Cientista de Dados: o que o mercado deseja

Conference Paper · October 2020

CITATIONS
0

READS
83

3 authors, including:

Marcos Nascimento
University of Brasília
2 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Ari Melo Mariano
University of Brasília
255 PUBLICATIONS **283** CITATIONS
SEE PROFILE

Cientista de Dados: o que o mercado deseja

Marcos Lopes do Nascimento
dept. Ciência da Computação
Universidade de Brsília
Brasília, Brasil
https://orcid.org/0000-0002-8995-303X

Vitor Sillos Alonso
dept. Engenharia de Produção
Universidade de Brsília
Brasília, Brasil
130137499@aluno.unb.br

Ari Melo Mariano
dept. Engenharia de Produção
Universidade de Brsília
Brasília, Brasil
https://orcid.org/0000-0002-7987-5015

***Abstract* - The purpose of this study is to identify the main skills required by the market for a data scientist. Companies need professionals with very different profiles and skills to be able to carry out activities related to the Big Data area. To identify these necessary skills, 7287 job descriptions were searched with the search term "Data Scientist" using web scraping techniques in "indeed.com" and "catho.com". The data was treated with the python programming language. An exploratory analysis of the data was performed, and the Latent Semantic Analysis (LSA) algorithm was applied. The most relevant terms were identified and skills such as teamwork, communication skills and understanding of the business are highly desirable by the market.**

Keywords - Data Science, job Description, python, LSA

I. INTRODUCÃO

Os dias atuais têm proporcionado para as empresas uma verdadeira avalanche de dados, decorrente do aumento do poder computacional e da crescente redução de custo de armazenamento, porém a capacidade de análise não acompanhou na mesma velocidade [1]. Para extrair valor desta grande quantidade de dados, chamada Big Data, as empresas precisam de pessoal altamente qualificado capaz de fazer as análises necessárias para trazer resultados ao negócio. Portanto, este estudo tem como problema de pesquisa: Quais são as habilidades de um profissional de ciência de dados para atuar com Big Data atualmente?

Compreender o perfil do profissional para lidar com a realidade do Big Data passou a ser de grande importância. [2] estudaram as principais características dos profissionais do Big Data, e encontraram que tais profissionais devem ter conhecimentos que vão além daqueles do campo da computação e estatística, como habilidades voltadas ao conhecimento do negócio, capacidades de contar histórias e manter um bom relacionamento com outros setores da empresa. Os autores definem, ainda, a característica principal destes profissionais como uma curiosidade que os leva a entender os dados, cruzando informações distintas da empresa e do mercado a fim de obter insights. Assim, o objetivo geral deste trabalho é identificar as habilidades do profissional de ciência de dados com base nos anúncios de emprego online.

II. MÉTODO

A metodologia deste trabalho tem duas fases principais [3]: coleta de dados e análise baseada em aprendizagem de máquina. Para que este trabalho possa ser útil para outros pesquisadores e interessados na área apresenta-se em detalhes os passos necessários e a abordagem analítica utilizada.

A. Coleta de Dados

A forma mais tradicional para contratação de empregados é por meio de anúncios de empregos. Na última década houve uma migração destes anúncios do jornal impresso para plataformas online, seja sites especializados ou redes sociais. Baseado nesta nova funcionalidade, este trabalho utilizou os sites 'Catho.com' e 'Indeed.com' como fonte de anúncios de empregos. O primeiro por ser o serviço líder em recrutamento na América Latina [4] e o segundo por ser o site de maior tráfego nos Estados Unidos [5]. Além disso, ambos têm acesso aberto, fornecem uma boa descrição dos empregos disponíveis e possuem o maior número de empregos disponíveis.

Os anúncios de emprego do trabalho foram selecionados, em ambos os sites, usando como termo de pesquisa '*Data Scientist*', selecionando todo anúncio que possuía no título o referido termo. Os dados foram coletados por meio de *Web Scraping* no dia 16 de outubro de 2019. Neste dia, um total de 7287 anúncios de emprego continham a descrição '*Data Scientist*'. Foram incluídas todas as ofertas de emprego que atendiam ao termo da pesquisa independente do idioma e da localidade, e excluídos aquelas cujas descrições apresentavam o mesmo corpo textual. Os procedimentos de coleta de dados, limpeza, tratamento e análise foram realizados por meio da linguagem de programação Python e as bibliotecas necessárias. Os códigos utilizados podem ser acessados na plataforma GitHub no link https://github.com/vsa-alonso/LSA-job-descriptions. Para a coleta de dados utilizou-se as bibliotecas Selenium (automação dos procedimentos) e BeautifulSoup4 (raspagem dos dados). Para o processo foi acionado um driver para conexão com o Google Chrome. Neste driver foram definidas as URL dos sites em questão, e assim o script criado encontra os corpos de descrição de emprego buscado no *header* da estrutura html. Os dados coletados formaram um corpo textual único salvo em um arquivo texto.

