## COMPARAÇÃO ENTRE LINGUAGENS DE PROGRAMAÇÃO NO MERCADO, ENSINO SUPERIOR E PESQUISAS CIENTÍFICAS NA CIDADE DE SÃO PAULO

## COMPARISON BETWEEN PROGRAMMING LANGUAGES IN THE MARKET, HIGHER EDUCATION AND SCIENTIFIC RESEARCH IN THE CITY OF SÃO PAULO

Mayko Rodrigues Ribeiro da Costa 1 Me. Cleber Silva de Oliveira <sup>2</sup>

O mundo atual passa por rápidas mudanças tecnológicas. E, com isso, as tecnologias para desenvolvimento de sistemas são atualizadas constantemente. E por esta razão diversas linguagens de programação são criadas enquanto outras perdem relevância no mercado de profissionais em desenvolvimento de sistema. Por isso é importante criar formas de mensurar as tecnologias de desenvolvimento que são mais utilizadas pelo mercado de trabalho, pela produção de trabalhos científicos e pelas instituições de ensino. Com isso, pode-se comparar e saber se as instituições de ensino e os trabalhos científicos produzidos estão alinhados com as linguagens de programação que o mercado exige. Como nesse trabalho não foi encontrado base de dados sobre vagas do mercado de trabalho para desenvolvimento de sistemas. Foi decidido obter os dados sobre vagas direto dos sites de emprego. Dados sobre trabalhos científicos na API (Interface de Programação de Aplicativos) da IEEEXPLORE e dados sobre as instituições de ensino nas ementas ou projeto pedagógico. Assim pode-se comparar se as instituições de ensino e as produções de trabalhos acadêmicos acompanham as exigências do mercado de trabalho. Na comparação de dados foi descoberto que há divergência entre o que é ensinado com o que é exigido pelas empresas. Por exemplo PHP e Javascript são as linguagens mais exigidas pelo mercado pela amostra analisada. Mas o foco das instituições de ensino superior foram as linguagens C e C++. Desta forma foi possível analisar um período e contribuir com futuras pesquisas sobre o tema.

Palavras-chave: Linguagem de Programação. Mercado de Programação. Vagas em Programação.

Today's world is undergoing rapid technological changes. And with that, systems development technologies are constantly updated. And for this reason several programming languages are created while others lose relevance in the system development market. It is therefore important to devise ways to measure development technologies that are most commonly used by the labor market, the production of scientific work and educational institutions. With this, one can compare and know if the educational institutions and the scientific work produced are aligned with the programming languages that the market demands. As in this work was not found database on job vacancies for systems development. It was decided to get job data direct from job sites. Data on scientific work in the IEEEXPLORE API (Application Programming Interface) and data on teaching institutions in the menus or pedagogical project. Thus one can compare whether educational institutions and academic work productions match the demands of the education market. job. In comparing data it was found that there is a divergence between what is taught and what is required by companies. For example PHP and Javascript are the languages most demanded by the market for the analyzed sample. But the focus of higher education institutions was the C and C ++ languages. Thus it was possible to analyze a period and contribute to future research on the subject.

Keywords: Programming Language. Programming Market. Vacancies in Programming.

## 1 INTRODUÇÃO

O mundo atual passa por rápidas mudanças tecnológicas. E, com isso, as tecnologias para desenvolvimento de sistemas são atualizadas constantemente. Por esta razão diversas linguagens de programação são criadas enquanto outras perdem relevância no mercado de profissionais em desenvolvimento de sistema. Nesse cenário as instituições de ensino precisam manter-se atualizadas constantemente com as linguagens de programação que estão sendo mais exigidas pelo mercado. Senão haverá uma lacuna entre aquilo que é ensinado e aquilo que realmente o mercado precisa. (ABES, 2019).

Diante desse contexto, a problemática levantada, repercute de que maneira as linguagens de programação utilizadas no mercado, índices de popularidade e da academia podem contribuir nas decisões quanto à capacitação de novos profissionais ou mesmo estabelecer metas de autocapacitação?

