Digerindo Artigos Acadêmicos e Autores

Paulo Henrique da Costa
São Paulo, 2018

Background

Eis o mapa mundo da produção científica!

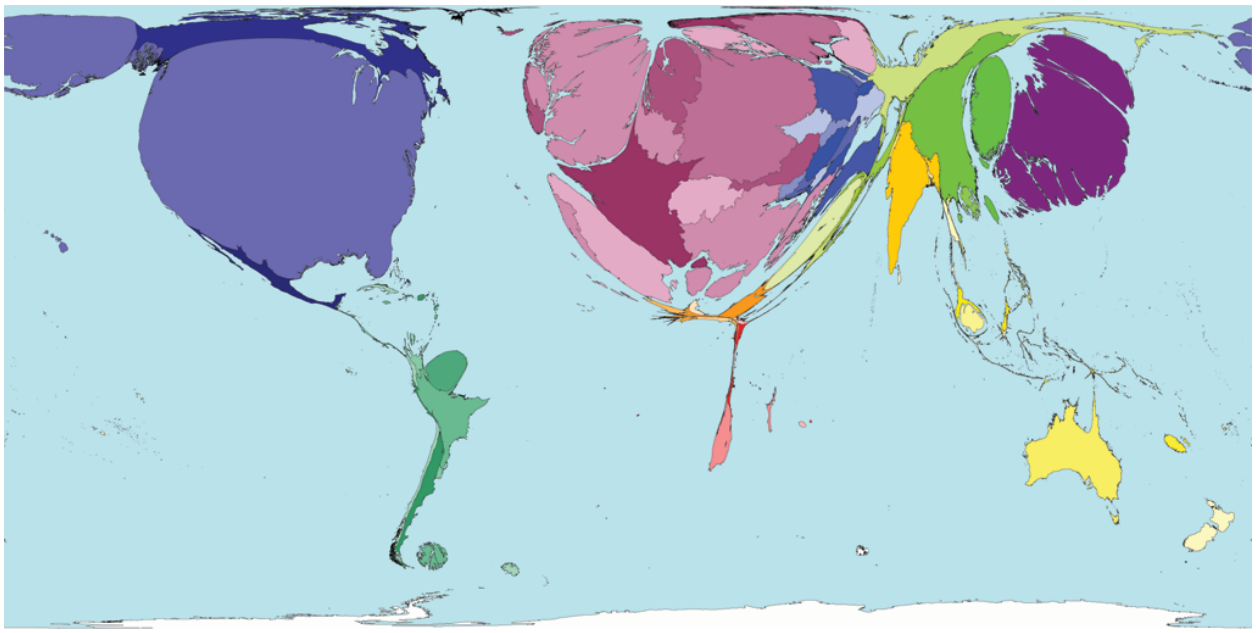

Figura 1 - Mapa mundo da produção científica [1]

Todos os dias pesquisadores, estudantes de graduação, mestrado, doutorado, curiosos estão descobrindo e escrevendo sobre nossa visão do mundo, artigos sobre astronomia, psicologia, informática, engenharia mecânica, entre outros, questionam e encontram informações sobre pequenas, grandes, estranhas, importantes ou inúteis descobertas. Se separarmos uma certa área específica, será que uma pessoa é capaz de ler, entender, conversar com os autores daquele assunto? Num mundo tão grande, com uma produção tão vasta, é difícil.

Em 2011, por exemplo, os pesquisadores brasileiros publicaram 49.664 artigos [2], olhando o mapa acima, talvez no mesmo ano, os Estados Unidos tenham publicado cerca de 20 vezes mais. Mas a disputa aqui não é sobre quem publica mais, mas sim, sobre seria possível consumir uma quantidade grande de produção científica.

O trabalho acadêmico tem por critério de qualidade, geralmente, a originalidade do trabalho, os pesquisadores citam em seus projetos descobertas originais, ou aplicação de métodos em novos tipos de dados, entre outras maneiras, porém de alguma forma estes pesquisadores estão ligados, por seus assuntos ou por métodos utilizados, o mundo científico também adotou a interdisciplinaridade, por vezes alguns artigos de medicina irão usar técnicas de ciências da computação ou matemática para comprovar uma descoberta, a arte pode usar a física, a física pode usar a psicologia.

Então se alguém escreve, ou se vai escrever sobre um assunto, quem são as outras pessoas da área dela, ou não, que estão produzindo sobre assuntos que dividem características com as essas publicações, se uma universidade ou revista quiser saber quais autores tem mais similaridade nos métodos usados, ou nos dados pesquisados, como unir essa informação sem ter que ler toda a volumosa produção científica que o mundo produz?

