Nanodegree Engenheiro de Machine Learning

Projeto final

Fábio Corrêa Cordeiro
Agosto de 2018

I. Definição

Resumo

O presente projeto final consiste na utilização de algoritmos de aprendizado não supervisionados conhecidos como Doc2Vec para representar documentos de um corpus através de vetores (Le et al., 2014). Esses vetores serão utilizados para treinar algoritmos de aprendizado supervisionado e gerar modelos de classificação nos subdomínios de Óleo e Gás. Desta forma será possível identificar quais documentos estão semanticamente relacionados, além de classificá-los nos subdomínios.

Esse projeto é uma evolução do trabalho "Word Embeddings em português para o domínio específico de Óleo e Gás" (Gomes, 2018), disponível em https://github.com/diogosmg/wordEmbeddingsOG. Esse projeto consistiu em testar algoritmos de vetorização de palavras (Mikolov et al., 2013) de forma que fosse possível identificar similaridades semânticas entre palavras de um mesmo domínio de conhecimento.

Descrição do problema

Neste projeto são resolvidos dois problemas. O primeiro é, dado um determinado documento, identificar quais outros documentos de um determinado corpus são similares. O segundo problema é classificar esse documento em um dos subdomínios de Óleo & Gás.

O corpus utilizado é composto por 290 projetos finais (monografias de graduação, dissertações de mestrado e teses de doutorado) do Programa de Recursos Humanos da Agência Nacional do Petróleo, Gás Natural e Biocombustível (PRH-ANP). Esses documentos estavam originalmente em formato PDF que, em alguns casos, haviam sido digitalizados. Foram utilizadas técnicas de reconhecimento de

caracteres (*OCR - optical character recognition*) para extrair os textos para o formato TXT. Todos esses documentos contém, após serem preprocessados, um total 3.308.466 palavras e um vocabulário de 52.880 palavras únicas.

Como o conjunto de documentos não é exageradamente grande, a extração dos subdomínios foi feita manualmente. Em geral a indústria de Óleo & Gás é dividida em "Upstream" (atividades que vão desde estudos geológicos, descoberta do petróleo até a sua produção) e "Downstream" (atividades que vão do refino do petróleo, produção dos derivados, distribuição e comercialização). Também foi incluído o subdomínio "Interdisciplinar" que contempla os documentos relacionados às disciplinas que afetam à toda a cadeia de Óleo e Gás (meio ambiente e desenvolvimento de novos materiais, por exemplo).

Métricas

A avaliação de similaridade entre os documentos foi feita através da comparação empírica principalmente utilizando a visualização dos resultados. Para reforçar as comparações criamos alguns documentos sintéticos a partir de documentos originais pertencentes ao corpus. A expectativa foi encontrar o documentos sintéticos e os originais próximos.

Já para a avaliação da classificação dos documentos foram calculadas as métricas de acurácia, precisão, revocação e F1-Score. Como não há nenhuma preferência entre minimizar os falsos positivos ou falsos negativos usamos o F1-Score como métrica principal para as tomadas de decisão.

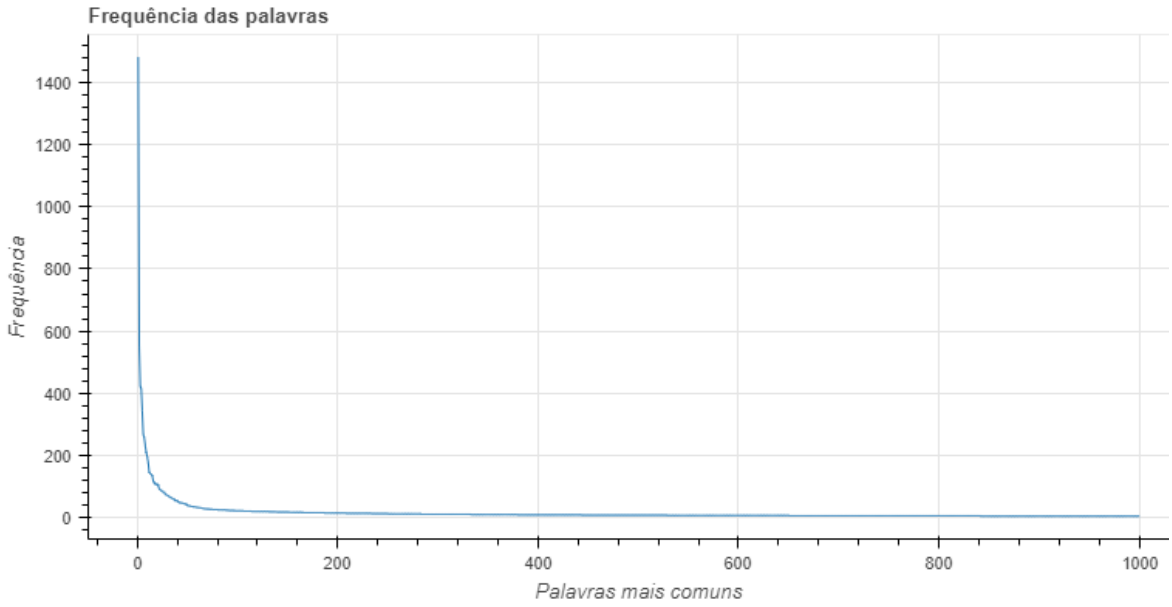
II. Análise

Exploração e Visualização dos Dados

O corpus que iremos trabalhar possui 290 documentos. Inicialmente vamos analisar apenas um documento como exemplo, o arquivo "PRH/20120904-MONOGRAFIA_0.txt". Para esse exemplo o número total de palavras (também chamado de tokens) é 20.395 e o tamanho do vocabulário (tokens únicos) é 5.271.

Os documentos em geral possuem a característica de possuirem relativamente poucas palavras que aparecem muitas vezes no texto, enquanto existem uma

quantidade muito grande de palavras que são raras. Podemos notar essa característica ao plotar um gráfico com a frequência de cada palavra no texto e verificar o efeito de "cauda longa".



