



EXTRAÇÃO E ANÁLISE DOS DADOS DE PUBLICAÇÕES CIENTÍFICAS DA ÁREA DE BIOENGENHARIA COM AUXÍLIO DE TÉCNICAS DE TEXT MINING

André Luiz Carlota

andrecarlota@gmail.com

Jhonatan Fernando de Oliveira

jhonatan.oliveira@uemg.br

UEMG - Divinópolis, Avenida Paraná, 3001, Jardim Belvedere, 35501-170, Divinópolis, Minas Gerais, Brasil.

Thiago Magela Rodrigues Dias

thiagomagela@gmail.com

CEFET – MG - Rua Álvares de Azevedo, 400, Bela Vista, 35503-822, Divinópolis, Minas Gerais, Brasil.

Abstract. *Nas últimas décadas com a popularização dos meios de comunicação e disseminação de conteúdo cada vez mais massivo na Web, surge a necessidade de processos automatizados para descobertas de informações relevantes que envolvem o grande volume de dados criados a todo momento. Este trabalho tem por finalidade organizar e transformar os dados da plataforma PubMed/MEDLINE na área de bioengenharia de forma verificar os autores que mais contribuem com a área e as instituições envolvidas nestas publicações. Com o auxílio de algoritmos de busca, armazenamento de dados e processamento de texto será possível a extração, integração e processamento dessas informações. Desta forma serão extraídos dados das publicações científicas de forma geral (Revisão de Tema, Short Communication, etc...), especificamente da área de bioengenharia, realizando diversas análises a fim de entender os principais estudos (publicações) de pesquisa da área citada. Além disso, diversos outros estudos também serão realizados, como por exemplo, a elaboração de uma rede colaboração entre os autores, indicando os maiores especialistas da área em análise, em contrapartida obteremos as principais instituições onde as pesquisas estão sendo realizadas.*

Keywords: *Bioengenharia, Extração e processamento de dados, Mineração de texto.*

CILAMCE 2017

Proceedings of the XXXVIII Iberian Latin-American Congress on Computational Methods in Engineering
P.O. Faria, R.H. Lopez, L.F.F. Miguel, W.J.S. Gomes, M. Noronha (Editores), ABMEC, Florianópolis, SC, Brazil,
November 5-8, 2017.

INTRODUÇÃO

Segundo Woszezenki e Gonçalves (2013), com tecnologias inovadoras, a pesquisa é altamente dinâmica mesmo em campos consolidados, podendo sofrer mudanças a todo momento, impossibilitando que pesquisadores conheçam as últimas informações da área de bioengenharia em uma quantidade de tempo aceitável.

De acordo com Han e Kamber (2001), a ampla disponibilidade de imensas bases de dados, aliada à eminente necessidade de transformar tais dados em informação e conhecimento úteis para o suporte à decisão, têm demandado investimentos consideráveis da comunidade científica e da indústria de software.

A Plataforma PubMed/MEDLINE (Revista de Publicações Científicas da Área da Saúde) possui mais de 27 milhões de citações e publicações científicas na área de medicina, enfermagem, odontologia, medicina veterinária, biologia, bioquímica, evolução molecular, entre outros (NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION; 2017). Utilizando de processos mineração de texto será possível processar informações referentes a bioengenharia e organizá-las de maneira a separar e agrupar instituições, pesquisas relacionadas e trabalhos correlatos entre pesquisadores.

A mineração de texto, é um processo de descoberta de conhecimento, que utiliza técnicas de análise e extração de dados a partir de textos, frases ou apenas palavras (MORAIS & AMBRÓSIO; 2007). Este processo utiliza de algoritmos computacionais, que analisam/processam textos e abstrai informações úteis, estes não organizados de forma estruturada. Assim pode-se definir descoberta de conhecimento em texto como sendo processo de extrair padrões ou conhecimento, interessante e não triviais, a partir de documentos textuais.

Com a necessidade de transformar tais dados em informações uteis, surge a criação de processos automatizados auxiliares para suporte a decisão.