Problema e Justificativa

O projeto de pesquisa se concentra em realizar um agrupamento de dados, para que no final, resolva o seguinte problema:

Utilizando Processamento de Linguagem Natural e Aprendizado de Máquina será construído um algoritmo para agrupar artigos publicados em revistas científicas, e saber o que está mais sendo publicado nelas, sem ter que ler todas elas, e encontrar a revista com mais afinidade para aceitar um novo artigo.

Para atingir este resultado será necessário coletar uma quantidade de resumos de artigos que são disponibilizados em formatos de arquivos específicos, estes resumos possuem atributos importantes como o nome dos autores, a revista em que o artigo foi publicado e palavras chave, não esquecendo do próprio texto do resumo em si. O texto do resumo por sua vez deverá passar por um processamento de linguagem natural, para que seja possível abstrair informação útil.

Na saída do agrupamento do projeto espera-se como resultado uma tabela como o seguinte exemplo:

Título	Atributo 1	Atributo n	Revista	Grupo
Artigo 1	Valor 1	Valor n	Revista X	Grupo a
...	...	...	...	...

Tabela 1 - Exemplo de tabela de saída do algoritmo de agrupamento (clustering)

Com este resultado de agrupamento poderemos analisar quais grupos uma revista mais publica, como por exemplo, poderemos ter a seguinte análise:
“Na Revista X, 70% das publicações são do grupo a, 20% do grupo b, e 10% do grupo c.”

Por fim, utilizaremos um algoritmo para prever em qual grupo um novo artigo pertence, e assim analisar em que revistas este artigo tem mais probabilidade de sucesso de aceitação, sabendo quais grupos as revistas mais publicam. Este algoritmo poderia ser o mesmo utilizado para agrupar os dados, no entanto, para explorar mais ferramentas de aprendizado de máquina, será utilizado um algoritmo para classificação, utilizando como variável alvo o grupo encontrado na etapa de agrupamento.

Conjunto de dados e inputs

A produção científica, no geral, está bem organizada, a maioria dos sites que mantém biblioteca de resumos de publicações espalhadas por revistas ao redor do mundo, por exemplo, no site (https://www.ncbi.nlm.nih.gov/pubmed/), estão concentrados artigos do mundo todo sobre a área da saúde. Neste site, são disponibilizados resumos de todos os artigos, também é disponibilizado um arquivo de extensão “.nbib”, este arquivo contém informações gerais sobre o artigo, além de seu resumo.

Exemplo:

```
PMID- 29415229
OWN - NLM
STAT- Publisher
LR  - 20180207
IS  - 1930-613X (Electronic)
IS  - 0026-4075 (Linking)
DP  - 2018 Feb 5
TI  - Military Personnel Who Seek Health and Mental Health Services Outside the
```

Military.
LID - 10.1093/milmed/usx051 [doi]
AB - Background: Although research conducted within the military has assessed the health and mental health problems of military (etc …)
FAU - Waitzkin, Howard
AU - Waitzkin H
AD - Health Sciences Center and Department of Sociology, University of New Mexico, 801 Encino Place NE, Suite C-14, Albuquerque, NM 87102.
FAU - Cruz, Mario
AU - Cruz M
AD - Department of Psychiatry and Behavioral Sciences, University of New Mexico School of Medicine, MSC09 5030, 1 University of New Mexico, Albuquerque, NM 87131.
FAU - Shuey, Bryant
AU - Shuey B
AD - Department of Psychiatry and Behavioral Sciences, University of New Mexico School of Medicine, MSC09 5030, 1 University of New Mexico, Albuquerque, NM 87131.
FAU - Smithers, Daniel
AU - Smithers D
AD - Boston University School of Medicine, 72 East Concord St, Boston, MA 02118.
FAU - Muncy, Laura
AU - Muncy L
AD - Civilian Medical Resources Network, P.O. Box 2965, Taos, NM 87571.
FAU - Noble, Marylou
AU - Noble M
AD - Civilian Medical Resources Network, P.O. Box 2965, Taos, NM 87571.
LA - eng
PT - Journal Article
DEP - 20180205
PL - England
TA - Mil Med
JT - Military medicine
JID - 2984771R
OTO - NOTNLM
OT - Access
OT - Mental health
OT - Military
OT - Multi-method research
OT - Suicide
OT - War
EDAT- 2018/02/08 06:00
MHDA- 2018/02/08 06:00
CRDT- 2018/02/08 06:00
PHST- 2017/05/23 00:00 [received]
PHST- 2018/02/08 06:00 [entrez]
PHST- 2018/02/08 06:00 [pubmed]
PHST- 2018/02/08 06:00 [medline]
AID - 4838357 [pii]

AID - 10.1093/milmed/usx051 [doi]
PST - aheadofprint
SO - Mil Med. 2018 Feb 5. pii: 4838357. doi: 10.1093/milmed/usx051.