Frequência das 1000 palavras mais comuns do arquivo "PRH/20120904-MONOGRAFIA_0.txt"

O próximo passo é a identificação do número de tokens e o vocabulário total de todos os documentos. Para isso é necessário unificar todos os documentos para analisá-los em conjunto.

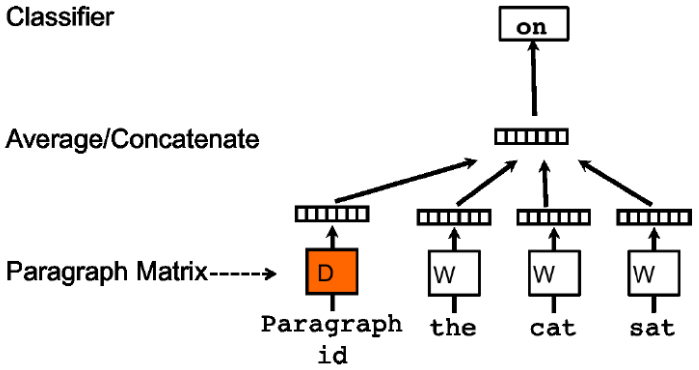
O corpus que estamos trabalhando é composto por cerca de 5,8 milhões de palavras e um vocabulário de aproximadamente 381 mil palavras diferente. No entanto, ao olharmos para as palavras mais comuns notamos que, em sua grande maioria, são compostas por termos com pouco significado semântico, ou seja, trazem pouca informação sobre o conteúdos dos textos de onde foram retiradas. É comum encontrar nas palavras mais frequentes artigos, preposições e pronomes. Também é possível notar que "a" e "A" são tratadas como duas palavras diferentes. Os números e caracteres especiais também estão presente nesse nosso vocabulário inicial.

Portanto, o próximo passo é realizar um preprocessamento do texto antes de começar a trabalhar com ele.

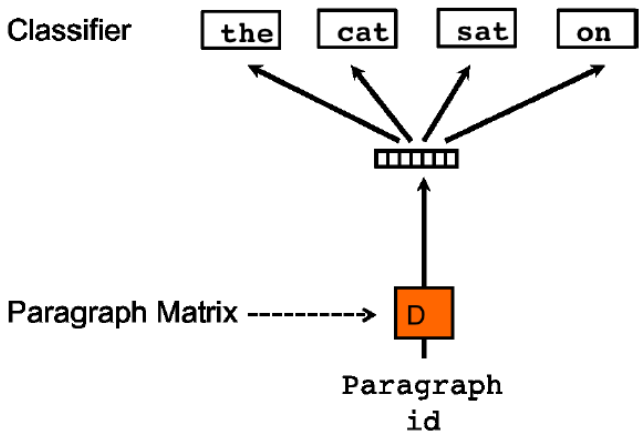
Núvem de palavras dos documentos unificados. O tamamnho das palavras é proporcional à frequência no corpus.

Algoritmos, Técnicas e Bechmark

O principal algoritmo utilizado neste trabalho é o Doc2Vec. Este algoritmo é a implantação dos modelos de "Document Embeddings" proposto por Le et al (2014). Foram propostos dois modelos de vetorização de frases (ou documentos), um é o "Phrase Vector – Distributed Memory" (PV-DM) e o outro é o "Phrase Vector – Distributed bag of words". O dois modelos são estruturas de redes neurais que utilizam vetores de palavras e frases (ou documentos) para prever as palavras seguintes.



Estrutura da rede neural do modelo PV-DM em Le et al (2014)



Estrutura da rede neural do modelo PV-DBOW em Le et al (2014)

Após Mikolov et al (2013) propor o algoritmo de vetorização de palavras, outros trabalhos foram desenvolvidos para generalizar os métodos de vetorização para frases e documentos. Nos trabalhos de Le et al (2014) e Dai et al (2015) foram utilizados conjuntos de documentos previamente anotados para comparar o algoritmo de vetorização de frases com outras técnicas similares.

Em Le et al (2014) foram usados o "Stanford Sentiment Treebank Dataset" para testar uma única frase e 100.000 críticas de filmes do site IMDb para testar textos com algumas sentenças. Já em Dai et al (2015) os algoritmos foram testados em textos mais longos, como artigos da Wikipedia e artigos acadêmicos do arXiv.

Mais recentemente, Nooralahzadeh et al (2018) realizou experimentos com vetorização de palavras para o domínio específico de Óleo e Gás, mas com documentos na língua inglesa. Tanto Nooralahzedeh quanto o nosso trabalho (Gomes, 2018) tenta demonstrar que o uso de corpus específicos de um domínio técnico retornam modelos melhores do que quando é usado corpus gerais.

Os experimentos realizados nesses artigos tentaram recuperar e classificar documentos usando os vetores. O presente projeto irá se basear nesses experimentos para realizar a recuperação e classificação dos 290 projetos finais do PRH-ANP. Como benchmark de comparação, nesse projeto vamos confrontar o modelo gerado com um classificador ingênuo. Esse classificador ingênuo irá distribuir os documentos aleatoriamente usando as probabilidades de ocorrência de cada subdomínio.