Os benefícios da mineração de textos podem estender-se a qualquer domínio que utiliza palavras ou texto, justificando-se a necessidade de se obter informações qualitativa que contenha alto valor agregado, sendo que suas principais contribuições estão relacionadas à busca de informações específicas em documentos, possibilitando melhor compreensão do

conteúdo disponível, de forma a alcançar seus objetivos e possibilitando a obtenção de resultados mais precisos.

A bioengenharia consiste na aplicação de conceitos e métodos de engenharia, baseando-se na resolução de problemas relacionados à saúde, esta agrega-se cada vez mais a tecnologia para realizar tarefas das mais variadas, como monitoramento de órgãos, equipamentos de precisão, operações, transplantes ou mesmo na busca por tratamentos alternativos (CARVALHO, A. C. A. *et al.*, 2010).

Com a mineração em textos das publicações da área de bioengenharia será possível obter informações relevantes de forma mais rápida, com o foco em melhorias dos processos já existentes. Em estudos relacionados como no trabalho de (CARVALHO, A. C. A. *et al.*, 2010) que visa na melhoria do aspecto estético de regeneração das células através da bioengenharia tecidual, a mineração de texto pode alavancar a obtenção de informações mais precisas e seguras sobre estudos de novas formas e meios para a resolução de problemas encontrados no dia a dia.

Este trabalho tem como propósito gerar uma rede de colaboração entre pesquisadores, que atuam na área de bioengenharia. Com isto será possível obter informações sobre o pesquisador que mais colabora com outros pesquisadores e as principais instituições onde ocorrem estas pesquisas.

METODOLOGIA

Os dados da plataforma PubMed/MEDLINE foram obtidos através de um extrator desenvolvido na linguagem de programação Python. Os principais motivos pela escolha da linguagem Python constam na praticidade de codificação e no alto poder de processamento, sendo esta uma linguagem de código fonte aberto o que facilita a obtenção de informações referentes a mesma (ROSSUM, 1991), e também por ser uma linguagem versátil, fantástica e de fácil entendimento (LAYTON, 2015). Através de recurso estudados no livro (LAYTON, 2015) foi possível a extração das informações.

Essa extração foi subdividida em 3 etapas para que fosse possível obter resultados apenas na área de bioengenharia e seus respectivos autores.

1. Foi necessário a extração dos identificadores de cada publicação desde a data de 01/09/1945 até 10/07/2017 relacionada a área de bioengenharia totalizando 193.830 registros.
2. Em seguida foram extraídas as publicações em acordo com os identificadores extraídos anteriormente.
3. E, por conseguinte foram extraídos autores e instituições relacionadas as publicações.

Para o armazenamento das informações adquiridas foi necessário a criação e utilização de um banco de dados relacional em linguagem SQL (*Structured Query Language* ou Linguagem de Consulta Estruturada) para posteriormente ser feita a análise destes dados. A Figura 1 exemplifica a estrutura de extração dos dados.

