Análise de Resumos de Artigos Científicos

Paulo Henrique da Costa
São Paulo, 2018

Definição

Visão Geral

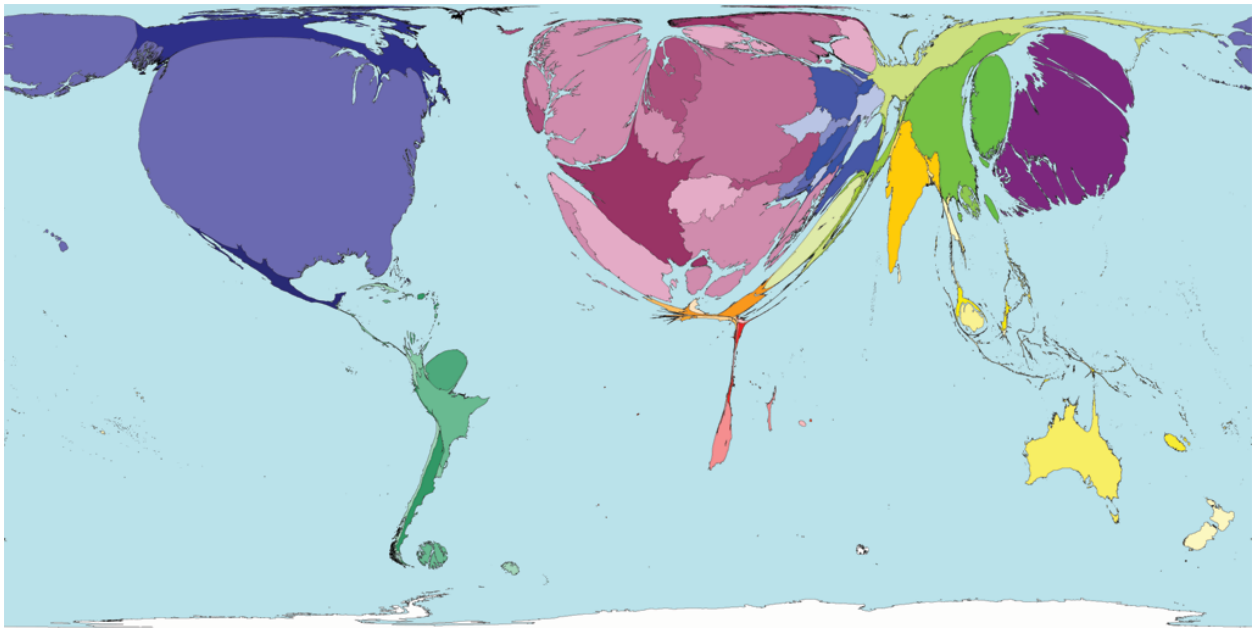

Figura 1 - Mapa mundo da produção científica [1]

A produção científica é muito vasta, mesmo em uma específica área é difícil acompanhar todas as pesquisas e descobertas que o mundo está produzindo, em 2011, por exemplo, os pesquisadores brasileiros publicaram 49.664 artigos [2] em diversas revistas nacionais e internacionais.

Considerando que os números de publicações são altos, devemos notar que existe um número ainda maior de artigos que foram escritos para disputar essas “vagas” de publicações. E o que será que têm agradado as revistas na escolha dos artigos publicados?

Neste projeto de pesquisa foi aplicado um algoritmo de machine learning na hipótese de se criar um preditor capaz de dizer em que revista um artigo têm mais chance de ser

publicado. Diversos artigos foram obtidos no site https://www.ncbi.nlm.nih.gov/pubmed/ e seu conteúdo foi extraído e analisado utilizando também processamento de linguagem natural.

Declaração do Problema

Considerando o grande número de publicações veiculados pelas revistas, decidiu-se criar um algoritmo para auxiliar os autores de artigos científicos a encontrar quais revistas tem mais afinidade sem que tivessem a necessidade de ler tudo o que as revistas publicam.

Dentro deste problema foi teorizado dois processos importantes:
O primeiro deles seria criar um classificador, cuja variável alvo fosse a revista publicada, este classificador teria a capacidade de analisar os resumos dos artigos através de processamento de linguagem e detectar em qual revista um novo artigo tem maior probabilidade de ser publicado, tendo como parâmetro os artigos que a revista vêm publicando.