III. Metodologia

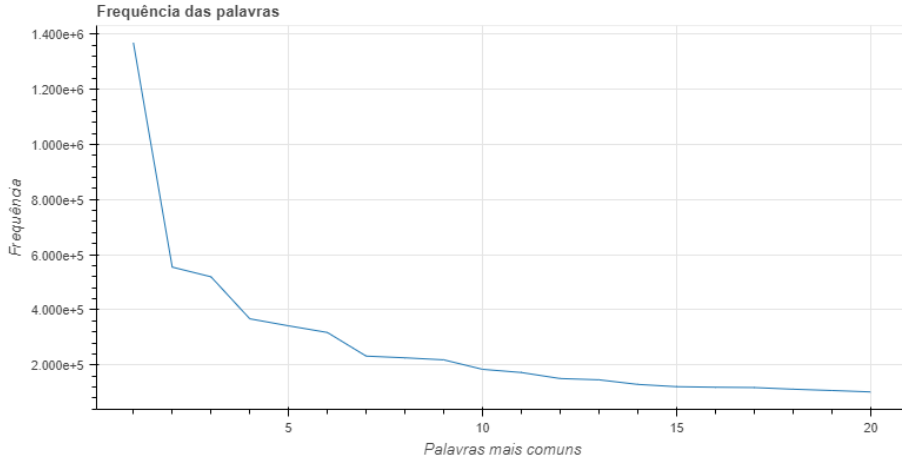
Preprocessamento dos Dados

Antes de começarmos a etapa de vetorização das palavras e dos textos é necessário preprocessar o texto original. Nesse preprocessamento foram extraidas as palavras com pouca representação semântica, além de tentar minimizar os efeitos de possíveis erros de digitalização.

A etapa de preprocessamento contou com as seguintes etapas:

- Substituir todas as letras pelo seu formato minúsculo. Dessa forma não há diferenciação das palavras grafadas em maiúsculo ou minúsculo.
- Retirar todas as acentuações, pois essa é uma grande fonte de problemas nos documentos digitalizados.
- Eliminar os caracteres de pontuação e números, já que esses podem atrapalhar na identificação da semântica pelos métodos utilizados.
- Por fim, subtrair as palavras com pouco significado semântico, as chamadas "stopwords".

As "stopwords" são as palavras que aparecem com maior frequência nos textos em geral. Por elas ocorrerem com uma probabilidade maior em qualquer documento, a ocorrência delas em um documento específico acrescenta pouca informação nova. Se observarmos as palavras mais frequentes no nosso corpus vamos identificar a presença de vários artigos e preposições.




```
[('de', 1367392), ('a', 554872), ('e', 519760), ('do', 367452), ('o', 342004), ('da', 317888), ('em', 232212), ('que', 22604
4), ('para', 218672), ('.', 184304), ('com', 172756), ('na', 150860), ('é', 146280), ('os', 129640), ('um', 121400), ('no', 11
9348), ('uma', 118228), ('dos', 112260), ('as', 107548), ('A', 102392)]
```

As 20 palavras mais comuns e sua frequência

Para extrair as "stopwords" vamos recorrer às bibliotecas da “Natural Languages Toolkit” (NLTK) com stopwords em português, inglês e espanhol. Também iremos considerar "stopwords" os caracteres únicos, ou seja, as letras do alfabeto de "a" a "z". Nesse caso, as letras isoladas, em geral, representam quebras de palavras no processo de digitalização. Também acrescentamos uma lista com os erros mais comuns identificados.

Após todas as etapas de preprocessamento o curpus final ficou com 5 milhões de tokens e um vocabulário de 115 mil palavras. Segue abaixo a nuvem de palavras resultante.

Núvem de palavras dos documentos preprocessados. O tamanho das palavras é proporcional à frequência no corpus.

Implementação

Vetorização de palavras

Nesta sessão são explorados os modelos de vetorização de palavras e documentos. Os modelos utilizados são o Word2Vec (Mikolov et al, 2013) e o Doc2Vec (Le et al, 2014). Esses modelos buscam representar cada termo do vocabulário (Word2Vec), ou cada documento de um corpus (Doc2Vec), em um vetor. A distância entre esses vetores tende a representar a distância semântica dessas palavras ou documentos.

Primeiramente analisamos o modelo de vetorização de palavras utilizando o arquivo com o corpus unificado ("corpus_unificado_preprocessados.txt"). Usamos

os hiperparâmetros utilizados por Nooralahzadeh (2018). Em seu trabalho Nooralahzadeh analisou os efeitos dos hiperparâmetro em um corpus do domínio Óleo & Gás em inglês. A única alteração que fizemos nos hiperparâmetros foi diminuir a dimensão dos vetores finais de 400 para 100, pois o nosso corpus é relativamente pequeno. Utilizamos o mesmo tamanho dos vetores escolhido para o nosso trabalho anterior (Gomes, 2018).

Após treinar o modelo, cada palavra do vocabulário é representada por um vetor com 100 dimensões. Vejamos como ficou representado a palavra "petroleo".

```
array([ 2.3618758 , -1.8521181 , -1.3786144 , -2.0969584 , -1.2961518 ,
       -0.19831924, -1.9641155 ,  1.3643969 , -0.94847447,  0.42566538,
        1.1727434 ,  0.61650807,  3.4209597 ,  0.64802796,  0.13120443,
        2.8254073 ,  1.044286  ,  2.491462  ,  0.14032774,  1.2577726 ,
       -2.8273807 ,  0.1637695 ,  0.27993608,  1.6882681 , -0.37941235,
       -1.2170286 , -0.5623705 ,  0.80301076,  1.6129153 ,  0.91989946,
        2.9003024 ,  0.87515396, -0.98913276,  2.0391393 ,  0.02073847,
        1.0636206 , -3.1396883 ,  2.0963116 , -0.30587026,  1.8977717 ,
       -0.6259287 ,  1.6571442 ,  0.12174448, -3.1434042 ,  0.4846722 ,
       -0.3831097 , -0.9079251 ,  2.5127857 ,  2.2798963 ,  2.9179122 ,
        3.7650893 ,  0.30976075, -1.811834  , -0.15404795,  0.4961172 ,
        3.2810218 ,  0.04964071, -0.9961865 , -3.6445386 ,  1.3767942 ,
       -1.9187531 , -2.3481607 , -0.8661051 , -2.8126738 , -2.6076248 ,
       -0.24992612, -3.333937  , -0.11924378, -2.3401976 ,  1.4035327 ,
        0.26816496,  2.8288977 ,  2.2638257 , -0.15736865, -1.4177445 ,
       -1.8775516 , -0.21513961,  0.1670739 , -1.5636092 , -2.9990675 ,
        1.3286327 ,  0.7140871 ,  2.3317897 ,  0.7493984 , -0.38732418,
       -1.8412697 ,  1.0193577 ,  0.5342287 ,  1.8354077 , -1.4226781 ,
       -3.05088   ,  1.9273531 , -1.0576828 , -0.8862126 , -2.4238605 ,
       -1.402633  ,  0.8842688 ,  0.31546402, -3.045116  , -2.7274137 ],
      dtype=float32)
```

Vetor que representa a palavra “petroleo”

Outra característica dos modelos vetoriais é a possibilidade, dado uma palavra, de identificar os vetores mais próximos. Vejamos as 10 palavras mais próximas do termo "petroleo".