B. Análise de Dados

As análises foram feitas segundo a metodologia de *Latent Semantic Analysis (LSA)*. Esta técnica é capaz de identificar informações relevantes em textos, com identificação de palavras chave, atribuição de pesos à termos e representações baseadas em vetores derivados das ocorrências de palavras sobre os documentos [6]. O fluxo de trabalho tem três etapas principais [7]. Na primeira etapa ocorre uma coleta dos dados e um pré-processamento. Esta etapa envolve a remoção de documentos duplicados e termos que agregam pouca informação. É realizada uma análise exploratória de dados, onde são realizadas análises de ocorrências de palavras, identificação de uni-grams (termos de palavra individual), bi-grams (termos de duas palavras) e n-grams (termos de n palavras), bem como extração de outras informações. A terceira etapa consiste na análise e interpretação. Esta etapa, requer atenção especial, uma vez que ela demanda de análises estatísticas não padronizadas e - mais importante - da opinião de especialistas.

O desenvolvimento das análises se deu exclusivamente por meio da IDE Jupyter Lab. As bibliotecas utilizadas para realizar a limpeza e análise dos dados foram NLTK, string, scikitlearn e wordcloud. A biblioteca NLTK é uma ferramenta

para tratar linguagem natural, assim o texto foi decomposto em palavras únicas, calculado a frequência de cada palavra e o índice de relação entre palavras e removidas as *stopwords*. Em seguida, para manipulação e análise dos dados foi utilizado a biblioteca Pandas, finalizando com a interpretação. Para a realização das análises, foram aplicados procedimentos de Lemmatization e Stemming. Estes processos envolvem a remoção do sufixo das inclinações das palavras para sua forma base. Lemmatization é o processo de encontrar a forma normalizada de uma palavra [8]. No processo consideramos as palavras "working", "Works" e "worked", e suas formas alteradas para o infinitivo "work". Já o Stemming é utilizado para reduzir todas as palavras a uma forma comum. No exemplo anterior, para as variações do verbo "work", o algorítimo de Stemming reduziria o vocábulo para "wor", ou seja, é um processo que reduz todas as palavras a sua forma base, enquanto o anterior remove seu sufixo [9].

III. Resultados

Os resultados serão apresentados de acordo com as fases principais do trabalho: iniciando com uma análise exploratória dos dados discorrendo aspectos da fase de coleta de dados e pré-processamento obtendo as principais característica e em seguida verificando os resultados da aplicação do LSA sobre as descrições de emprego obtidas.

A. Análise Exploratória de Dados (AED)

A base de dados coletada a partir dos sites de procura de emprego Indee.com e Catho.com com o termo de busca "Data Scientist" resultou em 7287 descrições de emprego. Para uma primeira análise foram desconsideradas as divisões entre diferentes descrições de emprego. Dessa maneira considerou-se o arquivo como um único corpo de texto. As atividades descritas foram realizadas através da linguagem python, utilizando o notebook Jupyter Lab. Para o desenvolvimento do trabalho foram usadas as bibliotecas Pandas, Numpy, Nltk e Matplotlib. Segundo [10] a linguagem de programação Python oferece um rico aparelho de estruturas de dados de alto nível, possuindo poderosas bibliotecas e ferramentas para análise de dados, como Pandas e Numpy.

1) Uni-grams mais frequentes: Os uni-grams são as estruturas composta por uma palavra que aparecem com mais frequência. Na tabela I estão os 15 uni-grams principais, após os processos de Lemmatization e Stemming. O termo "data" possui a maior frequência, em parte por estar associado a outros termos pertinentes, tal como "big data", "data science", dentre outros. Já "experience" está relacionado a necessidade de o profissional ter experiência em diversos campos e habilidades. Os termos "learn" e "machine" que esperava-se que tivessem um número aproximado de ocorrências devido a estrutura "machine learning", apresentou "learn" com quase 9.000 ocorrências a mais. Isto ao fato da necessidade do profissional de aprender novas técnicas e ferramentas. A habilidade de trabalhar em equipe, foi evidenciado no termo "team". A capacidade de realizar modelagens sobressaiu nos termos "Science", "model" e "engineer". Já os termos "requir" e "product" fazem menção ao desenvolvimento de modelos voltados aos requisitos e peculiaridades dos produtos. Os termos subsequentes "build", "program", "analysis", "solution" e "software", estão altamente relacionados. Elas remetem à parte operacional das atividades do cientista de dados (CD), na qual serão desenvolvidos programas e soluções para auxiliar na tomada de decisão de diversas áreas.