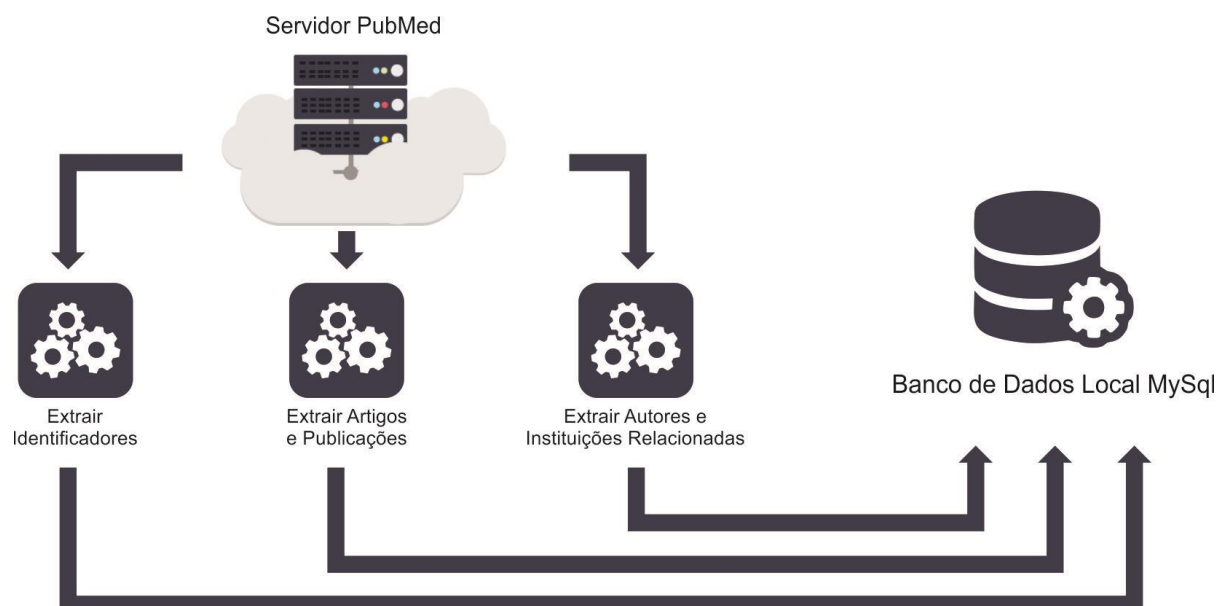


Figura 1 - Arquitetura da plataforma de extração das informações e armazenamento.

Com a obtenção dos dados foi necessário o desenvolvimento de uma estrutura paralela para remover “ruídos” e informações desnecessárias, sendo estas os artigos que não constam autores ou instituições vinculadas.

Em primeiro momento foram transformados todos os dados em letras maiúsculas para uma padronização das informações adquiridas e aplicado o algoritmo de distância de Levenshtein ou distância de dois “strings”.

O algoritmo de Levenshtein mede o número de operações necessárias para que uma palavra seja igual a outra, sendo essas operações de deleção, substituição ou inserção de caracteres (LEVENSHTEIN, 1966). Assim foi possível igualar registros que continham as mesmas informações escritas de forma diferente utilizando uma taxa de 10% de alteração necessário que para que uma palavra se igualasse a outra.

Com base nos dados coletados e normalizados foram realizadas as seguintes etapas para o desenvolvimento e verificação da rede de colaboração.

Seleção dos dados: consiste na definição de um filtro para os dados que deseja trabalhar, por exemplo, somente pesquisadores. Através desta definição, novos arquivos de pesquisadores e colaborações são gerados.

Arquivo de pesquisadores: como o foco do trabalho foi verificar os autores que trabalharam entre si e quais os que mais contribuíram com a área de bioengenharia foi necessário o desenvolvimento de outro algoritmo que analisasse os artigos que houveram colaboração e inserir estes em um outro tipo de arquivo CSV (*Comma Separated Values*).

Análise e processamento dos dados: Nesta etapa ocorre a geração das informações tabulares (*rankings* e matriz de colaboração) dos arquivos a serem importados no *Gephi*, sendo um arquivo de nós, relativo aos pesquisadores do presente estudo e um arquivo de arestas não direcionado (colaborações).

Mapa de colaborações: através do *Gephi*, ocorre a geração do grafo de colaborações entre os pesquisadores.

O *Gephi* é um software livre de código aberto e gratuito de exploração e visualização de todos os tipos de rede, mantido por um consórcio sediado na França com inúmeras aplicações em áreas como ciências biológicas, economia, exatas e etc... (JACOMY, 2009).

A Figura 2, mostra como foi o processo de análise, separação e geração da rede de colaboração através do *Gephi*.