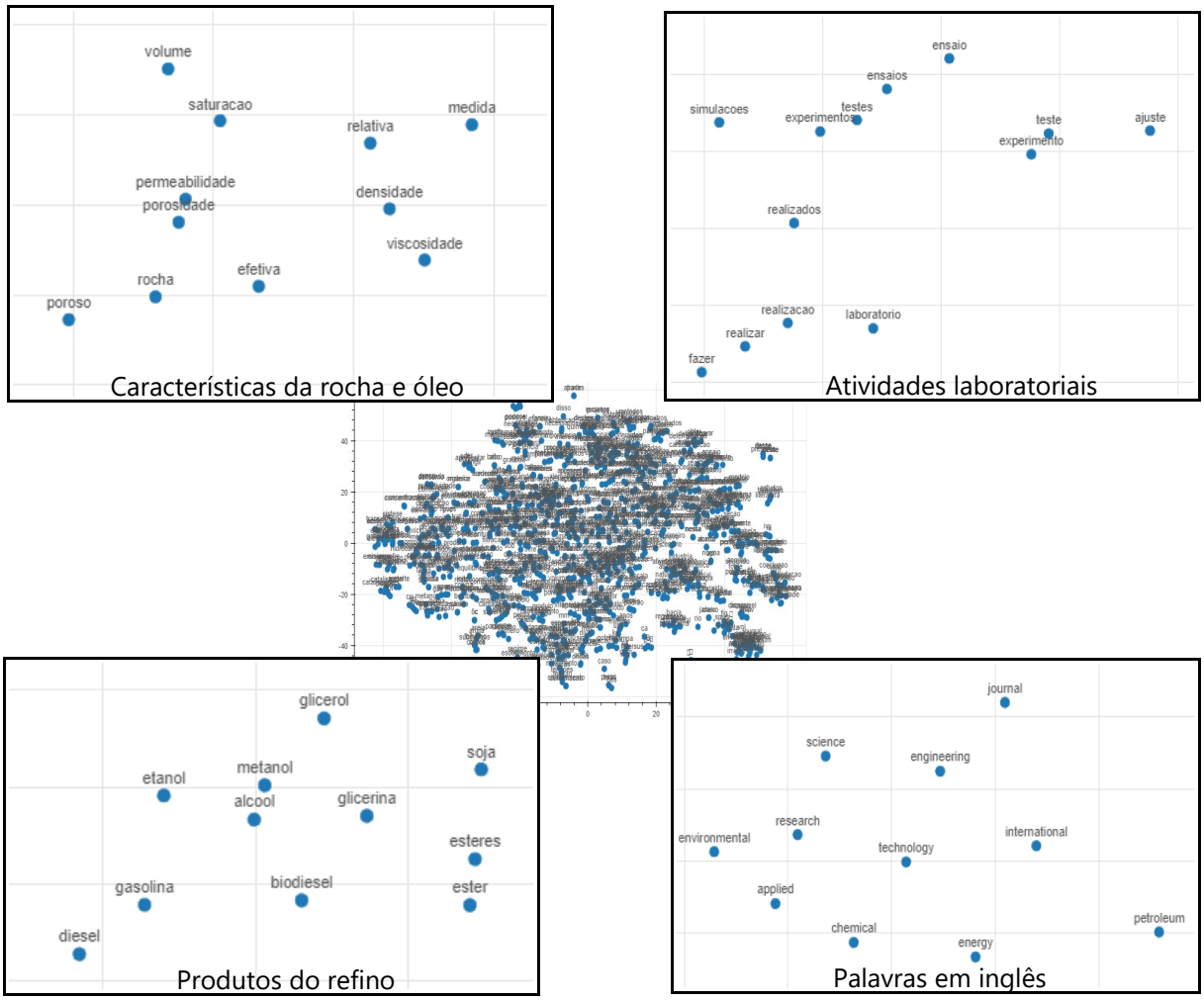
```
[('petrolifera', 0.6948655247688293),
 ('oleo', 0.6198599338531494),
 ('secundaria', 0.6076905727386475),
 ('exploracao', 0.586421549320221),
 ('primaria', 0.5741363763809204),
 ('oleos', 0.569688081741333),
 ('petroquimica', 0.5549637079238892),
 ('reservatorios', 0.5548878312110901),
 ('presal', 0.5527384281158447),
 ('automobilistica', 0.548987865447998)]
```

10 palavras mais próximas do termo “petroleo” e a sua respectiva distância cosseno.

Para ter uma ideia melhor de como os termos estão relacionados podemos visualizá-los em um único gráfico. Como o vocabulário é muito grande para plotar todas as palavras, escolhemos as 1000 palavras mais frequentes.

Para plotar as palavras em um gráfico de duas dimensões tendo vetores de 100 dimensões precisamos aplicar algum algoritmo para reduzir a dimensionalidade. Utilizamos o algoritmo t-SNE para reduzir a dimensionalidade dos vetores.