2) Bi-Grams mais frequentes: Os bi-grams são as estruturas compostas por duas palavras mais citadas conforme tabela I. "Machine learn", "data scientist", "data scienc", e "comput scienc" estão muito relacionados às características gerais das vagas de emprego. Habilidades envolvendo machine learning são muito cobiçadas, possuindo o maior número de ocorrências. Outro termo, "deep learning" mostra a necessidade que as empresas possuem de empregar técnicas mais avançadas de machine learning. O termo "yea experi" traz a importância dada aos profissionais possuírem experiência anterior na área. Os Bi-Grams "commu skill" e "team member" denotam um ponto importante do trabalho de um CD, que envolve a interação com diversas áreas havendo a execução de tarefas multidisciplinares, para tanto o profissional deve ter capacidade de se comunicar e expressar pontos com clareza [11]. Alinha-se a este ponto os Bi-Grams "equal opportun", "nation origin", sexual orient" e "gender ident". As empresas estão preocupadas em montar suas equipes de maneira heterogênea, buscando pessoas com diversos históricos e visões, que já ficou evidenciado em pesquisas anteriores como uma vantagem na análise [12].

3) Tri-Grams mais frequentes: Os termos de 3 palavras mais frequentes podem ser vistos na tabela I. Com o uso de Tri-grams, é possível perceber que muitos dos termos apresentados são voltados para estabelecer o ponto de igualdade de oportunidades independente da origem, gênero ou orientação sexual do profissional, visto nos itens 1, 14, 15 e 16. Novamente apareceram termos relacionados ao aprendizado de máquinas, com os Tri-Grams 4, 5, 6, 7, 12, 17, ressaltando a necessidade de o profissional deter conhecimentos técnicos sobre machine learning. Devido aos maiores números de termos do Tri-Gram, começam a aparecer combinações que agregam pouco valor na identificação de habilidades demandadas por profissionais, como os Tri-Grams 2 e 5. Dessa maneira, a análise de sequências de mais termos não traria benefícios ao trabalho, sendo a próxima análise a word cloud.

4) Word-Cloud: Para compreender quais foram os termos mais pertinentes no texto e assim avaliar o corpo textual de maneira dinâmica, utilizou-se uma *word cloud* [13]. Para isso, utilizou-se a biblioteca *wordcloud*. A Fig. 1 apresenta o *word cloud* gerado.

TABELA I. Uni-Grams/Bi-Grams/Tri-Grams mais frequentes

	Uni-Gram	Bi-Grams	Tri-Grams
1º	data	machin learn	equal employ opportun
2º	experi	data scientist	locat one locationbenefit
3º	learn	data scienc	team data scientist
4º	develop	comput scienc	knowledg machin learn
5º	machin	deep learn	appli data scientist
6º	team	yea experi	network technolog data
7º	scienc	appli data	data store integr
8º	model	commu skill	store integr connect
9º	engin	team member	integr connect secur
10º	requir	equal opportun	connect secur connect
11º	product	pierc associ	secur connect platform
12º	comput	data analysi	understand network technolog
13º	includ	predict model	technolog data traffic
14º	opportun	team data	sexual orient gender
15º	analyt	big data	race color religion
16º	build	softwar engin	orient gender ident
17º	progr	nation origin	machin learn model
18º	analys	sexual orient	feder state local
19º	solut	gender ident	one locationbenefit health
20º	soft	basic understand	health insurancedent insurancevis

Na nuvem de palavras gerada há uma predominância de Bi-Grams em sua representação, ou seja, foram considerados os mais relevantes para a construção do modelo. O principal termo representado na figura é "machine learn", resultado alinhado com a análise dos Bi-Grams. Os termos "deep learn", "predict model", "neural networks" e "random forest" mostram uma demanda por conhecimento em modelos preditivos. A necessidade de trabalho em equipe e capacidade de comunicação foram evidenciados pelos termos "commun skill" e "team member". Alinhado assim, com estudo anterior em que a capacidade de contar histórias é uma habilidade fundamental, tendo em vista a necessidade de esclarecer informações, tendências e descobertas relevantes às partes interessadas [11]. Também é perceptível diversos termos que buscam esclarecer a questão de não haver quaisquer restrições relacionadas à origem, gênero ou orientação sexual da pessoa, podendo indicar o esforço das empresas para diversificar suas equipes em busca de uma visão mais holística do problema.