Evidencia-se que no decorrer desta pesquisa não foi encontrada nenhuma base de dados que compare popularidade com o uso acadêmico e absorção regional de vagas de emprego na área de desenvolvimento

- 1 Especialização em Gestão de Sistemas da Informação. < mayko.costa@aluno.ifsp.edu.br >. 2 Especialização em Gestão de Sistemas da Informação. < cleber@ifsp.edu.br >.

de sistemas. Dessa forma, o objetivo deste estudo é mapear as tecnologias de desenvolvimento de sistemas, focando nas linguagens de programação mais utilizadas no mercado e na área acadêmica. Todavia, para atingir esse objetivo será necessário extrair os dados de sites de emprego, API (Application Programming Interface) sobre artigos científicos e Ementa ou Projeto Pedagógico das Instituições de Ensino Superior que obtiveram nota 5 pelo IGC (Índice Geral de Cursos do INEP ) no Enade de 2017.

Entre as prováveis hipóteses, pontua-se que hoje, a quantidade de linguagens de programação está crescendo em um ritmo cada vez mais rápido. Isso faz com que pressione outras linguagens de programação a ser cada vez menos usadas. Por isso as instituições de ensino superior devem revisar as linguagens ensinadas para manter os alunos atualizados com as mais relevantes pelo mercado. Outro fato é que a falta de atualização das linguagens de programação ensinadas pode comprometer a inserção do aluno no mercado de trabalho. Além de criar uma lacuna na falta de profissionais para ocupar estas vagas. E, com isso, a empresa terá maior custo na contratação de profissionais porque terá que treiná-los na linguagem de programação utilizada pela empresa.

As contribuições desse estudo são importantes tanto para sociedade quanto para o meio acadêmico, visto que há uma escassez de material sobre esse tema. Por isso é importante criar formas de mensurar as tecnologias de desenvolvimento que são mais utilizadas pelo mercado de trabalho, pela produção de trabalhos científicos e pelas instituições de ensino, podendo, assim comparar e saber se as instituições de ensino e os trabalhos científicos produzidos estão alinhados com as linguagens de programação que o mercado exige.

A metodologia utilizada para estruturação desse estudo será a pesquisa exploratória, realizado através do levantamento de dados. Segundo Gil (2008) essa pesquisa tem como finalidade selecionar uma amostra significativa de determinado universo, a fim de estaticamente apresentar informações acerca de comportamentos, crenças e opiniões, considerando suas vantagens e limitações.

#### **2 DESENVOLVIMENTO**

# 2.1 Abordagem sobre os processo de exploração de dados (CRISP-DM, KDD e SEMMA)

O processo escolhido para o experimento é o KDD (Knowledge Discovery in Databases) já que esse se adapta plenamente ao experimento. De acordo com Shafique (2014) é um processo mais flexível que os outros dois mais usados. Já o Semma não se adapta porque é um processo menos flexível.

Ainda de acordo com o autor, esse processo foi criado especificamente para ser usado com a ferramenta SAS Enterprise Miner software da empresa SAS. Já o CRISP- DM (Cross-Industry Standard Process for Data Mining) é uma metodologia criada a partir de uma concessionária de empresas que definiram as melhores práticas a partir de problemas e cases reais compilando um manual de boas práticas adotado amplamente pelo mercado. Todavia como o foco do experimento é descobrir novos conhecimentos acadêmicos nas

bases de dados e não em implementar regras de negócios ou implementar dentro do contexto de negócio optamos pelo KDD. Em ambiente de negócio existe a metodologia CRISP-DM que tem o foco em regras de negócios e resultados para negócios inclusive o guia explica detalhadamente o fluxo de processos entre as fases (SHAFIQUE, 2014).

Segundo apontam Shafique (2014) e Fayyad (1996) as fases do processo KDD são: Análise da base de dados, Preparação dos dados, Limpeza dos dados, Análise exploratória dos dados. Os dados extraídos não vem com as informações prontas. É necessário passar pelo processo de seleção dos dados, pré-processamento, transformação, descoberta de padrões e descoberta das informações. É um ciclo que permite analisar os dados de forma que consiga extrair apenas os dados necessários para a descoberta da informação. Possibilitando remover dados excedentes e estruturá-los de forma que possa viabilizar o armazenamento em banco de dados e a posterior geração do conhecimento.