O segundo seria testar um agrupamento com um algoritmo de clusterização, para ter uma visão de quantos grupos de tipos de artigo estão sendo publicados ultimamente para testar se seu artigo é uma novidade dentro do que está sendo veiculado, ou se será apenas “mais do mesmo”.
Em ambos os casos, os artigos foram analisados por um algoritmo de processamento de linguagem natural, e convertidos em conjunto de dados binários.

Métricas

A avaliação do modelo foi feita utilizando os *scores* de avaliação da biblioteca de algoritmos de machine learning do Scikit Learn [3];
O agrupamento foi avaliado com o *silhuette_score*[4];
A classificação foi avaliada utilizando o *accuracy_score*[5];

Medidas que variam de -1 e 1, onde tipicamente, valores mais próximos de 1 indicam o melhor modelo, e valores negativos indicam erro.

Análise

Exploração dos dados

Para elaboração do projeto foi escolhida uma pequena amostragem de artigos que foram coletados manualmente no site https://www.ncbi.nlm.nih.gov/pubmed/. No entanto, em vez de utilizar os artigos no formato .nbib, como idealizado no projeto, foi utilizado arquivos .xml.
Eles estão no idioma inglês, 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.
Sendo eles:

- Número total de artigos: 124
    - Basic & clinical pharmacology & toxicology: 17
    - Child psychiatry and human development: 15
    - Clinical psychology & psychotherapy: 20
    - Drug and alcohol dependence: 17
    - Journal of affective disorders: 20
    - Neuropharmacology: 17
    - Psychiatry investigation: 18

Umas das grandes questões em escolher artigos com uma certa similaridade é justamente porque um pesquisador, ao pesquisar em qual revista pode submeter seu artigo, não irá buscar revistas muito distantes de seu ramo de atividade, então não faria sentido misturar artigos de física, medicina e engenharia mecânica por exemplo, pois seria muito óbvio predizer qual a melhor revista submeter, porém, inevitavelmente, a pontuação esperada nas métricas pode não ser a melhor esperada.
Outro fator importante é que os artigos estão entre os mais recentes publicados, tornando a análise bem pontual para o atual cenário, pois podem existir tendências em relação à uma época, não faz sentido pesquisar artigos que estas revistas publicaram dez anos atrás.

Algoritmos e Técnicas

O algoritmo usado no agrupamento foi o Modelo de Mistura Gaussiana, pois os dados são observados de forma binária. Como o k-means calcula a média dos valores, ele não seria adequado para a utilização neste conjunto de dados.

Para o algoritmo de classificação foi utilizado Árvore de Decisão, Regressão Logística e *Support Vector Machine*, foi feita uma comparação entre os três para a escolha do melhor algoritmo.

No processamento de linguagem natural foi utilizado as seguintes técnicas:

- Tokenização: conversão de um texto em uma lista de palavras;
- Stemmização: Conversão de palavras em seus radicais;
- Tagging: Processo de categorização capaz de identificar se uma palavra é um pronome ou verbo, por exemplo.