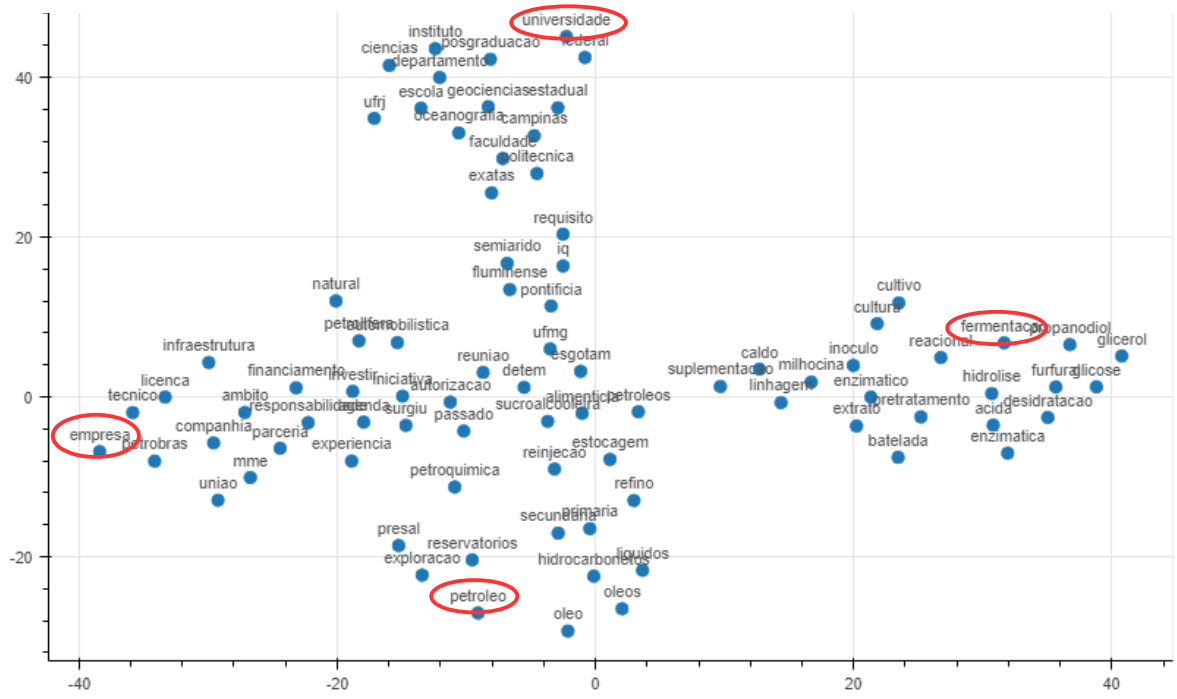
O gráfico é formado por uma núvem com as 1000 palavras mais comuns do vocabulário. Essas palavras foram plotadas de acordo com a proximidade dos seus vetores. Ao navegarmos por essa núvem utilizando o zoom podemos notar algumas relações semânticas entre os termos que compõem um mesmo aglomerado. Essa proximidade semântica é diretamente relacionada com corpus onde o documento foi treinado (Gomes, 2018).



Núvem com 1000 palavras mais comuns e alguns aglomerados destacados.

Para representar melhor às distâncias, tanto vetorial quanto semântica, escolhemos os termos "empresa", "universidade", "fermentacao", "petroleo". Para cada um dos

quatro termos buscamos os 20 vetores mais próximos e plotamos todos em um gráfico de duas dimensões.



Núvem com os 20 termos mais próximo das palavras “universidade”, “fermentacao”, “petroleo” e “empresa”.

Vetorização de documentos

A próxima etapa consiste em usar o algoritmo Doc2Vec para representar cada documento como um vetor (Dai et al, 2015). Esse vetor foi usado para calcular as distâncias entre os documentos e, desta forma, identificar os documentos que estão semanticamente relacionados. Os vetores também foram usados como atributos para treinar um algoritmo de classificação.

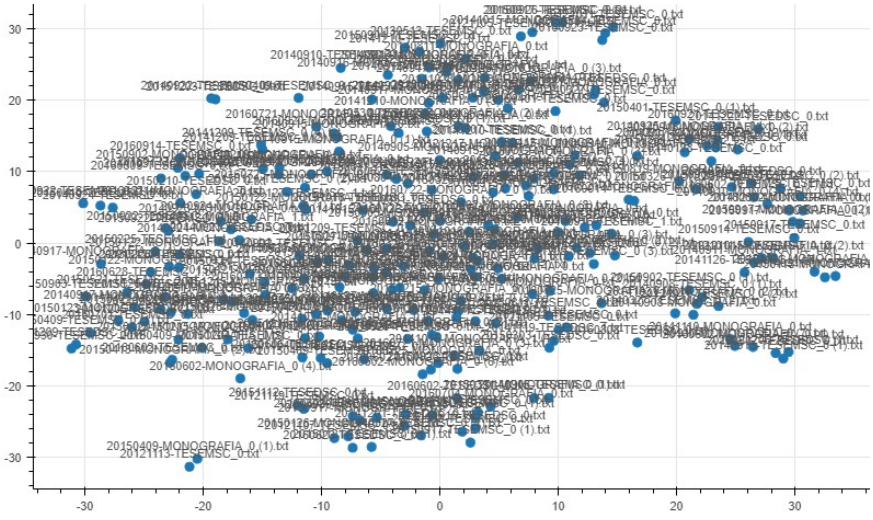
Geramos vetores do mesmo tamanho que usamos para a vetorização das palavras, 100 dimensões. A "janela" de palavras utilizadas é um hiperparâmetro importante para os algoritmos Doc2Vec. No nosso caso, como estamos trabalhando com documentos longos, escolhemos uma janela relativamente longa de 50 palavras. Por fim, usamos todas as palavras do vocabulário com frequência maior do que 10, e o número de épocas de treinamento foi 10.

Foi necessário fazer a anotação dos 290 documentos para que fosse possível gerar o modelo de classificação. Para cada documento identificamos o título e o subdomínio que ele pertencia. Com isso foi criado uma planilha como a apresentada abaixo.