5) Habilidades Específicas: Buscou-se identificar quais conhecimentos são interessantes aos profissionais em especial as linguagens de programação e a computação em nuvem. *Linguagens de Programação:* As linguagens de programação em maior demanda ficaram da seguinte forma: Python (37,12%); R (33,37%); Java (9,09%); C++ (6,63%); C (4,88%) e Matlab (4,60%). Observa-se uma predominância de Python e R, totalizando 73,89%. Explicado pela facilidade de progração e base estatística para análise dos dados. *Cloud Computing:* As ferramentas em nuvem com maior demanda são: AWS (23,23%); Azure (21,20%); IBM (1,4%) e Cloud (54,17%). Aqui também, existe uma forte predominância dos serviços AWS e Azure. Importante destacar que o termo *cloud* teve mais de 50% das observações consideradas, de forma que é necessário o domínio desta tecnologia, mesmo que não tratando de um serviço especifico.

B. Latent Semantic Analysis (LSA)

Com a realização da análise exploratória de dados foi possível conhecer os dados existentes no corpo textual, identificando termos mais comuns dentro de diferentes categorias. Porém, não foi possível identificar as relações entre os termos. No desenvolvimento da LSA primeiro foi gerada uma matriz TF-IDF (*Term Frequency – Inverse Document Frequency*) para então usar seu esquema de ponderação, que promove a ocorrência de termos raros e desconta a ocorrência de termos mais comuns [7].

Para obter a matriz TF-IDF primeiro foi realizada uma contagem de palavras por documento, gerando uma matriz de termos por documento. Sobre esta matriz será aplicado os pesos da matriz TF-IDF, assim termos que aparecem frequentemente em um documento, mas não aparecem no outro recebem um peso maior, pois assumimos que estes termos possuem maior significado. Como parâmetros desta

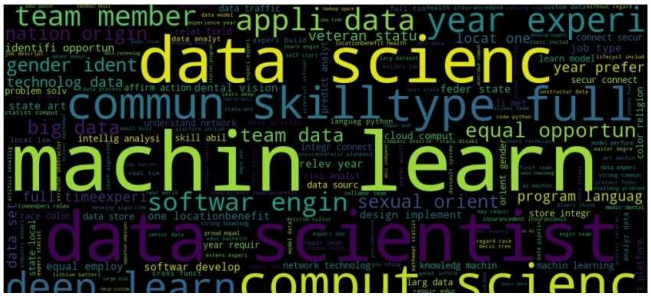

Fig. 1. Word cloud dos termos do corpo textual.

TABELA II. PRINCIPAIS TERMOS POR CLUSTERS

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
algorithm	technolog	busi	knowledg	ai
build	intellig	model	platform	engin
code	solut	scientist	secur	includ
engin	innov	data scientist	product	perform
system	analysi	opportun	tensorflow	network

transformação foi utilizada a maior frequência de termos em descrições a serem considerados de 80%, a frequência mínima considerada foi de 20% e foram considerados termos Uni-Gram, Bi-Grams e Tri-Grams. Esta etapa foi realizada utilizando-se a classe "TfidfVectorizer()" do módulo "sklearn.feature_extraction.text". Com a matriz TF-IDF pronta foi aplicado um algoritmo de K-means clustering para compreender a estrutura dos dados. Foram definidos 5 clusters para agrupar as descrições de emprego. O número de documentos dentro de cada cluster ficou distribuído da seguinte forma: cluster 1 (1226); cluster 2 (1211); cluster 3 (3649); cluster 4 (398) e cluster 5 (803). Com a utilização da classe "cluster_centers" do pacote "sklearn.externals.joblib" identificou-se os principais termos dentro de cada cluster. Estes termos são vistos na tabela II.