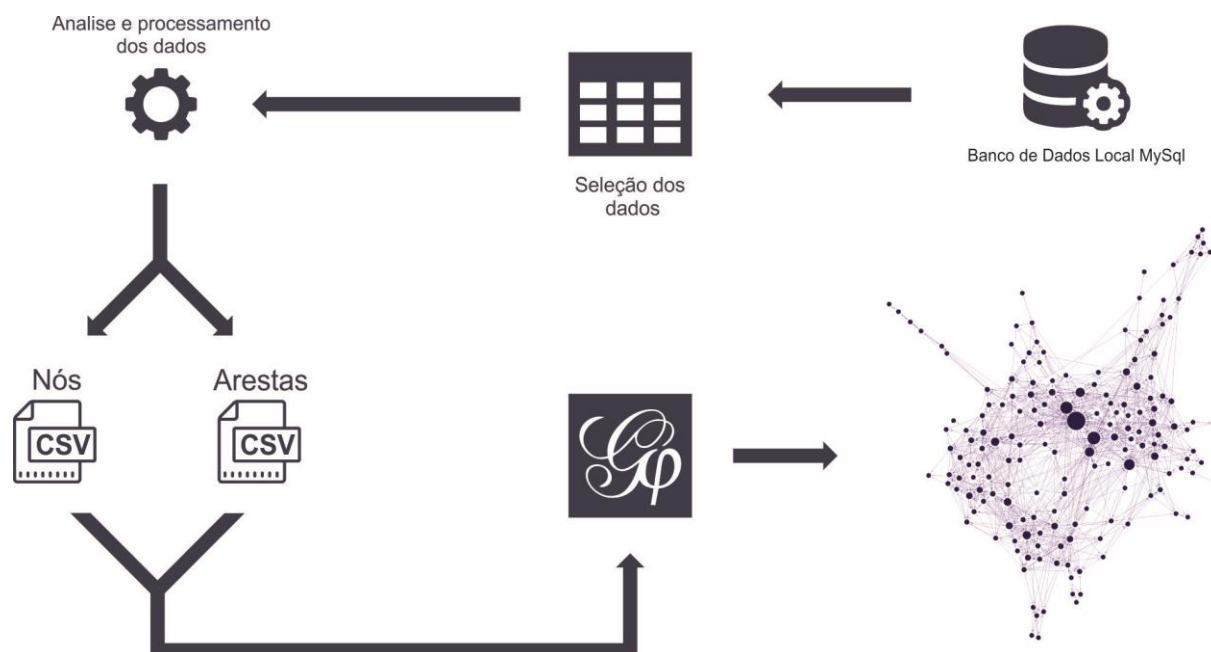


Figura 2 – Funcionamento da plataforma proposta.

RESULTADOS

Com a adoção da análise e mineração de texto em redes de colaboração entre pesquisadores, é possível identificar características que não são visualmente identificadas. Estas características são importantes porque podem revelar quais pesquisadores possuem mais afinidades, e quais pesquisadores podem estar agregando mais valor em pesquisas distintas.

Os resultados da análise dos dados extraídos da plataforma PubMed/MEDLINE apresentam a quantidade de publicações de um determinado pesquisador, e o grau de colaboração que ocorreu em determinadas publicações. A Figura 3, apresenta uma rede de colaboração entre pesquisadores.