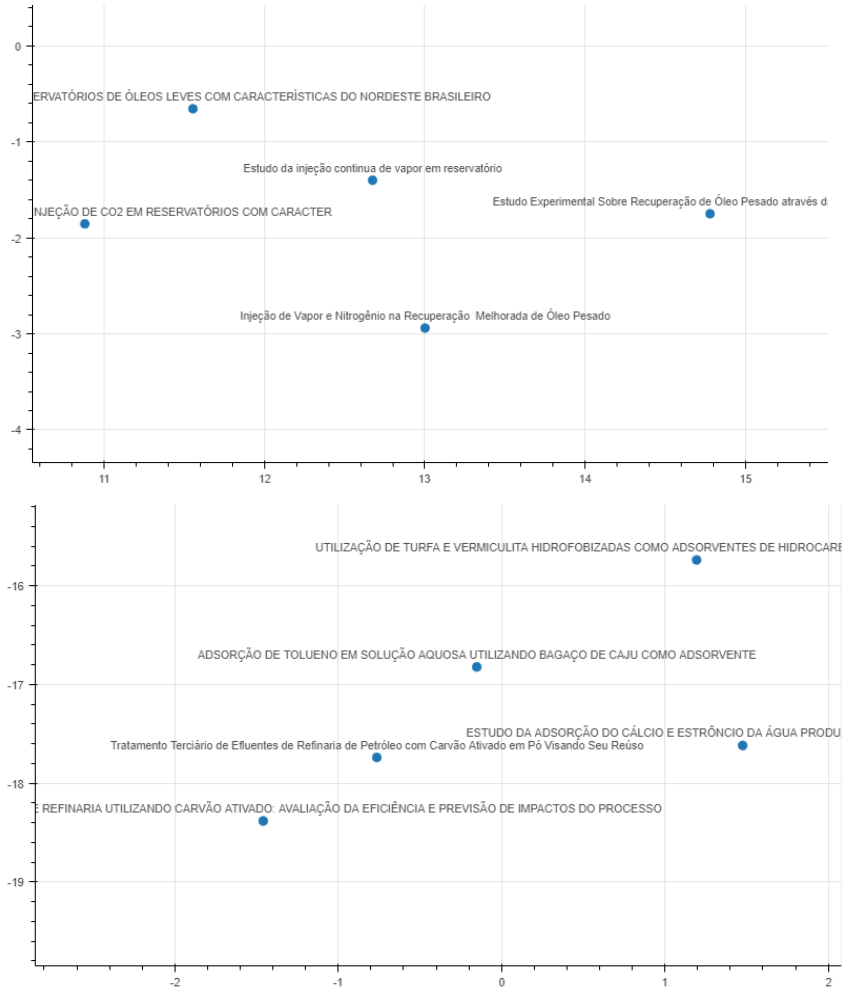
	nome_arquivo	titulo	label
0	20120904-MONOGRAFIA_0.txt	Estudo Tecnológico e Modelagem Reacional para ...	downstream
1	20121011-MONOGRAFIA_0 (2).txt	ESTUDOS AMBIENTAIS NAS ÁREAS DE ATUAÇÃO DA IND...	interdisciplinar
2	20121011-MONOGRAFIA_0.txt	A CONTRIBUIÇÃO DOS BIOMARCADORES NA GEOQUÍMIC...	interdisciplinar
3	20121011-TESEMSC_0 (2).txt	AVALIAÇÃO DO IMPACTO DA ATIVIDADE SÍSMICA SOBR...	upstream
4	20121011-TESEMSC_0.txt	Avaliação da Influência da Variabilidade Tempor...	interdisciplinar
5	20121015-TESEMSC_0.txt	DIAGNÓSTICO AMBIENTAL DO CONE DO RIO GRANDE – ...	interdisciplinar
6	20121105-MONOGRAFIA_0.txt	MODELO DE PREVISÃO DE DEMANDA POR COMBUSTÍVEIS...	downstream
7	20121105-TESEMSC_0 (1).txt	INFLUÊNCIA DE PAR ÂMETROS GEOMÉTRICOS NA RESIS...	upstream
8	20121105-TESEMSC_0.txt	ADAPTAÇÃO REGULATÓRIA NA INDÚSTRIA DE BIOCOMBU...	downstream
9	20121107-TESEMSC_0.txt	EXTRAÇÃO, PURIFICAÇÃO E IMOBILIZAÇÃO DE LIPASE...	downstream

Planilha com o nome do arquivo, título do documento e subdomínio de O&G

Da mesma forma como foi feito com os vetores de palavras é possivel plotar a nuvem de vetores que representam os documentos. Também foi realizada a redução de dimensionalidade utilizando o algoritmos t-SNE. Se navegarmos por essa núvem observando os títulos dos documentos podemos observar que documentos sobre temas parecidos estão próximos.



Vetores que representam os documentos



Destaque para aglomerados de documentos relacionados à “reservatórios” e “adsorção de produtos químicos em refinarias”

Usando os vetores para a identificação de similaridade entre documentos

Após a criação dos vetores é possível identificar quais documentos são similares entre si. Para testar essa identificação de similaridade, pegamos dois documentos e dividimo-los pela metade. Dessa forma criamos dois documentos sintéticos que, apesar de serem diferentes, são muito similares ao original. Ao procurar os documentos do corpus cujos vetores estão estão próximos aos documentos sintéticos, esperamos encontrar o documento original.

O documento sintético deve ser preprocessado da mesma forma que os originais. Em seguida, o modelo Doc2Vec previamente treinado é utilizado para inferir um

vetor para cada documento sintético. A partir desses vetores podemos identificar os 10 documentos mais próximos. Podemos notar que, como esperado, os documentos originais são os mais próximos dos sintéticos.

O título do primeiro documento sintético é "Estudo Tecnológico e Modelagem Reacional para Processo Fischer-Tropsch com Gás Natural".

