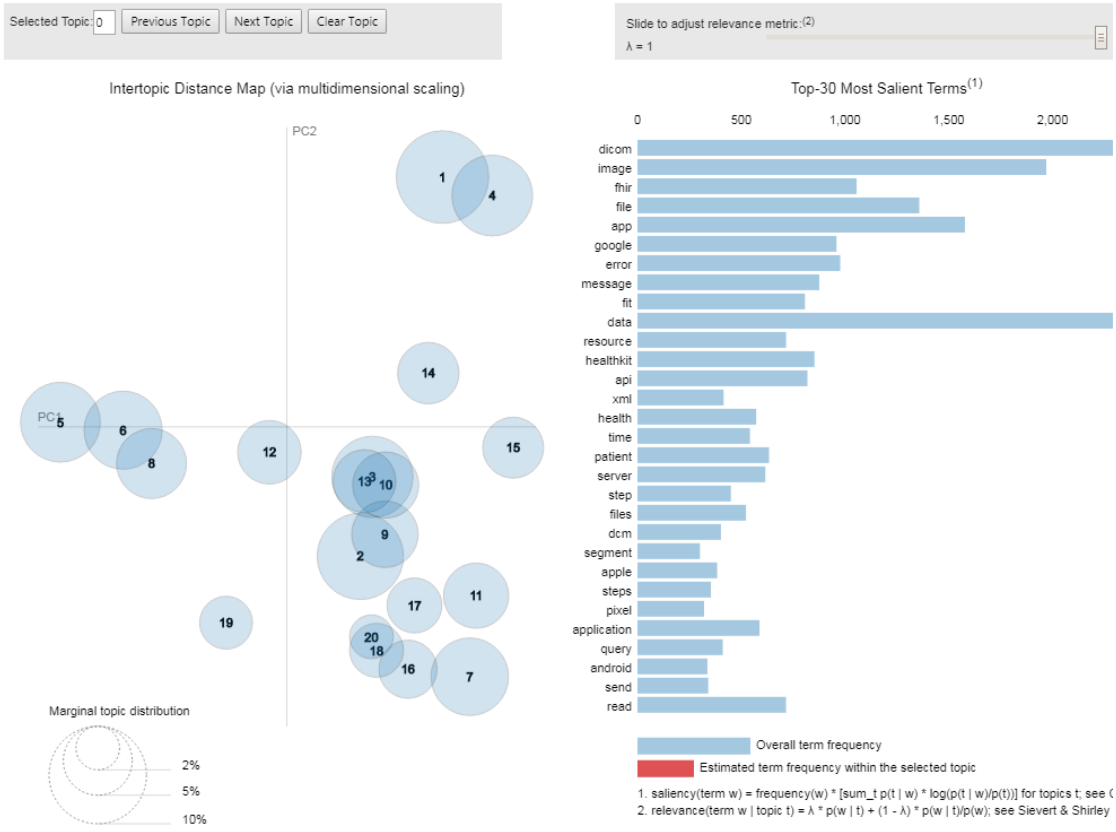

Figure 8. LDA clustering results.

Figure 9. Word cloud.

5. Validity Threats

In this work, we propose an investigation of the usage of data from smart objects (e.g., smartphones, smart...

6. Conclusion

In this work, we propose an investigation of the usage of data from smart objects (e.g., smartphones, smart...

7. Code availability

In order to improve the reproducibility of this work, all codes and data used are available on the Internet. The query to search the OS base posts can be accessed at the link: https://data.stackexchange.com/stackoverflow/query/1124873/ehealth-discussions. The codes used to build the visualizations presented in the article are available on Observable Notebook through the link: https://observablehq.com/@pedroalmir and the datasets can be found at: https://github.com/pedroalmir/datavis-course. All of these artifacts are licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s).

References

AIELLO, Luca Maria et al. Sensing trending topics in Twitter. IEEE Transactions on Multimedia, v. 15, n. 6, p. 1268-1282, 2013.

ALLAN, James. Topic detection and tracking: event-based information organization. Springer Science & Business Media, 2012.

BANDEIRA, Alan et al. We need to talk about microservices: an analysis from the discussions on StackOverflow. In: Proceedings of the 16th International Conference on Mining Software Repositories. IEEE Press, 2019. p. 255-259.

BLACK, Ashly D. et al. The impact of eHealth on the quality and safety of health care: a systematic overview. PLoS medicine, v. 8, n. 1, p. e1000387, 2011.

BLEI, David M.; NG, Andrew Y.; JORDAN, Michael I. Latent dirichlet allocation. Journal of machine Learning research, v. 3, n. Jan, p. 993-1022, 2003.

BLEI, David. Probabilistic topic models, 2012. Available at: http://icml. cc/2012/tutorials. Accessed in September 24, 2019.

BOSTOCK, Michael; OGIEVETSKY, Vadim; HEER, Jeffrey. D^3 data-driven documents. IEEE transactions on visualization and computer graphics, v. 17, n. 12, p. 2301-2309, 2011.