### **3 MÉTODO DE COMPARAÇÃO**

A forma como esse trabalho procede leva em consideração outros trabalhos já compilados e consolidados internacionalmente. Assim é feito a comparação das linguagens de Programação analisadas que estão sendo exigidas pelo mercado de trabalho para classificá-las conforme as exigências nas vagas de emprego disponibilizadas na rede mundial de computadores. Assim também são quantificadas as linguagens analisadas que estão sendo ensinadas nas instituições de nível superior coletando a ocorrência destas nas ementas ou Projetos Pedagógicos dos cursos. Por último, são mensuradas as linguagens analisadas nos trabalhos de pesquisa publicados. Assim pode-se classificá-las dependendo da ocorrência destas. Para validar o processo, foi explorado o "Estado da arte" das técnicas disponíveis no mercado para esse fim. Algumas das metodologias estudadas estão sendo descritas nesse trabalho.

# 3.1 Processo de coleta e mensuração de dados da pesquisa

Primeiramente foi definido que serão pesquisados dados no Estado de São Paulo conforme descrito no escopo desse trabalho, bem como coleta de vagas e informações sobre ementas de instituições de ensino de nível superior desse Município.

As instituições que terão as ementas utilizadas nesta pesquisa são aquelas que atingiram conceito 5 pelo Enade nos cursos voltados à Ciências da computação, Engenharia da Computação, Gestão Da Tecnologia Da Informação e Análise de Sistemas.

Essa coleta de vagas foi assim definida, porque os dados são muito volumosos e não há material válido para amparar a pesquisa em um contexto mais abrangente, além de estar fora da capacidade necessária para finalizar esse trabalho. De 2013 até 2017 foram contabilizadas 26475 vagas de emprego segundo (Seade, 2019).

Devido a falta de informações sobre vagas de emprego autônomas e sem vínculo empregatício

na pesquisa foi decidido por extrair os dados direto dos endereços eletrônicos disponibilizados pela rede mundial de computadores que divulgam vagas para programadores de sistemas. Assim poderá se ter acesso aos dados para mensurá-los e comparálos, mediante os dados extraídos das ementas das instituições de ensino de nível superior e pesquisas científicas disponíveis em portal especializado.

# 3.2 Método para mensurar a classificação das linguagens utilizadas por terceiros

O ranking da Redmonk leva em consideração dados do Github e do StackOverflow. O StackOverflow faz pesquisas online que os desenvolvedores respondem as questões e as classificações são efetuadas a partir das respostas coletadas. Em 2019 tiveram mais de 90 mil respondentes em 170 países. Tiobe utiliza o termo Linguagem Programming em 25 motores de buscas. O IEEE Spectrum rastreia 300 linguagens de programação em diferentes plataformas como Google Search, Google Trend, Twitter, GitHub, Stack Overflow, Reddit, Hacker News, CareerBuilder, Dice, IEEE Xplore Digital Library conforme endereço eletrônico https://spectrum.ieee.org/static/interactive-the-top-programming-languages-2019.

Como pode ser constatado há uma diferença entre os quatro índices porque cada pesquisa utiliza uma forma diferente para coletar os dados e mensurar os resultados. Além de que essas pesquisas levam em consideração dados internacionais e não reflete apenas a situação da região analisada.

#### 3.3 Ferramentas

O trabalho em sua primeira fase será a revisão da literatura, estudar a problemática e definir o escopo abordado e como os macros objetivos serão atingidos. Após será a pesquisa exploratória onde serão levantados os processos existentes na literatura e as técnicas que serão utilizadas para atingir os objetivos. Na terceira fase, serão mensurados quantitativamente os resultados, com vistas a realizar uma análise comparativa entre as linguagens que aparecem entre os 10 primeiros lugares nos três índices acima, são esses: Java, Javascript, Python, PHP, C#, C++, C, Ruby, Swift, VBA, Go e NodeJs. Com os dados será calculado o valor de cada termo que consta nos documentos a partir do cálculo TF-IDF (frequência do termo-inverso da frequência na extração do documento ).

Os termos foram classificados em ordem de relevância decrescentes no contexto dos documentos analisados e será comparado com a hipótese de pesquisa, analisado a partir do gráfico do tipo gráfico de barras. Após a explicação do resultado e a conclusão desse processo.