O cluster 3, que contém o maior número de observações, apresenta como primeiro termo "busi", relacionando-se ao business. Extrai-se disto a necessidade de o profissional possuir conhecimentos sobre o negócio. O cluster 1, apresenta termos voltados à programação, fazendo referência à necessidade do CD ter conhecimentos profundos sobre programação e criação de algoritmos para as tarefas. Já o cluster 2 evidencia a necessidade de que o profissional tome decisões baseadas em evidências, sendo capaz de analisar o cenário para construir soluções plausíveis [11]. Os clusters 4 e 5, trazem características gerais sobre as atribuições do profissional, apresentando termos relacionados ao aprendizado de máquinas, como tensorflow e ai, o que reforça a demanda por habilidades técnicas do profissional.

IV. DISCUSSÃO

Analisando os resultados obtidos, é perceptível identificar padrões sobre os termos levantados. O primeiro padrão refere-se aos repetidos termos sobre habilidades relacionadas a "machine learning", "deep learning", "big data", "predict model "e "data analysis". Dessa maneira, habilidades muito relacionadas ao aprendizado de máquinas e análises de dados, envolvendo todo o conhecimento e ferramentas relacionados. Outro grupo pertinente de habilidades está relacionado ao campo da ciência da computação, com termos como "computer Science", "software engineer" e "programing language", denotando a relevância deste campo. A AED nos mostrou que as principais línguas de programação procuradas são Python e R. Outra habilidade identificada como fator primordial é o conhecimento do negócio, evidenciado nos termos business, "product", "solution" e "innovation".

Alinhado a este ponto alguns termos retratam a necessidade do trabalho em equipe para o profissional. Termos como "Team member" e "communication skills" são volumosos e reforçam a sua importância. [11] ao explicar o trabalho de um CD, enfatizou a necessidade de equipes multidisciplinares, a fim de obter a maior quantidade possível de informações dos dados. Um indicador disto é a frequente ocorrência de termos que buscam profissionais com diversos perfis. Termos como "national origin", "sexual orientation", "gender identity", dentre outros, sugerem uma busca por

equipes heterogêneas, a fim de trazer diversos pontos de vista para a empresa, enriquecendo as análises e obtendo insights que não ocorreriam de outra maneira.

Isto posto, vemos uma tentativa das empresas de obter um profissional que possua todas as características listadas por [11] para um CD. Segundo o autor, as empresas buscam por profissionais com características abrangentes, possuindo a capacidade de codificar, conhecimentos em estatística, conhecimentos relacionados ao negócio, seja um tomador de decisões baseado em evidências e possua boas habilidades de comunicação. A tabela III agrupa os termos mais pertinentes identificados ao longo das análises segundo as categorias de um profissional de ciência de dados elencadas por [11].

Segundo a figura, vemos que as empresas buscam por profissionais completos, que abordem todas as característica elencadas por [11]. Os profissionais devem ter conhecimentos abrangentes e a capacidade de realizar diversas tarefas, envolvendo conhecimentos de programação, conhecimentos aplicados ao negócios e comunicação, dentre outros, ratificando [11]. Observando-se este complexo perfil exigido, torna-se evidente a necessidade de aumentar a participação de profissionais de diversas áreas, diversificando a experiência nas equipes e, portanto, enriquecendo as análises realizadas para obter resultados mais consistentes ao negócio.

V. CONCLUSÃO

O problema deste trabalho foi compreender quais são as habilidades de um profissional de ciência de dados para atuar com Big Data atualmente. O resultado obtido foram os principais termos usados no corpo textual, sendo possível identificar características de interesse e agrupá-las dentro das categorias estabelecidas por [11] para os principais atributos de um CD.

Além disso, a pesquisa mostrou que trabalhar em um ambiente multidisciplinar é uma característica da área, que possuir não apenas habilidades voltadas ao desenvolvimento de aprendizado de máquinas, mas também habilidades comportamentais, que permitam trabalhar em equipe são tão importantes quanto aquelas técnicas. Dessa forma, o objetivo geral do trabalho de identificar as principais habilidades requeridas por um profissional da ciência de dados, através de padrões de discursos encontrados em ofertas de emprego foi alcançado, buscando, ainda, uma comparação com as características elencadas por [11]. Sendo o estudo, particularmente, útil aos que desejam atuar nesta área de Big Data, pessoal de recursos humanos e demais pessoas envolvidas com a formação de material humano e que desejam uma formação mais holística para os alunos.

