Ferramenta de Web-Scraping: Impactos da COVID-19 na Indústria de Software

Wiliane Maria A. S. de Souza[1], Wladimir F. T. Filho [2], Wylliams B. Santos[1]

[1]Universidade de Pernambuco (UPE) – Caruaru, PE – Brazil

[2]Escola Politécnica (POLI) – Universidade de Pernambuco (UPE)
Recife, PE – Brazil

{wiliane.maria,wbs}@upe.br, wftf@ecomp.poli.br

Abstract. *The pandemic caused by COVID-19 brings challenges to the software industry. To study this context, the use of Grey Literature is of great relevance, although there is a gap with regard to research tools for this content. This work presents information about a Web Scraping tool, developed for the search of Grey Literature in the Agile Software Development context. We use the Design Science method and show preliminary results composed by more than 7900 documents scraped by the tool. We expect to bring improvements to this application, which already proves useful to the research field.*

Resumo. *A pandemia causada pelo COVID-19 traz desafios para a indústria de software. Para o estudo deste contexto, a utilização da Literatura Cinza é de grande relevância. No entanto, há uma lacuna no que diz respeito a ferramentas de busca para tal conteúdo. Esse trabalho apresenta informações acerca de uma ferramenta de Web Scraping, desenvolvida para busca de Literatura Cinza no contexto de desenvolvimento ágil de software. Utilizamos o método de Design Science e mostramos resultados preliminares compostos por mais de 7900 documentos capturados pela ferramenta, a qual deve ser aprimorada futuramente e já se mostra útil no meio acadêmico.*

1. Introdução

Enfrentamos, desde janeiro de 2020, uma pandemia do vírus SARS-CoV-2, causador da COVID-19, anunciada pela Organização Mundial de Saúde [WHO 2020]. O vírus vem causando enormes impactos, sendo recomendadas medidas restritivas, que incluem o isolamento social, afetando diretamente diversas áreas, inclusive a indústria de software.

Acreditamos que as empresas da Indústria de Software vêm tomando atitudes para contornar as barreiras trazidas pela situação da pandemia, sendo interessante entender os problemas enfrentados e as boas práticas empregadas por elas. A Literatura Cinza (LC) pode ser uma fonte de investigação desse contexto.

A LC é definida como múltiplos tipos de documentos produzidos em todos os níveis governamentais, acadêmicos e empresariais em formato impresso ou eletrônico, desde que não sejam controlados por editoras comerciais [Schöpfel 2010].

Na literatura branca pode-se encontrar plataformas que suportam *strings* de busca, filtros e exportação dos resultados de maneira unificada. Porém, na literatura cinza, não existe ainda esse processo. Nosso objetivo, portanto, é desenvolver tal ferramenta com o

uso de técnicas de raspagem de dados (*Web Scraping*), em múltiplas fontes de dado, para implementação em uma aplicação web

2. Fundamentação Teórica

2.1. A Literatura Cinza na Engenharia de Software

O uso da LC em pesquisas acadêmicas no campo de Engenharia de Software (ES) vem se mostrando relevante. Nos últimos anos, estudos envolvendo LC vêm atraindo atenção de pesquisadores do campo de ES [Neto et al. 2019].

Uma Revisão de Literatura *Multivocal* (MLR) é descrita em [Garouse et al. 2016] como uma forma de Revisão Sistemática de Literatura, a qual inclui Literatura Cinza em adição à Literatura Formal. Ele também afirma que um uso em potencial para MLRs é no fechamento da lacuna entre a academia e a prática profissional e conclui que poucas MLRs foram publicadas no campo de SE até agora, destacando a necessidade de produção de mais trabalhos de MLRs. Dessa forma, é evidente o importante papel da LC, trazendo a necessidade da criação de ferramentas que auxiliem o processo de busca dessa literatura.

2.2. Raspagem de Dados

A técnica de raspagem de dados vem sendo utilizada em estudos relacionados a SE. Em [Georgiou et al. 2020] é realizada raspagem de dados na plataforma StackOverflow, utilizando a linguagem de programação Python e a biblioteca Selenium. No trabalho [Fiallos et al. 2019] é utilizado *Web Scraping* para coleta de dados do Reddit, fazendo uso das bibliotecas Selenium e BeautifulSoup, na linguagem Python. Assim, a raspagem de dados se mostra uma prática útil para a coleta de dados. As tecnologias usadas nesses estudos são consideradas no presente estudo.

3. Metodologia

Neste trabalho é utilizado o método de pesquisa *Design Scien*ce (DS), que, segundo [Wieringa 2009], enfatiza a conexão entre conhecimento e campo prático. Ele afirma que, ao produzir coisas úteis, estamos também produzindo conhecimento científico. Esse método consiste em: 1) Investigação do problema; 2) Desenho da solução; 3) Desenho da implementação; e 4) Implementação da solução. A primeira etapa consiste na identificação da problemática, abordada na introdução. As demais etapas são brevemente descritas a seguir.

O desenho da solução é composto pelo desenvolvimento da ferramenta de *Web Scraping*, com finalidade de monitoramento e coleta de dados, os quais estão sendo utilizados em uma revisão de literatura cinza, visando compreender os impactos da COVID-19 na indústria de software.