### 3.4 Fluxo de trabalho do Tika

A biblioteca Tika é muito complexa em sua totalidade. É composta por diversos analisadores de arquivo html (Hyper Text Transfer Protocol), pdf (Formato de Documento Portável), xml (Linguagem de Marcação Extensiva ao HTML). E extratores para os mesmos formatos. O próprio Tika esconde a complexidade destas diversas ferramentas compondo um "pacote único" denominado composite. Porém quanto mais complexa a tarefa mais desses componentes deverão ser configurados e usados fazendo com que o uso da ferramenta fique difícil. Mesmo que utilizar estas ferramentas de forma avulsa seria ainda mais complexo do que pelo Tika. O conceito principal do Tika é justamente unir estas ferramentas e abstrair a complexidade de análise dos mais diversos formatos de arquivos. Sendo abstraído o processo de detecção conforme o tika.org e (MATTAMNN, 2012).

#### 3.5 FullText no Mysql

O Mysql tem funções de pesquisa FULL-TEXT que facilita na mineração da base de dados e mensura a relevância de cada termo dentro do banco de dados para cada registro com o cálculo TF-IDF conforme documentação do Mysql. Assim comparando os nomes das linguagens de programação e a ocorrência destas nas bases de dados conforme direto na consulta enviada ao banco de dados conforme dev.mysql.

#### 3.6 Projeto do Extrator de dados

Como não foi encontrado nenhuma base de dados, API ou fonte de dados para a coleta. Foi necessário criar extratores de dados conforme cada fonte. Já que a estrutura em que os dados estão dispostos é diferente. Ementas ou Projetos Pedagógicos são arquivos PDF, vagas de empregos são páginas HTML e as informações sobre artigos científicos estão em uma API que envia os dados como XML. Para posteriormente guardar esses dados extraídos foi criado um banco de dados com as informações que são em comuns entre as três fontes.

### **4 RESULTADOS**

# 4.1 Resultados da Pesquisa por vagas de empregos Analisadas

As vagas de emprego analisadas foram coletadas de sites indicados nas primeiras páginas do Google.com que tinha o foco em atender o mercado de Tecnologia da Informação, principalmente a àrea de informática. Foi coletado entre os dias 29 e 30 de Novembro de 2019 um total de 6935 vagas de emprego. Alguns sites detém uma quantidade maior de vagas proporcional à outros nesta coleta. O site Adzuna detém 55,51 % das vagas coletadas, ou seja, mais da metade das vagas. A pesquisa mostra que o PHP é a linguagem de programação mais aparece nas exigências das vagas com 25,37% do mercado paulista, seguida por Javascript 22,70%. Como fica evidente pelos gráficos algumas linguagens concentram a maior parte das vagas desta coleta mostrando que PHP e Javascript ficam com 48,7%. Mas, em contrapartida, Go, Ruby, Swift e Python tem baixa ou nenhuma relevância dentro da amostra coletada.

Figura 1 – Gráfico comparativo de linguagens de Programação em Vagas de Emprego

Fonte: Próprio Autor

Na coleta de dados para o ensino superior ficou evidenciado que os valores estão distribuídos entre as primeiras 8 linguagens C++, C, C#, Javascript, Java, VBA, NodejS E Python. Ruby, Swift e Go não teve relevância nenhuma nesta amostra.

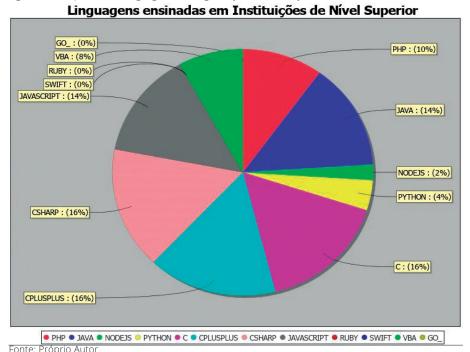
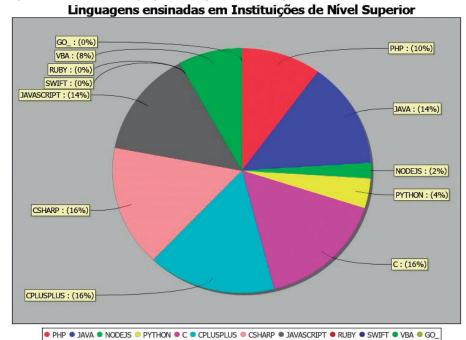


Figura 2 – Comparativo de linguagens de Programação em Instituições de Ensino

.