Visualização Exploratória

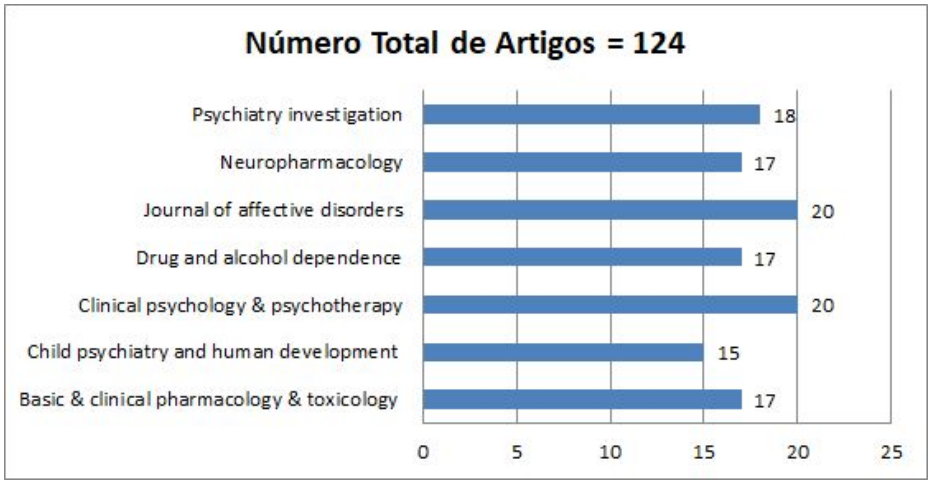

Figura 2 - Gráfico com a visão geral dos artigos utilizados no projeto

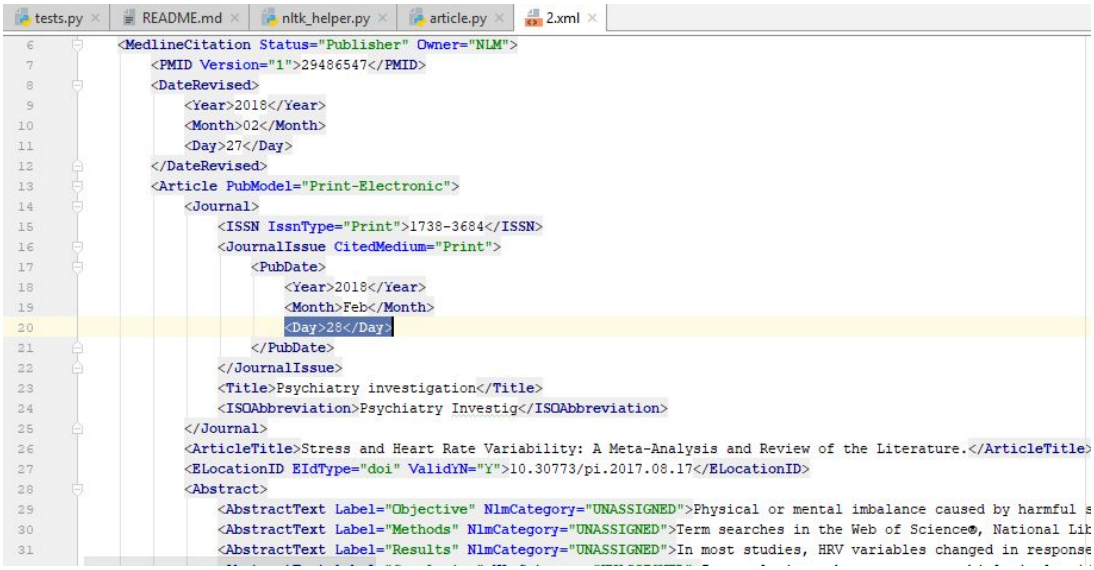

Figura 3 - Visualização de um artigo em formato .xml

Benchmark

A única ferramenta encontrada que tem o mesmo objetivo que o projeto proposto foi a disponibilizada pelo site http://www.edanzediting.com/journal_selector/.

Neste site o usuário pode inserir as palavras chave de seu artigo e tentar encontrar revistas que possuem similaridade com seu projeto. No entanto, esta ferramenta não oferece pontuações para que possa ser comparado os resultados com os obtidos pelo algoritmo deste projeto. Portanto a comparação de resultados é apenas qualitativa e observacional.

As pontuações dos algoritmos também foram comparadas com os outros projetos desenvolvidos ao longo do curso de machine learning.

Metodologia

Processamento de Dados

As seguintes ferramentas foram utilizadas com a IDE PyCharm:

- Scikit Learn [3];
- Nltk [6];

Conforme definido em projeto, e melhorado ao longo do projeto, foram adotadas as seguintes etapas de processamento:

Extração de Dados:

Após baixar todos os artigos, foi utilizada a biblioteca *xml.etree.ElementTree* para extrair os dados dos arquivos e converter em objetos definidos nas classes do pacote *custom_entities.*

Processamento de palavras chaves e resumo:

Foi adotado um modelo binário que verifica se o artigo tem uma palavra chave (1), e se não tem uma palavra chave (0).
No resumo foi aplicado processamento de linguagem natural com nltk [6], utilizando os seguintes passos para definir um conjunto de palavras consideradas importantes:

- Resumo foi quebrado em palavras e convertidas para letras minúsculas;
- Palavras foram tagueadas, para que fosse possível remover verbos de transição;
- Palavras reduzidas ao seu radical, com o objetivo de eliminar duplicidades, como por exemplo: *adult* e *adults* foram convertidas em *adult*;
- Foram removidas palavras comuns definidas na biblioteca stopwords do nltk;
- Foram removidas palavras com menos de 4 letras;
- Palavras foram convertidas em um conjunto de dados binário, assim como o adotado com as palavras chave;
- Apenas as 10 palavras mais importantes do resumo, e cinco palavras chave foram utilizadas no conjunto final;

Conversão em conjunto de dados:
Todos os artigos e suas respectivas palavras chaves filtradas foram convertidos em um grande conjunto de dados utilizando a biblioteca *pandas*[7]. Cada palavra foi convertida em uma coluna e cada artigo foi convertido em uma linha, e, quando um artigo possuía uma determinada palavra chave, foi inserido o valor de 1, quando não, foi inserido 0.

Exemplo visual do conjunto de dados final:

DOI	**abstin**	**access**	**accommod**	**Etc..**
doi 1	1	1	0	1
doi 2	0	1	1	0

Colunas de atributos (Primeiras 40):
['abstin', 'access', 'accommod', 'acetaminophen', 'across', 'action', 'activ', 'addict', 'adolesc', 'adrenerg', 'adult', 'adulthood', 'affair', 'affect', 'agent', 'aggress', 'alcohol', 'alexithymia', 'allianc', 'almost', 'although', 'aluminium', 'amazon', 'ambival', 'among', 'analges', 'analysi', 'androstan', 'anesthet', 'anti-epilept', 'antibiot', 'antidepress', 'anxieti', 'anxiou', 'apach', 'aptam', 'arson', 'assert', 'assess', 'associ']
Coluna-alvo: .JOURNAL

Implementação

O conjunto de dados foi separado em treinamento e teste na seguinte proporção:

- O conjunto de treinamento com 114 amostras.
- O conjunto de teste com10 amostras.

Esta separação foi usada tanto no classificador quanto no agrupamento dos artigos.

O classificador foi testado em três diferentes algoritmos na tentativa de escolher um melhor, sendo os resultados parciais:

Classificador 1 - LogisticRegression

Tamanho do Conj de Treinamento	**Tempo de Treinamento**	**Tempo de Estimativa (teste)**	**Pontuacao Accuracy (treinamento)**	**Pontuacao Accuracy (teste)**
114	0.0070 seg	0.0010 seg	0.9649.	0.4000.

Classificador 2 - DecisionTreeClassifier

Tamanho do Conj de Treinamento	**Tempo de Treinamento**	**Tempo de Estimativa (teste)**	**Pontuacao Accuracy (treinamento)**	**Pontuacao Accuracy (teste)**
114	0.0050 seg	0.0000 seg	0.9825.	0.3000.

Classificador 3 - SVC

Tamanho do Conj de Treinamento	**Tempo de Treinamento**	**Tempo de Estimativa (teste)**	**Pontuacao Accuracy (treinamento)**	**Pontuacao Accuracy (teste)**
114	0.0140 seg	0.0010 seg	0.2719.	0.2000.

Após a implementação dos algoritmos de classificação foi constatado que o que possui melhor desempenho foi o Regressor Logístico, o mesmo também foi escolhido pelos seguintes fatores:

- Com a árvore de decisão, apesar de se mostrar um bom algoritmo para o problema, os resultados apontados na estimativa geraram valores quase que absolutos, geralmente apontando o resultado para 100% em apenas uma revista da coluna alvo;
- Com o regressor logístico, o resultado foi melhor distribuído em questão de porcentagem de acerto, dando à pessoa que fosse analisar a possibilidade de ver e rankear as melhores opções de revistas para enviar seu artigo;
- O regressor se demonstrou mais equilibrado em relação de sobreajuste e subajuste, apesar de um resultado inicial de 40% de pontuação de acuidade.