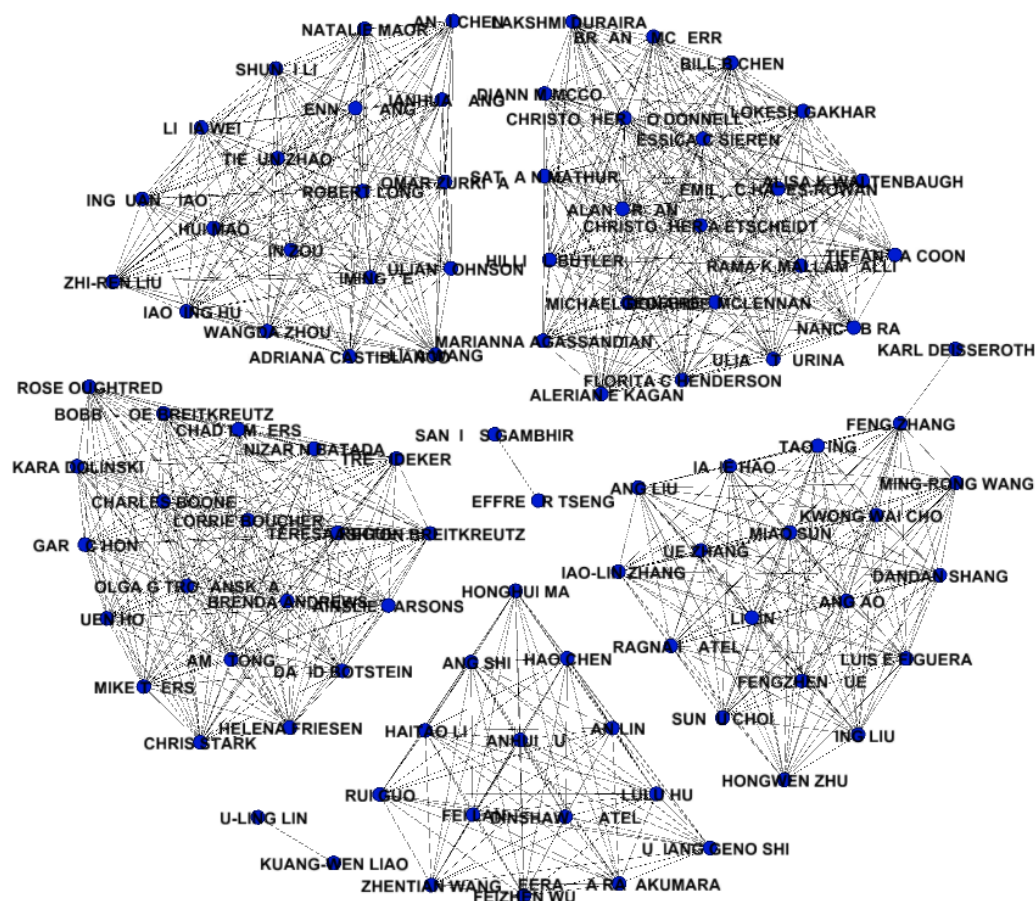


Figura 3 - Rede de pesquisadores onde ocorreu colaboração.

A Figura 3 mostra as publicações onde ocorreram mais colaborações entre os pesquisadores, sendo que cada nó representa um pesquisador e as arestas a ligação deste com outros pesquisadores. Nesta representação foi aplicado um filtro para mostrar os pesquisadores que mais colaboraram com a área de bioengenharia e com outros pesquisadores.

Tabela 1 - Autores que mais contribuíram com a área de bioengenharia e pesquisadores.

ID	Nº de publicações	Nome
1	9	H CHARLES MANNING
2	9	J J HOLST
3	9	JOSEPH M DESIMONE
4	8	ANDREA NATALE
5	8	C D A STEHOUWER
6	8	CATO T LAURENCIN
7	7	A K AZAD KHAN
8	7	A K BARBEY
9	7	ADIB KEIKHOSRAVI
10	7	ANGELO H ALL

Na Tabela 1 podemos ver quais os pesquisadores que mais contribuirão com a área de bioengenharia, sendo que em cada uma de suas publicações ele pode ter trabalhado com vários outros autores diferentes. Nesta tabela foram utilizados apenas os pesquisadores que colaboram com outros autores descartando registros de publicações onde não houve colaboração.

Outros dados que foram obtidos com a análise e mineração de texto das publicações foi a menção das instituições vinculadas aos pesquisadores, atentando-se para as que mais possuem pesquisadores colaborando com a área de bioengenharia e outros pesquisadores. A Tabela 2 nos mostra a instituição que mais teve participação nessas publicações.

Tabela 2 - Instituições que tiveram mais participação nas publicações.

ID	Nº de ocorrência	Instituição
1	10	INSTITUTE FOR MOLECULAR MEDICINE, DIVISION OF GYNECOLOGIC ONCOLOGY, UNIVERSITY OF BRESCIA, BRESCIA, ITALY.
2	9	CENTRE FOR EPILEPSY AND PARKINSON SURGERY OSPEDALE NIGUARDA CA' GRANDA, MILAN, ITALY.
3	8	FOUNDATION EPILEPSY INSTITUTES NETHERLANDS (SEIN), ACHTERWEG 5, 2103 SW HEEMSTEDE, THE NETHERLANDS.
4	7	DEPARTMENT OF BIOSCIENCE TECHNOLOGY, CHUNG YUAN CHRISTIAN UNIVERSITY TAOYUAN CITY 32023, TAIWAN.
5	7	DEPARTMENT OF BIOSCIENCE TECHNOLOGY, CHUNG YUAN CHRISTIAN UNIVERSITY TAOYUAN CITY 32023, TAIWAN.
6	6	CENTRE FOR EPILEPSY AND PARKINSON SURGERY A.O. OSPEDALE NIGUARDA CA' GRANDA, MILAN, ITALY.
7	5	PROJECT LABORATORY OF BIOMECHANICS AND TISSUE REPAIR, BIOENGINEERING COLLEGE, CHONGQING UNIVERSITY, CHONGQING 400044, CHINA; KEY LABORATORY OF BIORHEOLOGICAL SCIENCE AND TECHNOLOGY (CHONGQING UNIVERSITY), MINISTRY OF EDUCATION, BIOENGINEERING COLLEGE, CHONGQING UNIVERSITY, CHONGQING 400044, CHINA.
8	4	DEPARTMENT OF PHYSICAL THERAPY AND HUMAN MOVEMENT SCIENCES, FEINBERG SCHOOL OF MEDICINE, NORTHWESTERN UNIVERSITY, CHICAGO, IL 60611, USA.
9	3	IMAGE SCIENCES INSTITUTE, UNIVERSITY MEDICAL CENTER UTRECHT, 3508 GA UTRECHT, THE NETHERLANDS.
10	2	DEPARTMENT OF SURGERY, NARA MEDICAL UNIVERSITY , KASHIHARA, NARA , JAPAN.