Nesta última coleta ficou evidenciado no quadro acima, a relevância das linguagens de programação ficaram concentradas nas linguagens C#, C++, C e Java. As outras linguagens ficaram abaixo da média tendo pouca relevância ou nenhuma de acordo com as amostras.

Figura 3 – Comparativo de linguagens de Programação em Artigos Científicos



-onte: Próprio Autor

### **5 MÉDIA GERAL**

# 5.1 Resultados da Pesquisa por vagas de empregos Analisadas

As vagas de emprego analisadas foram coletadas de sites indicados nas primeiras páginas do Google.com que tinha o foco em atender o mercado de Tecnologia da Informação, principalmente a àrea de informática. Foi coletado entre os dias 29 e 30 de Novembro de 2019 um total de 6935 vagas de emprego. Alguns sites detém uma quantidade maior de vagas proporcional à outros nesta coleta. O site Adzuna detém 55,51 % das vagas coletadas, ou seja, mais da metade das vagas.

A pesquisa mostra que o PHP é a linguagem de programação mais aparece nas exigências das vagas com 25,37% do mercado paulista, seguida por Javascript 22,70%. Como fica evidente pelos gráficos algumas linguagens concentram a maior parte das vagas desta coleta mostrando que PHP e Javascript ficam com 48,7%. Mas, em contrapartida, Go, Ruby, Swift e Python tem baixa ou nenhuma relevância dentro da amostra coletada.

Na coleta de dados para o ensino superior ficou evidenciado que os valores estão distribuídos entre as primeiras 8 linguagens C++, C, C#, Javascript, Java, VBA, NodejS E Python. Ruby, Swift e Go não teve relevância nenhuma nesta amostra.

Nesta última coleta ficou evidenciado no quadro acima, a relevância das linguagens de programação ficaram concentradas nas linguagens C#, C++, C e Java. As outras linguagens ficaram abaixo da média tendo pouca relevância ou nenhuma de acordo com as amostras.

De todos os dados já analisados até o momento. Extraímos um gráfico com a média simples de todos os dados e pode ser evidenciado que a linguagem de programação C# obteve a melhor média dentre todas as linguagens estudadas. As linguagens C, C++ e C# somadas possuem 49% do mercado. Ou seja, quase a metade da média.

### **6 ANÁLISE DOS DADOS**

A partir dos dados coletados fica evidente que as linguagens ensinadas nas Instituições de Nível Superior são diferentes das exigidas pelo mercado. Enquanto as linguagens mais exigidas pelo mercado foram PHP, Javascript e C#. As instituições de Ensino Superior e artigos de pesquisa científica teve maior relevância nas linguagens C, C++ e C#. Swift e Go não tem relevância dentro da amostra analisada sobre as vagas de emprego

analisadas, nem pelas instituições de Ensino Superior analisada mas teve uma baixa relevância em artigos de pesquisa científica analisados. Isso evidencia que a linguagem não tem espaço no mercado paulista e nem nas Instituições de Ensino Superior mas há pouco interesse em produzir contribuições acadêmicas à mesma dentro da amostra analisada. Não há evidências claras para tendências futuras por falta de dados históricos. Então esse estudo equivale à apenas aos meses de Outubro e Novembro. Conforme os gráficos se pode confirmar que C# está dentro da média nas três comparações. PHP não há contribuições relevantes nesse ano pela comunidade acadêmica do IEEEXPLORE mas se mantém bastante relevante nas Instituições de Ensino Superior situandose acima da média e muito relevante no mercado dentro das amostras analisadas.

#### 7 CONCLUSÃO

Esta pesquisa não explora todo o assunto. Apenas aborda de acordo com a metodologia citada que foi criada a partir da análise exploratória dos dados e levantamento bibliográfico sobre o assunto numa perspectiva de curto prazo do quadro atual das principais linguagens de programação dentro do escopo apresentado. Inclusive como proposta para futuros trabalhos a continuação desta mesma análise comparando dados coletados em períodos para comparar dados e a evolução histórica desses. Incluir outras tecnologias na pesquisa como frameworks, outras linguagens e mudar ou ampliar a área geográfica da pesquisa. Além de poder coletar de outras fontes de dados como Springer, Scielo, ResearchGate para artigos científicos. Trampos.co, Linkedin, e outras fontes de vagas para o mercado de trabalho. Outras Instituições de Ensino de Nível Superior de forma a ampliar o estudo e criar um quadro mais completo sobre as linguagens de programação.