O agrupamento foi implementado utilizando o Modelo de Mistura Gaussiana, pois o K-means não era recomendado para conjunto de dados binários, conforme previsto no projeto.

Durante a etapa de redução de dimensionalidade, foram escolhidos os 7 primeiros componentes identificados pelo algoritmo de PCA, que representavam cerca de 85% da variância total do conjunto.

O agrupamento foi testado com diversos números de clusters, sendo seu resultado parcial:

- Numero de cluster: 5 score: 0.305383342073
- Numero de cluster: 6 score: 0.354630535747
- Numero de cluster: 7 score: 0.269748677378
- Numero de cluster: 8 score: 0.423426297449
- Numero de cluster: 9 score: 0.418020165928

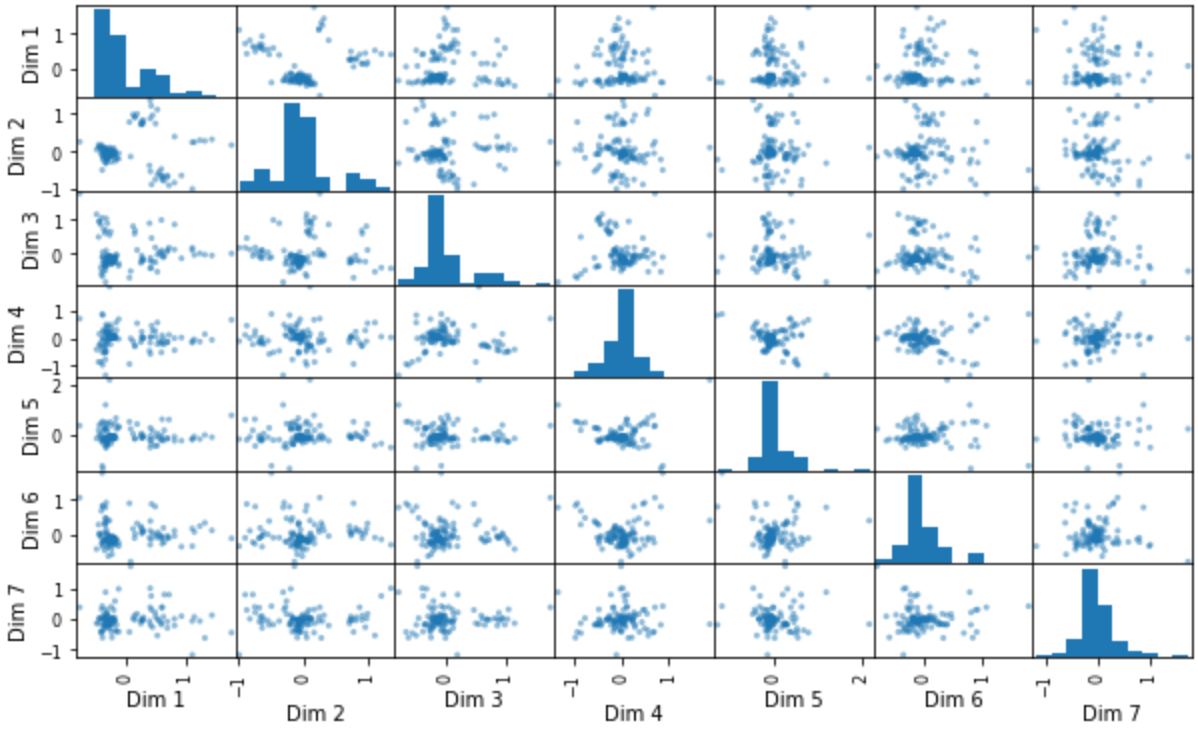

Figura 4 - Gráfico de relacionamento entre os componentes identificados no PCA

O número de clusters escolhido foi 8, primeiramente devido a sua pontuação de silhueta ser a melhor observada e por, considerando apenas a observação dos dados, o número 8 representa bem a divisão dos artigos usados no projeto.

Refinamento

O classificador foi refinado rebuscando os melhores parâmetros ***c*** e ***class_weight***:

Antes: LogisticRegression

Tamanho do Conj de Treinamento	Pontuacao Accuracy (treinamento)	Pontuacao Accuracy (teste)
114	0.9649.	0.4000.