Descrição de alguns itens importantes:

- TI: Título do artigo;
- AB: Resumo do artigo;
- FAU: Autor do artigo;
- OT: Palavras chave;
- JT: Revista onde foi publicado;

Como não foi encontrada nenhuma API pública que disponibilizasse arquivos do tipo .nbib ou quaisquer outros arquivos similares, todos os arquivos serão coletados manualmente.
Serão coletados em torno de 300 resumos de artigos, todos do site especificado acima, no idioma inglês, para evitar erros de viés durante a elaboração dos algoritmos, todos os artigos deste website são sobre medicina, de modo que possamos reduzir o grupo de observação e não coletar dados extremamente distantes que não sejam efetivos no momento da construção do algoritmo.

Os dados deverão ser transformados para que seja possível o agrupamento, será adotado a transformação dos atributos obtidos na extração do resumo e nas palavras chave OT em atributos binários, avaliando a presença ou não de um atributo, como por exemplo:

Título	**Access**	**Mental Health**	**Military**	**Etc..**
Artigo 1	1	1	0	1
Artigo 2	0	1	1	0

Tabela 2 - Exemplo de conjunto de dados binários para input no algoritmo

Nesta etapa, serão desconsiderados os atributos autores e revistas onde eles foram publicados, o esperado é que seja usado no agrupamento apenas as palavras chave já definidas no resumo e as palavras chaves obtidas no processamento de linguagem natural do resumo, assim será possível montar um conjunto de dados binários baseado apenas no conteúdo do artigo.

Solução

A solução e objetivos propostos são:

- Utilizar processamento de linguagem natural para ler resumos de artigos científicos e encontrar palavras importantes para classificar seus autores;
- Clusterizar os autores e revistas onde os artigos foram publicados para encontrar grupos naturalmente formados;
- Prever a possibilidade de um autor fazer parte de um grupo de autores ou de publicar em uma revista;

Com um clusterizador será possível identificar grupos de autores e revistas, enquanto com um classificador, usando dados clusterizados, poderemos fazer previsões com artigos não utilizados como dados anteriormente.

As seguintes ferramentas serão utilizadas com a IDE PyCharm:

- Scikit Learn [3];
- Nltk [4];
- Outras bibliotecas pyhton;

O algoritmo usado no agrupamento será o Modelo de Mistura Gaussiana, pois parte dos dados serão observados de forma binária. Como o k-means calcula a média dos valores, ele não seria adequado para a utilização neste conjunto de dados, a melhor escolha seria o agrupamento com a Mistura Gaussiana. Para o algoritmo de classificação será utilizado Árvore de Decisão, escolhido também pelo uso de um conjunto de dados com atributos binários.

Benchmark

Existe no mercado algumas APIs de processamento de linguagem natural que fazem classificação de textos, uma delas por exemplo é a Monkey Learn [5].
Esta API é capaz de fazer classificações de textos e descobrir uma categoria de assunto a que ele se refere, como neste exemplo utilizando dados do Twitter [6].
Outras APIs como dialogflow, wit.ai ou luis.ai também exploram o processamento de linguagem natural como um serviço. Estas três ferramentas exploram mais especificamente para conversação, enquanto a aplicação do Monkey Learn está mais ligada à realidade deste projeto, que é extrair de um texto algumas classificações ou palavras chaves que ajudam a simplificar o objeto de pesquisa do artigo.

No website da biblioteca do nltk também existe uma descrição detalhada da utilização da ferramenta, de modo a utilizá-la da maneira mais simples e efetiva para um bom resultado do projeto.

Os exemplos acima, no entanto, não podem ser usados para comparação de resultados, pois não fornecem uma métrica explícita para comparação com o projeto proposto, portanto, para comparação métrica de resultados será utilizado o score e avaliação dos projetos anteriores do curso, comparando os resultados obtidos na etapa de agrupamento e na etapa de classificação com o resultado e com o que foi aprendido durante os projetos 2 e 3 do curso. Nestes projetos foram utilizados os mesmos algoritmos de classificação e clusterização propostos para este projeto, embora o escopo do projeto seja diferente, a comparação será embasada observando o conjunto de dados e sua qualidade em si.

Avaliação

A avaliação do modelo será feita utilizando os scores de avaliação da própria biblioteca scikit:
O agrupamento será avaliado com o score do Modelo de Mistura Gaussiano.[7]
A classificação será avaliada utilizando o f1 score.[8]

Será também utilizado um modelo de validação cruzada e boas práticas de uso de dados de teste e treinamento. O conjunto de dados será separado em 80% para treinamento e 20% para teste.
Também será usado como avaliação a experiência real de pessoas convidadas para avaliar os resultados não mensuráveis, no entanto, sem qualquer modelo heurístico.