	Similaridade	Título
0	0.792535	Estudo Tecnológico e Modelagem Reacional para Processo Fischer-Tropsch com Gás Natural
1	0.637237	Síntese de Dimetil Éter a Partir de Halometano
2	0.546248	Viabilidade técnica, econômica e ambiental de processo de reuso Offshore de CO2 de Gás Natural por Reforma a Seco
3	0.540003	O Mercado de Gás Natural Veicular no município de Natal
4	0.535890	Avaliação da Indústria Petroquímica no Brasil: Desenvolvimento de Modelo via Programação Matemática
5	0.515322	SÍNTESE DE FISCHER-TROPSCH SOBRE CATALISADORES CONVENCIONAIS E ESTRUTURADO PARA OBTENÇÃO DE COMBUSTÍVEIS LÍQUIDOS
6	0.503120	ANÁLISE DOS NOVOS CONDICIONANTES DA OFERTA NACIONAL DE GÁS NATURAL E A DEMANDA TERMELÉTRICA NO PRÓXIMO DECÊNIO
7	0.481491	USO EFICIENTE DE BIOGÁS DE ESGOTO EM MOTORES GERADORES
8	0.477969	Remoção de metais da água de produção utilizando tensoativo
9	0.469127	Identificação e avaliação de alternativas para a produção de plásticos convencionais a partir de matérias pr...

Vetores mais próximos do documento sintético “Estudo Tecnológico e Modelagem Reacional para Processo Fischer-Tropsch com Gás Natural”.

O título do segundo documento sintético é “Integração de técnicas computacionais como contribuição para o mapeamento dos índices de sensibilidade fluvial a derrames de óleo na região de Coari (AM)”.

	Similaridade	Título
0	0.891850	INTEGRAÇÃO DE TÉCNICAS COMPUTACIONAIS COMO CONTRIBUIÇÃO PARA O MAPEAMENTO DOS ÍNDICES DE SENSIBILIDADE FLUVIAL A DER...
1	0.723982	METODOLOGIA DE AVALIAÇÃO DA SUSCETIBILIDADE A INUNDAÇÕES EM ZONAS COSTEIRAS TROPICAIS POR INTEGRAÇÃO DE DADOS FISIOG...
2	0.722304	MAPEAMENTO MULTI-TEMPORAL DA SENSIBILIDADE AMBIENTAL A DERRAMES DE ÓLEO EM ÁREAS INUNDADAS NA REGIÃO DOS RIOS URUCU ...
3	0.558682	ANÁLISE MULTICRITERIAL PONDERADA NO ESTUDO DE SUSCETIBILIDADE EROSIVA NO ENTORNO DE DUTOS; ESTUDO DE CASO: DUT...
4	0.479779	AVALIAÇÃO DO MÉTODO DE CLASSIFICAÇÃO BASEADA NO OBJETO EM IMAGENS DE ALTA RESOLUÇÃO ESPACIAL APLICADO PARA O MONITOR...
5	0.447548	Visão computacional em meio subaquático:Um estudo sobre detecção de pontos de interessee classificação utilizando co...
6	0.412396	RESPOSTA DA BAÍA DE GUANABARA A EVENTOS EXTREMOS
7	0.411806	Interpretação Aeromagnética Sobre Áreas Proximais das Bacias de Campos e Santos Utilizando Inversão Compacta
8	0.410445	VARIABILIDADE ESPAÇO-TEMPORAL DA CONCENTRAÇÃO DE CLOROFILA-A NA PLATAFORMA CONTINENTAL DAS BACIAS DE CAMPOS E DO ESP...
9	0.405924	INTEGRAÇÃO ENTRE DADOS MORFOLOGICOS DO FUNDO E TERMO-HALINOS DA COLUNA D'ÁGUA DA PLATAFORMA EXTERNA E TALUDE DAS BAC...

Vetores mais próximos do documento sintético “Integração de técnicas computacionais como contribuição para o mapeamento dos índices de sensibilidade fluvial a derrames de óleo na região de Coari (AM)

Classificação de Documentos

A última tarefa executada foi o treinamento de um algoritmo de classificação que use como atributos os vetores gerados. Os vetores que criamos contém 100 atributos e utilizar todos esses atributos poderia gerar um sobretreinamento do classificador (“*overfitting*”). Portanto realizamos uma seleção para manter apenas os atributos que mais influenciavam na classificação.

O classificador “*decision tree*” possui uma função que permite identificar quais atributos contribuem mais para a identificação das classes, dessa forma iniciamos o nosso treinamento usando esse algoritmo. Usamos 90% dos dados para treino e 10% para teste com dez subconjuntos usando o algoritmo "*StratifiedKFold*".

Foram calculadas as métricas F1-Score, acurácia, precisão e revocação para cada subconjunto de treino e teste, e em seguida calculadas as médias dos dez subconjuntos. As decisões foram tomadas com base nos melhores valores médios de F1-Score. Para avaliar a qualidade do nosso resultado também foi gerado um classificador ingênuo que escolhe aleatoriamente uma das classes.

F1-Score	0.318442
Acurácia	0.317241
Precisão	0.328414
Revocação	0.317241

Métricas do classificador ingênuo

Após treinar o algoritmo "decision tree" recuperamos os atributos mais importantes para a classificação dos subdomínios. Os três principais atributos são responsáveis por 28% do resultado. Considerando os 10 e 20 principais atributos temos a explicação de, respectivamente, 58% e 80% dos resultados. Finalmente, se usarmos apenas os 40 principais atributos conseguimos 100% da explicação do resultado. Ou seja, 60% dos atributos não influenciam na classificação dos subdomínios de Óleo & Gás. Portanto, apesar de termos a disposição um vetor de

100 dimensões para treinamento, podemos usar apenas as melhores dimensões evitando o *"overfitting"*.

F1-Score	0.554298
Acurácia	0.554311
Precisão	0.583289
Revocação	0.554311

Métricas do classificador "*Decision tree*" utilizando apenas as 10 principais dimensões.