Depois: LogisticRegression

Tamanho do Conj de Treinamento	Pontuacao Accuracy (treinamento)	Pontuacao Accuracy (teste)
114	0.9035	0.9000

O agrupamento não passou por refinamento.

Resultados

Avaliação do Modelo

Resultado final do classificador:

Tamanho do Conj de Treinamento	Pontuacao Accuracy (treinamento)	Pontuacao Accuracy (teste)
114	0.9035	0.9000

Pontuação de silhueta final do algoritmo de agrupamento:

- Melhor Score: 0.423426297449 para 8 clustrers.

O classificador final atingiu boas pontuações de acuidade, a pontuação de treinamento teve uma ligeira queda, mas isso também devido a redução do sobreajuste do modelo ao refinar os parâmetros, enquanto a pontuação de teste teve uma drástica melhora. Vale ressaltar que, como se trata de um grupo de teste bem reduzido, as pontuações podem ter drásticas variações.

A pontuação de silhueta do algoritmo de agrupamento não foi refinada, no entanto a pontuação de 0,42 para 8 clusters pode ser considerada boa devido ao fato da análise ter sido feito sobre artigos de um mesmo ramo de pesquisa.

Justificativa

Não é possível comparar o resultado do algoritmo com a ferramenta do site http://www.edanzediting.com/journal_selector/, no entanto, considerando que a

pontuação de acuidade é de 0,90, quando 1 é a melhor pontuação, temos um algoritmo refinado para realizar predições de possíveis revistas em que um artigo pode ser publicado.

Comparando com os projetos anteriormente desenvolvidos no curso de machine learning, no caso do classificador do projeto “Student Intervention” obtivemos, inclusive com o mesmo algoritmo de regressão logística:

LogisticRegression:

Tamanho do Conj de Treinamento	**Pontuacao F1 (treinamento)**	**Pontuacao F1 (teste)**
300	0.8267	0.8243

Neste projeto foi utilizado a pontuação F1, pois a coluna alvo do preditor era binária, porém esta pontuação também tem o máximo valor em 1, podemos analisar que por esta pontuação o modelo deste projeto pode ser considerado ótimo.

Já no agrupamento, os resultado do Projeto “Segmentos de Clientes” foi muito similar ao do projeto atual, tendo aproximadamente 0,42 de pontuação de silhueta também, porém com apenas 2 clusters contra 8. Ou seja, mesmo o agrupamento não sendo o foco principal deste projeto, conseguimos obter uma pontuação boa.

Conclusão

Visualização

Tabela de resultados do classificador e do agrupamento com os dados de teste:

	Basic & clinical pharmacology & toxicology	**Child psychiatry and human development**	**Clinical psychology & psychotherapy**	**Drug and alcohol dependence**	**Journal of affective disorders**	**Neuropharmacology**	**Psychiatry investigation**	**GROUP**
10.1111/bcpt.12993	0.516771	0.0488785	0.1652	0.0626269	0.0563479	0.0707318	0.0794439	5.0
10.1016/j.drugalcdep.2017.12.011	0.0759722	0.0499999	0.0621991	0.617456	0.0592309	0.0742879	0.0608543	1.0
10.1111/bcpt.12996	0.551419	0.0463863	0.14961	0.0589464	0.0532289	0.0662842	0.0741253	5.0
10.1016/j.jad.2018.02.036	0.0487957	0.0798272	0.129255	0.0419091	0.600076	0.0575488	0.0425874	4.0
10.30773/pi.2017.06.25	0.150278	0.107421	0.192618	0.122045	0.13101	0.192597	0.104031	1.0

10.30773/pi.2017.06.07	0.121372	0.0735008	0.0914014	0.102974	0.0906829	0.114892	0.405176	1.0
10.1002/cpp.2172	0.0770654	0.0346589	0.684535	0.0560383	0.0443351	0.0488891	0.0544783	5.0
10.1111/bcpt.12994	0.597335	0.0520134	0.0613375	0.0783978	0.0713569	0.0755102	0.0640497	1.0
10.1016/j.drugalcdep.2018.01.006	0.157472	0.0892861	0.114038	0.259763	0.113024	0.14789	0.118528	1.0
10.1016/j.neuropharm.2018.02.031	0.0516219	0.0467636	0.0508501	0.142017	0.0411067	0.588818	0.0788235	1.0

Nesta tabela, na última coluna se encontra o grupo do artigo, definido pelo algoritmo de agrupamento, nas demais colunas as revistas e a porcentagem de relevância que o conteúdo tem com o que é publicado nas revistas.