Na Tabela 2 podemos verificar as instituições que mais possuem pesquisadores que contribuem com a área de bioengenharia. Para a geração dessa tabela foram analisadas as instituições nas quais os pesquisadores pertenciam, com isso é possível verificar e classificar de onde são os pesquisadores que mais publicam dados na área de pesquisa.

CONCLUSÃO

Em virtude dos fatos mencionados, a análise da rede de colaboração entre pesquisadores, pode-se obter informações bastantes relevantes quanto as publicações, e verificar o quão significativo estas foram para o meio científico.

O método proposto neste trabalho visa compreender o grau de colaboração entre pesquisadores, onde pode-se atinar que os pesquisadores que mais contribuem com outros pesquisadores são educadores da área de bioengenharia e afins. Já as instituições podem ter ou não relação com esses pesquisadores sendo que as instituições mencionadas são instituições de grande prestígio no meio acadêmico e científico.

Com a análise gráfica de pesquisadores e seus colaboradores obtemos informações relevantes onde esta colaboração foi benéfica e verificar o grau de aceitação destes trabalhos em meio científico tratando-se de uma maior gama de pessoas envolvidos.

Além disto é possível obter conhecimento sobre as instituições mais ligadas a área de pesquisa no meio de bioengenharia, assim como os pesquisadores que mais participam destas publicações sendo elas em colaboração com outros autores ou não.

REFERÊNCIAS

- CARVALHO, A. C. A. et al. Estratégias regenerativas da bioengenharia tecidual e aspectos éticos. **Revista de Ciências Médicas e Biológicas**, v. IX, n. 1, p. 20-27, 2010. ISSN 1677-5090.
- JACOMY, M.; M. B.; S. H. Gephi: An Open Source Software for Exploring and Manipulating Networks. **Association for the Advancement of Artificial**, Paris, France, p. 361-362, 2009.
- KOHN, H. C.; TAN, G. Data Mining Applications in Healthcare. **Journal of Healthcare Information Management**, v. XIX, n. 2, p. 64-72, 2005. ISSN 2. Acesso em: 5 jun. 2017.
- LAYTON,. **Learning Data Mining with Python**. Birmingham - Mumbai: Packt Publishing, 2015.
- LEVENSHTAIN, I. V. **Binary Codes Capable of Correcting Deletions, Insertions and Reversals**. [S.l.]: Soviet Physics Doklady, v. 10, 1966. Cap. 8, p. 707.
- MORAIS, E. A.; AMBRÓSIO, A. P. L. **Mineração de Textos**. Universidade Federal de Goiás. Goiás, p. 30. 2007. (INF_005/07).
- NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION. PubMed/MEDLINE. **PubMed**. Disponível em: <<https://www.ncbi.nlm.nih.gov/pubmed/>>. Acesso em: 1 jun. 2017.
- ROSSUM, G.. Python. **Python**, 1991. Disponível em: <<https://www.python.org/>>. Acesso em: 01 Janeiro 2016.
- WOSZEZENKI, C. R.; GONÇALVES, A. L. Mineração de textos biomédicos: uma revisão bibliométrica. **Revista Perspectivas em Ciência da Informação**, v. XIII, n. 3, p. 23-44, Julho/Setembro 2013.