TABELA III. CATEGORIAS E HABILIDADES DE DAVENPORT E DO HABILIDADES DO TRABALHO

	Categorias	Habilidades	Trabalho
1	Hacker	▪ Habilidade de escrever código ▪ Entender arquiteturas do Big Data	▪ Computer Science ▪ Programming Language ▪ Software Engineer ▪ Code ▪ Algorithm
2	Cientista	▪ Tomar decisões baseadas em evidências ▪ Improvisação	▪ Problem Solving ▪ Knowledge ▪ Engineer
3	Conselheiro de Confiança	▪ Habilidade comunicação ▪ Compreender o processo	▪ Communication Skills ▪ Team Member
4	Analista Quantitativo	▪ Análise estatística ▪ Aprendizagem de máquina e análise de dados não estruturados	▪ Machine Learning ▪ Deep Learning ▪ Product Model ▪ Data Analysis
5	Expert em Negócios	▪ Compreender o negócio ▪ Noção de onde aplicar o Big Data	▪ Business ▪ Product ▪ Solution ▪ Innovation

Este trabalho possuiu dois fatores limitantes. Primeiro, este trabalho utilizou as descrições das vagas de emprego encontradas em sites, porém empresas que não utilizam este serviço podem desejar um perfil diferente de profissionais de ciência de dados. Assim, uma sugestão a trabalhos futuros é o estudo dos perfis de profissionais destas empresas. O segundo fator limitante refere-se à aplicação do algoritmo clusterização k-means, não havendo uma segmentação clara das características para alguns clusters. Para futuras pesquisas seria interessante a reaplicação do modelo proposto, mas com outro algoritmo, comparando a acurácia e resultados. Outra possível limitação diz respeito ao período curto de coleta de dados, podendo faz a pesquisa por um período mais longo.

REFERÊNCIA

[1] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "The KDD process for extracting useful knowledge from volumes of data," *Communications of the ACM*, vol. 39, no. 11, pp. 27–34, 1996.

[2] T. H. Davenport and D. Patil, "Data scientist," Harvard business review, vol. 90, no. 5, pp. 70–76, 2012.

[3] J. L. Jimenez-Marquez, I. Gonzalez-Carrasco, J. L. Lopez-Cuadrado, and B. Ruiz-Mezcua, "Towards a big data framework for analyzing social media content," International Journal of Information Management, vol. 44, pp. 1–12, Feb. 2019.

[4] A. Cardoso, F. Mourão, and L. Rocha, "A characterization methodology for candidates and recruiters interaction in online recruitment services," in Proceedings of the 25th Brazillian Symposium on Multimedia and the Web - WebMedia '19, 2019, pp. 333–340, doi: 10.1145/3323503.3349541.

[5] E. Schonfeld, "Indeed Slips Past Monster, Now Largest Job Site By Unique Visitors | TechCrunch." https://techcrunch.com/2010/11/17/indeed-monster-largest-job-site/ (accessed Aug. 31, 2020).

[6] P. Wiemer-Hastings, K. Wiemer-Hastings, and A. Graesser, "Latent semantic analysis," in Proceedings of the 16th international joint conference on Artificial intelligence, 2004, pp. 1–14.

[7] S. Debortoli, O. Müller, and J. vom Brocke, "Comparing Business Intelligence and Big Data Skills," Business & Information Systems Engineering, vol. 6, no. 5, pp. 289–300, Oct. 2014.

[8] J. Plisson, N. Lavrac, and Dr. D. Mladenić, "A rule based approach to word lemmatization," *Proceedings of the 7th International Multiconference Information Society (IS'04)*, pp. 83–86, 2004, [Online]. Available: http://eprints.pascal-network.org/archive/00000715/.

[9] T. Korenius, J. Laurikkala, K. Järvelin, and M. Juhola, "Stemming and lemmatization in the clustering of finnish text documents," in *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, 2004, pp. 625–633.

[10] S. van der Walt, S. C. Colbert, and G. Varoquaux, "The NumPy Array: A Structure for Efficient Numerical Computation," *Computing in Science & Engineering*, vol. 13, no. 2, pp. 22–30, Mar. 2011.

[11] T. Davenport, *Big Data at Work: Dispelling the Myths, Uncovering the Opportunities*. Harvard Business Review Press, 2014.

[12] H. Sword, M. Blumenstein, A. Kwan, L. Shen, and E. Trofimova, "Seven ways of looking at a data set," *Qualitative Inquiry*, vol. 24, no. 7, pp. 499–508, 2018.

[13] G. Tibaná-Herrera, M. T. Fernández-Bajón, and F. de Moya-Anegón, "Mapping a Research Field: Analyzing the Research Fronts in an Emerging Discipline," in *Scientometrics*, InTech, 2018, pp. 49–64.