Artigo exemplo para reflexão:

DOI: 10.1016/j.jad.2018.02.036
PUB DATE: 28/Feb/2018
JOURNAL: Journal of affective disorders
TITLE: The prevalence and correlates of severe depression in a cohort of Mexican teachers.
ABSTRACT:
Depression is among the 10 major causes of disability in Mexico. Yet, local contextual factors associated to the disorder remain poorly understood. We measured the impact of several factors on severe depression such as demographics, pharmacotherapy, multimorbidity, and unhealthy behaviors in Mexican teachers.
A total of 43,845 Mexican female teachers from 12 Mexican states answered the Patient Health Questionnaire (PHQ9). Data were part the Mexican Teacher's Cohort prospective study, the largest ongoing cohort study in Latin America. Unadjusted and adjusted estimates assessed the impact of several contextual factors between severe versus mild-no depression cases.
In total 7026 teachers (16%) had a PHQ9 score compatible with severe depression. From them, only 17% received psychotropics, compared to 60% for those with a formal diagnosis. Less than 5% of teachers with PHQ9 scores compatible with severe depression had a formal diagnosis. Adjusted analysis reported higher odds of pharmacotherapy, having ≥ 3 comorbidities, higher levels of couple, family and work stress, fewer hours of vigorous physical activity, higher alcohol consumption, and smoking as risk factors for severe depression. Also, rural residents of northern and center states appeared more severely depressed compared to their urban counterparts. On average, the PHQ9 scores differed by ~ 10 points between severe and mild-no depressed teachers.
A cross-sectional design. Also, the study focused on female teachers between ages 25 and 74 years old, reducing the generalizability of the estimates.
Under-diagnosis of clinical depression in Mexican teachers is concerning. Unhealthy behavior is associated with severe depression. The information collected in this study represents an opportunity to build prevention mechanisms of depression in high-risk subgroups of female educators and warrants improving access to mental care in Mexico.

Observando este artigo de exemplo, que foi usado para testar os algoritmos tanto de classificação quando de agrupamento pudemos observar que:

- Foi previsto com 60% de chance de ser publicado na revista que realmente foi publicado;
- A segunda revista, com quase 13% de probabilidade também publica artigos similares a revista em que o artigo realmente foi publicado;
- Foi incluído no grupo 4, um grupo distante dos demais grupos, o que têm uma certa relevância, por ser um estudo longitudinal específico com professores mexicanos;
- Estudo de coorte também não são muito comuns, podemos relevar então que este grupo 4, pode conter conteúdo mais inovador, por isso difere um pouco da maioria.