CHEN, Hui; COOGLE, John; DAMEVSKI, Kostadin. Modeling stack overflow tags and topics as a hierarchy of concepts. Journal of Systems and Software, v. 156, p. 283-299, 2019.

CHIARINI, Giovanni et al. mHealth technologies for chronic diseases and elders: a systematic review. IEEE Journal on Selected Areas in Communications, v. 31, n. 9, p. 6-18, 2013.

CULOTTA, Aron. Towards detecting influenza epidemics by analyzing Twitter messages. In: Proceedings of the first workshop on social media analytics. acm, 2010. p. 115-122.

DROSATOS, George; KAVVADIAS, Spiros E.; KALDOUDI, Eleni. Topics and trends analysis in ehealth literature. In: EMBEC & NBC 2017. Springer, Singapore, 2017. p. 563-566.

EYSENBACH, Gunther. What is e-health?. Journal of medical Internet research, v. 3, n. 2, p. e20, 2001.

FARAHANI, Bahar et al. Towards fog-driven IoT eHealth: Promises and challenges of IoT in medicine and healthcare. Future Generation Computer Systems, v. 78, p. 659-676, 2018.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From data mining to knowledge discovery in databases. AI magazine, v. 17, n. 3, p. 37-37, 1996.

GAGNON, Marie-Pierre et al. m-Health adoption by healthcare professionals: a systematic review. Journal of the American Medical Informatics Association, v. 23, n. 1, p. 212-220, 2015.

HEER, Jeffrey et al. A tour through the visualization zoo. Commun. Acm, v. 53, n. 6, p. 59-67, 2010.

ISLAM, SM Riazul et al. The internet of things for health care: a comprehensive survey. IEEE Access, v. 3, p. 678-708, 2015.

KEIM, Daniel et al. Visual analytics: Definition, process, and challenges. In: Information visualization. Springer, Berlin, Heidelberg, 2008. p. 154-175.

KUCHER, Kostiantyn; KERREN, Andreas. Text visualization techniques: Taxonomy, visual survey, and community insights. In: 2015 IEEE Pacific Visualization Symposium (PacificVis). IEEE, 2015. p. 117-121.

LIPSCOMB, Carolyn E. Medical subject headings (MeSH). Bulletin of the Medical Library Association, v. 88, n. 3, p. 265, 2000.

MCCALLUM, Andrew Kachites. Mallet: A machine learning for language toolkit. http://mallet. cs. umass. edu, 2002.

NGUYEN, Thin et al. Prediction of population health indices from social media using kernel-based textual and temporal features. In: Proceedings of the 26th International Conference on World Wide Web Companion. International World Wide Web Conferences Steering Committee, 2017. p. 99-107.

PAUL, Michael J.; DREDZE, Mark. Discovering health topics in social media using topic models. PloS one, v. 9, n. 8, p. e103408, 2014.

RAGKHITWETSAGUL, Chaiyong et al. Toxic code snippets on stack overflow. IEEE Transactions on Software Engineering, 2019.

ROBBINS, Timothy David; KEUNG, Sarah N. Lim Choi; ARVANITIS, Theodoros N. E-health for active ageing; a systematic review. Maturitas, v. 114, p. 34-40, 2018.

SAHAMA, Tony; SIMPSON, Leonie; LANE, Bill. Security and Privacy in eHealth: Is it possible?. In: 2013 IEEE 15th International Conference on e-Health Networking, Applications and Services (Healthcom 2013). IEEE, 2013. p. 249-253.

SUKHIJA, Nitin et al. Topic modeling and visualization for big data in social sciences. In: 2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC / ATC / ScalCom / CBDCom / IoP / SmartWorld). IEEE, 2016. p. 1198-1205.

THOMAS, Stephen W.; HASSAN, Ahmed E.; BLOSTEIN, Dorothea. Mining unstructured software repositories. In: Evolving Software Systems. Springer, Berlin, Heidelberg, 2014. p. 139-162.

TREUDE, Christoph; BARZILAY, Ohad; STOREY, Margaret-Anne. How do programmers ask and answer questions on the web?: Nier track. In: 2011 33rd International Conference on Software Engineering (ICSE). IEEE, 2011. P. 804-807.

ULLAH, Muhammad; FIEDLER, Markus; WAC, Katarzyna. On the ambiguity of Quality of Service and Quality of Experience requirements for eHealth services. In: 2012 6th International Symposium on Medical Information and Communication Technology (ISMICT). IEEE, 2012. p. 1-4.

WATTENBERG, M.; VIEGAS, F. Tag clouds and the case for vernacular visualization. Interactions, v. 15, p. 49-52, 2008.

WORLD HEALTH ORGANIZATION et al. mHealth: new horizons for health through mobile technologies. mHealth: new horizons for health through mobile technologies., 2011.

WU, Yuhao et al. How do developers utilize source code from stack overflow? Empirical Software Engineering, v. 24, n. 2, p. 637-673, 2019.