Concluí-se que as questões que o trabalho respondeu as questões propostas e, a <u>partir</u> das respostas. Foi criado a análise do quadro atual das linguagens de programação apresentado.

#### **REFERÊNCIAS**

AMARAL, F. **Introdução à Ciência de dados:** Mineração de dados e Big Data. Rio de Janeiro: Alta Books, 2016.

BROWNEL, D. Sax2: **Processing XML Efficiently with Java.** Gravenstein Highway North, USA. O'Reilly, 2002.

CALDEIRA, D; DAYANE, C.M.F; et al. Mineração de Dados no LinkedIn para Definição do Perfil Profissional com a Metodologia MineraSkill. (2017). Disponível em: < https://www.researchgate.net/profile/Gabriel\_Spadon/publication/318413888\_Data\_mining\_on\_LinkedIn\_data\_to\_define\_professional\_profile\_via\_MineraSkill\_methodology/l inks/596a19650f7e9b8091945b4a/Data-mining-on-LinkedIn-data-to-define-professional-profile-via-MineraSkill-methodology.pdf?origin=publication\_detail>.Acesso em:5 nov. 2019.

DOUMOND, J.L. **Choosing the Right Graph** (2001). Disponível em: <a href="https://docplayer.net/storage/100/144128779/1575929412/4u0">https://docplayer.net/storage/100/144128779/1575929412/4u0</a> CwQ7DTiOle\_FCOj4a3A/144128779.pdf>. Acesso em: 7 nov. 2019.

FAYYAD, U., PIATETSKY-SHAPIRO, G., SMYTH, P. From Data Mining to Knowledge Discovery in Databases (1996). Disponível em: < https://doi.org/10.1609/aimag.v17i3.1230 >. Acesso em: 19.nov. 2019.

IEEXPLOREAPI.ORG. Disponível em: <a href="http://www.ieeexploreapi.ieee.org">http://www.ieeexploreapi.ieee.org</a> Acesso em: 3.nov. 2019.

INEP – INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA. Tabela com O IGC dos cursos em território nacional de 2017. (2019). Disponível em: < http://portal.inep.gov.br/web/guest/educacao-superior/indicadores-de-qualidade/resultados >. Acesso em: 18.nov. 2019.

JSOUP.ORG. Disponível em: < https://www./jsoup.org/ >. Acesso em: 7.nov.2019.

MATTMAN, C; et al. (2012) Tika in Action. Shelter Island NY. Manning Publications Co, 2012.

MYSQL.COM. Disponível em: < https://www.dev.mysql.com/doc/internals/en/full-text-search.html >, Acesso em: 2.nov. 2019.

POULAKOS, J. **Charts & Graphs:** Choosing the Right Kind of Chart or Graph. UETS Digital Media Group. (2005). Georgia United State, USA. Disponível em: <a href="https://www.perceptualedge.com/articles/ie/the\_right\_graph.pdf">https://www.perceptualedge.com/articles/ie/the\_right\_graph.pdf</a> Acesso em: 17. nov. 2019.

SEADE - SISTEMA ESTADUAL DE ANÁLISE DE DADOS. Número de vínculos, da ocupação Estado de São Paulo - 2013 à 2017. São Paulo. 2019. Disponível em: < http://www.seade.gov.br/profissoes/view/graficos.php?tipo=1&loc\_fam=3171%20 -%20T%E9cnicos%20de%20desenvolvimento%20 de%20sistemas%20e%20aplica%E7%F5es >. Acesso em: 5.nov. 2019.

SHAFIQUE, U. QAISER, H. A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA). International Journal of Innovation and Scientific Research, 2014, P. 12.

SPRINGERNATURE.COM. Disponível em: < https://www.dev.springernature.com/ >. Acesso em: 6 nov. 2019.

STACKOVERFLOW DEVELOPER SURVEY RESULTS. (2019). Disponível em: < https://insights.stackoverflow.com/survey/2019/ >. Acesso em: 18.nov. 2019.

TIKA.ORG. Disponível em: < https://www.tika.apache.org/. Acesso em: 9.nov. 2019.

TIOBE.COM. Disponível em: < https://www.tiobe.com/tiobe-index/ >, Acesso em: 12.nov. 2019.