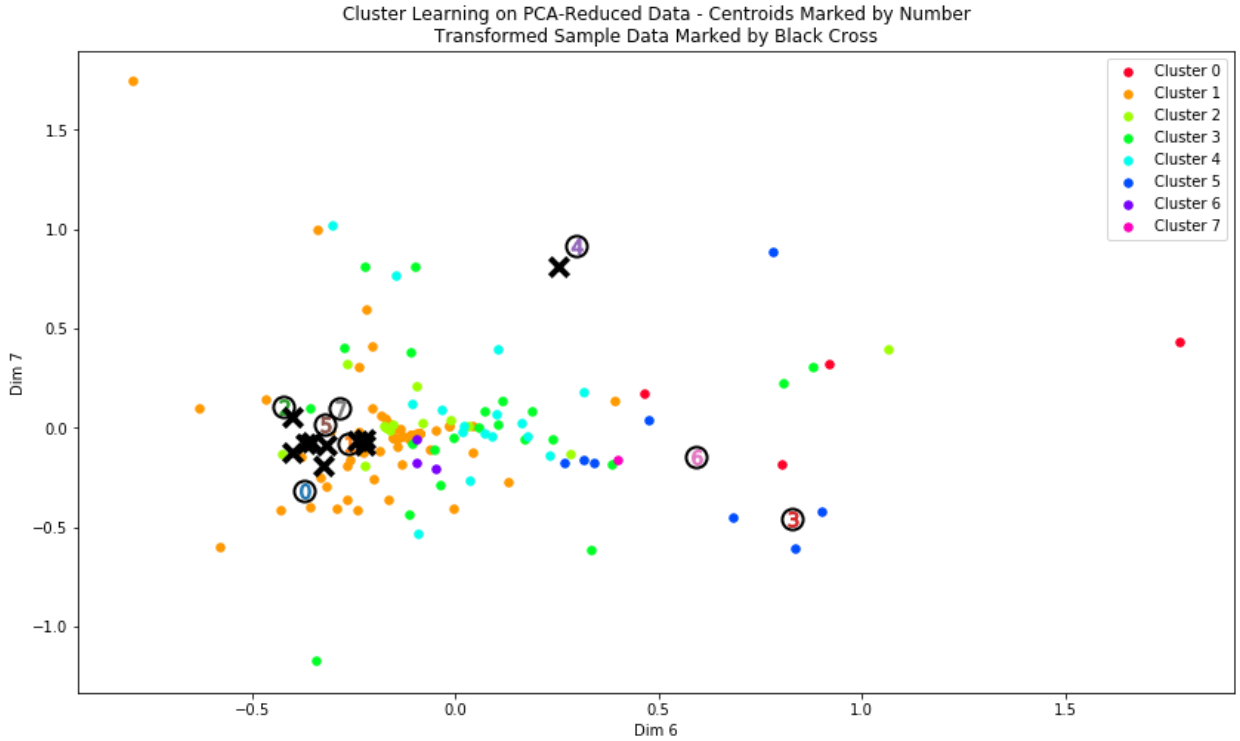

Figura 5 - Grafico de agrupamento pelas coordenadas Dim 6 e Dim 7

Reflexão

No presente projeto foi possível criar um preditor de publicações com uma pontuação ótima, considerando o número de amostras e a nível de tratamento dos dados.

O classificador nos retorna uma lista com a probabilidade de similaridade que um artigo têm com uma revista, dando a possibilidade se analisar as duas ou três revistas que estão mais próximas do artigo analisado.
O agrupamento foi interessante para avaliar se o grupo do artigo em questão é um grupo muito comum, ou pode ser caracterizado como incomum, ou inovador, o que pode ser muita vantagem na hora de publicar. Por exemplo, os grupos 0, 1, 2, 5 e 7, por exemplo, como observado na figura 5, são grupos de artigos similares e que representam a maior parte dos artigos publicados. Saber se seu artigo é um destaque dentro da maioria é uma grande vantagem na hora de submeter para uma revista.

Melhorias

Apesar dos resultados satisfatórios, existe uma série de melhorias que foram identificadas ao longo do projeto, sendo as mais importantes:

- Criar uma base de consulta mais robusta, para que os algoritmos possam ser testados com maior número de objetos de treinamento e teste;
- Melhorar o processamento de linguagem, criando uma biblioteca maior de palavras comuns, para que palavras importantes tenham ainda mais ênfase;
- Organizar ainda mais as funções e classes para tornar uma futura análise mais fácil.

Por fim, com as melhorias citadas, seria possível construir uma API para elaborar um serviço que pudesse ser compartilhado e usado pela comunidade científica.

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. sklearn.metrics.silhouette_score — scikit-learn 0.19.1 documentation. Available at: http://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html. (Accessed: 19th March 2018)
5. sklearn.metrics.accuracy_score — scikit-learn 0.19.1 documentation. Available at: http://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html. (Accessed: 19th March 2018)
6. Natural Language Toolkit — NLTK 3.2.5 documentation. Available at: http://www.nltk.org/. (Accessed: 8th February 2018)
7. Python Data Analysis Library — pandas: Python Data Analysis Library. Available at: https://pandas.pydata.org/. (Accessed: 8th February 2018)