Esboço do Projeto

Extração de Dados:

A primeira parte do projeto consiste em criar um script capaz de ler arquivos .nbib e extrair os dados os convertendo em um pandas dataset [9].

Processamento de palavras chaves:

Apenas lendo as palavras chaves dos artigos já podemos categorizar um artigo sem a necessidade do processamento de seu resumo, neste caso, iremos adotar um modelo binário que verifica se o artigo tem uma palavra chave (1), e se não tem uma palavra

chave (0). Essas palavras chaves já são definidas no resumo pelo autor, só será necessário realizar a coleta destes dados, geralmente os artigos científicos trazem de três a cinco palavras chave.
Este procedimento irá gerar uma grande quantidade de colunas para cada objeto de estudo, o que precisará ser tratado posteriormente com o método de redução de dimensionalidade (PCA).

Processamento do resumo:

Aplicando processamento de linguagem natural com nltk [4], serão estudado métodos como tokenização ou stemmização para extrair palavras importantes ou classificações dos resumos. Essas palavras ou classificações também serão tratadas como binárias ao enviar para o dataset de artigos observados.
Outra alternativa para o nltk será a utilização do TD-IDF para extrair termos frequentes ao invés de mergulhar em um complexo problema de processamento de linguagem natural, o que não é o problema real do projeto.

PCA das ocorrências binárias:

Nesta etapa do projeto, os dados dos artigos consistirão em uma tabela binária de ocorrências de classificações, e poderão necessitar de uma redução de dimensionalidade para melhor observar o conjunto de dados, bem como clusteriza-los em grupos.

Variável de saída da etapa de agrupamento

É esperado, nesta etapa do projeto, os grupos de artigos como variável de saída do clusterizador, conforme tabela 1, na seção Problema e Justificativa.

Avaliação de revistas:

Nesta etapa será realizada uma contagem percentual de quais grupos de artigos as revistas mais publicam, conforme previsto no problema, será possível avaliar qual grupo de artigos uma revista mais publica.

Previsão de um novo artigo:

Com os dados agrupados, espera-se no final poder prever em que grupo um novo artigo pertenceria, para que seja possível avaliar a possibilidade de um determinado artigo ser publicado em uma determinada revista.
Usando o grupo encontrado como variável alvo, será construído um classificador utilizando um algoritmo de árvore de decisão.
O algoritmo de clusterização não será usado para prever novos artigos apenas por fins acadêmicos, para que o projeto cubra o máximo do que foi exposto no curso.

Resultado final esperado

Pesquisar um novo artigo, do mesmo site utilizado para coletar dados, e prever qual revista tem mais afinidade com o artigo.

Referências

1. Romanzoti, N. Estranho mapa do mundo baseado na produção científica. *HypeScience* (2015). Available at: https://hypescience.com/mapa-mundo-ciencia-producao-cientifica/. (Accessed: 8th February 2018)
2. Brasil cresce em produção científica, mas índice de qualidade cai - 22/04/2013 - Ciência - Folha de S.Paulo. *Folha de S.Paulo* (2013). Available at: http://www1.folha.uol.com.br/ciencia/2013/04/1266521-brasil-cresce-em-producao-cientifica-mas-indice-de-qualidade-cai.shtml. (Accessed: 8th February 2018)
3. scikit-learn: machine learning in Python — scikit-learn 0.19.1 documentation. Available at: http://scikit-learn.org/stable/. (Accessed: 8th February 2018)
4. Natural Language Toolkit — NLTK 3.2.5 documentation. Available at: http://www.nltk.org/. (Accessed: 8th February 2018)
5. MonkeyLearn - Natural Language Processing. Available at:

https://monkeylearn.com/. (Accessed: 8th February 2018)

6. monkeylearn. monkeylearn/twitter-post. *GitHub* Available at: https://github.com/monkeylearn/twitter-post. (Accessed: 8th February 2018)
7. sklearn.mixture.GaussianMixture — scikit-learn 0.19.1 documentation. Available at: http://scikit-learn.org/stable/modules/generated/sklearn.mixture.GaussianMixture.html#sklearn.mixture.GaussianMixture.score. (Accessed: 13th February 2018)
8. sklearn.metrics.f1_score — scikit-learn 0.19.1 documentation. Available at: http://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html. (Accessed: 13th February 2018)
9. Python Data Analysis Library — pandas: Python Data Analysis Library. Available at: https://pandas.pydata.org/. (Accessed: 8th February 2018)