Para o desenvolvimento da ferramenta são utilizados como fontes websites relacionados ao contexto ágil de software: Scrum.org, AgileConnection.com e AgileAlliance.org. Através de técnicas de *Web Scraping*, utilizando a linguagem Python e as bibliotecas Selenium e BeautifulSoup, são capturadas todas as publicações existentes nas fontes, as quais são armazenadas em banco de dados. O banco de dados em questão é utilizado na aplicação *web* (figura 1), cujo *Back End* utiliza Python e Flask para implementar filtros de acordo com as opções selecionadas no *Front End*. Os filtros definem: um intervalo de data, tipos, fontes desejadas e *string* de busca, que suporta

expressões com ou sem parênteses utilizando os operadores lógicos “AND” e “OR”. Por sua vez, o *Front End* utiliza JavaScript e o framework React.js. Ele captura as opções de busca do usuário e as envia à API, a qual retorna os resultados, que são exibidos em tabela HTML, podendo ser exportados nos formatos xlsx ou csv.

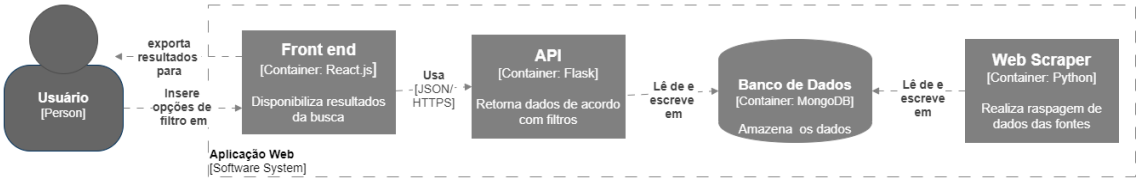

Figura 1. Diagrama de container C4 da Aplicação Web.

O desenho da implementação e a Implementação da Solução compreendem etapas futuras, responsáveis pela disseminação do conhecimento produzido. Na primeira, pretende-se aplicar as boas práticas, que foram identificadas na etapa anterior, em empresas de desenvolvimento de software do agreste pernambucano e observar suas adaptações. Na etapa de Implementação da Solução, o objetivo é aplicar, novamente, as boas práticas, juntamente com as adaptações observadas, até que um conjunto de boas práticas seja observado pelas empresas.

4. Resultados Preliminares da Ferramenta de Raspagem de dados

A tabela 1 exibe o total de postagens capturadas e armazenadas em banco de dados. A versão de testes pode ser acessada por meio do endereço https://search-for-grey.netlify.app/.

Tabela 1. Número de dados capturados pela ferramenta de Web Scraping.

Dados por fonte	
Fonte	*Dados capturados*
agilealliance.org	2618
agileconnection.com	2988
scrum.org	2337
total	7943

A *string* de busca (Figura 3) apresenta resultados preliminares de uma pesquisa de mestrado que realiza uma revisão de literatura *multivocal*, que foi aplicada à ferramenta e retornou 59 resultados. Tais resultados estão sob análise, sendo incluído um subconjunto durante etapa inicial de classificação dos critérios de inclusão e exclusão.

("software development" **OR** "software Project" **OR** "software engineering" **OR** "software team" **OR** "project management" **OR** "project manager" **OR** "Agile" **OR** "Extreme Programming" **OR** "XP" **OR** "Lean Software Development" **OR** "SCRUM" **OR** "Kanban") **AND** ("SARS-CoV-2" **OR** "COVID-19" **OR** "Coronavirus" **OR** "2019-ncov")

Figura 3. String de busca

5. Conclusões e Considerações Finais

A ferramenta apresentada neste trabalho mostrou resultados quantitativos relevantes, com 7943 documentos armazenados em banco de dados. Ela permite qualquer pessoa utilizar *strings* de busca a fim de coletar Literatura Cinza de forma eficaz. A ferramenta continua rodando em um servidor online, de maneira contínua, capturando novos dados diariamente, de forma a atualizar o banco de dados.

Pretende-se utilizar a ferramenta desenvolvida em trabalhos futuros que envolvam revisões de literatura cinza ou *multivocal*, coletando mais *feedbacks* acerca dela. É esperado também que ocorram aprimoramentos futuros. Além disso, é importante destacar que as técnicas de *Web Scraping* são necessariamente específicas para cada fonte, podendo ser necessária uma atualização do código, em casos de reformas nos *websites*. Outra expectativa é a de expansão do número de fontes utilizadas.

Agradecimentos

Os autores agradecem o apoio da Universidade de Pernambuco (UPE) e do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

Referências

Fiallos, A. and Jimenes, K. (2019) “Using Reddit Data for Multi-Label Text Classification of Twitter Users Interests”, In: 2019 Sixth International Conference on eDemocracy & eGovernment (ICEDEG).

Garousi, V., Felderer, M. and Mäntylä, M. V. (2016) “The need for multivocal literature reviews in software engineering”, In: Proceedings of the 20th International Conference on Evaluation and Assessment in Software Engineering - EASE ’16.

Georgiou, K., Mittas, N., Angelis, L. and Chatzigeorgiou, A. (2020) “A Study of Knowledge Sharing related to Covid-19 Pandemic in Stack Overflow”. arXiv preprint arXiv:2004.09495.

Neto, G. T. G., Santos, W. B., Endo, P. T. and Roberta Fagundes, A. A. (2019). “Multivocal literature reviews in software engineering: Preliminary findings from a tertiary study”, In: 2019 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM).

Organization, World Health. (2020) “WHO announces COVID-19 outbreak a pandemic”, http://www.euro.who.int/en/health-topics/health-emergencies/coronavirus-covid-19/news/news/2020/3/who-announces-covid-19-outbreak-a-pandemic. Acesso em: 02 abr. 2020.

Schöpfel, J. (2010), “Towards a Prague Definition of Grey Literature”, In: Twelfth International Conference on Grey Literature: Transparency in Grey Literature. Grey Tech Approaches to High Tech Issues.

Wieringa, R. (2009), “Design Science as Nested Problem Solving”. In: Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology - DESRIST ’09.