Refinamento

Até o momento, os parâmetros utilizados para gerar os modelos foram baseados em trabalhos anteriores ou na literatura. Com isso chegamos a um valor médio de F1-Score de 0.55. Os parâmetros utilizados foram:

- Algoritmo de treinamento do Word2Vec: PV-DBOW
- Dimensão do vetor: 100
- Janela de predição do algoritmo Word2Vec: 50
- Épocas de iteração do algoritmo Word2Vec: 10
- Algoritmo de classificação: Decision Tree
- Redução de dimensionalidade: 90%

Nesta etapa de refinamento geramos novamente os vetores dos documentos e treinamos o algoritmo de classificação alterando as principais escolhas que fizemos até então. O objetivo foi identificar a combinação de hiperparâmetros e algoritimos de classificação que retornasse a maior valor para F1-Score. Os parâmetros que foram testados são:

- Algoritmo de treinamento do Word2Vec: PV-DM ou PV-DBOW
- Dimensão do vetor: 25, 50, 100, 200 ou 400
- Janela de predição do algoritmo Word2Vec: 5, 10, 50, 100
- Épocas de iteração do algoritmo Word2Vec: 5, 10 ou 20

- Algoritmo de classificação: Decision Tree, Suport Vector Machine, Near Centroid, Gaussian Nayve Bayes, Random Forest e Neural Network.
- Redução de dimensionalidade: 0%, 50%, 80%, 90% ou 95%

Não foi possível testar todas as combinações de parâmetro, seriam 3600 combinações diferentes. Portanto, para cada parâmetro, testamos todas as opções mantendo os demais constantes (*"ceteris paribus"*). Após encontrar os melhores valores para cada parâmentro fizemos uma nova rodada testando cada parâmetro. Após essa nova série de rodadas escolheremos a combinação com maior F1-Score.

IV. Resultados

Após o processo de refinamento o conjunto de parâmetros selecionados para o modelo de classificação foi:

- Algoritmo de treinamento do Word2Vec: PV-DBOW
- Dimensão do vetor: 200
- Janela de predição do algoritmo Word2Vec: 5
- Épocas de iteração do algoritmo Word2Vec: 5
- Algoritmo de classificação: Near Centroid
- Redução de dimensionalidade: 80%

O valor F1-Score obtido foi de 70%, acima do classificador ingênuo (35%) que usamos para comparação.

F1-Score	0.700224
Acurácia	0.699720
Precisão	0.715054
Revocação	0.699720

Métricas final do classificador

V. Conclusões

Reflexões

As técnicas de vetorização, ou "*embeddings*", são ferramentas muito úteis no processamento de linguagem natual (NLP). Nesse trabalho utilizamos a vetorização de documentos ("*document embeddings*") para identificar similaridades entre dois arquivos, além de treinar um algoritmo de classificação.

A identificação de similaridade entre arquivos pode ser usada por bibliotecas, bases de arquivos ou livrarias virtuais para sugerir novos documentos para seus usuários. Essa função também pode ser usada para melhorar os resultados de sistemas de buscas ("*information retrival*"). Já o algoritmo de classificação de subdomínios de Óleo & Gás pode ser usado para gerar *"tags"* automáticas, aprimorando os metadados dos arquivos, o que também ajudaria a melhorar os resultados de sistemas de buscas.

Para exemplificar vamos usar como exemplo a introdução e o prólogo do livro "A Busca" do conhecido escritor de geopolítica do petróleo e energia Daniel Yeergin. Ao procurar quais documentos são similares, encontramos a lista abaixo. Quando usamos o nosso modelo para classificar esse texto ele é classificado como um documento "Interdisciplinar".

	Similaridade	Título
0	0.463018	A GOVERNANÇA AMBIENTAL DA PREVENÇÃO E CONTROLE DE INCIDENTES COM ÓLEO NAS ATIVIDADES MARÍTIMAS DE PETRÓLEO NO BRASIL
1	0.447471	ADAPTAÇÃO REGULATÓRIA NA INDÚSTRIA DE BIOCOMBUSTÍVEIS
2	0.425155	A GEOPOLÍTICA PETROLEIRA DO MERCOSUL COM A ENTRADA DA VENEZUELA: CONTINUIDADE OU MUDANÇA?
3	0.417927	O Mercado de Gás Natural Veicular no município de Natal
4	0.414272	O IMPACTO DA EXPORTAÇÃO DE ENERGIA ELÉTRICA DAS USINAS HIDRELÉTRICAS BINACIONAIS NO CRESCIMENTO ECONÔMICO DO P
5	0.412901	A EMERGÊNCIA DE MODELOS DE NEGÓCIOS INOVADORES PARA APOIAR O DESENVOLVIMENTO DA ELETRIFICAÇÃO VEIC
6	0.410752	TEORIA DAS OPÇÕES REAIS: A APLICAÇÃO DESSA FERRAMENTA NA ANÁLISE DE INVESTIMENTOS DE PESQUISA E DESENVOLVIMENTO NO S...
7	0.405294	O PAPEL DO PROGRAMA NACIONAL DE PRODUÇÃO E USO DE BIODIESEL COMO INSTRUMENTO DE POLÍTICA DE REDUÇÃO DAS DESIGUALDADE...
8	0.398831	UM ESTUDO COMPARATIVO DOS SISTEMAS DE INOVAÇÃO DO BRASIL E DA CHINA NA ÁREA DE COMBUSTÍVEIS LÍQUIDOS ALTERNATIVOS
9	0.376779	ANÁLISE ENTRE PROCESSOS E MATÉRIAS-PRIMAS PARA A PRODUÇÃO DE BIODIESEL

Documentos similares a introdução e prólogo do livro "A Busca" de Daniel Yeergin

Ao procurar o livro "A Busca" em sites de vendas online podemos notar como é comum a comparação entre livros do catálogo. No entanto, os sites costumam usar

a probabilidade de compra como fator mais importante, a abordagem aqui apresentada prioriza a similaridade semântica dos textos.

Comparação entre o livro “A Busca” e restante do catálogo da Amazon

Melhorias

Os resultados alcançados foram satisfatório para a quantidade de dados disponíveis. A quantidade de documentos, 290 arquivos, é grande suficiente para gerar os vetores, mas relativamente pequena para treinar um algoritmo de classificação mais preciso. Deixamos como sugestão de continuidade e melhoria a replicação desse trabalho com uma base de arquivos maior.

Outra técnica que pode ser útil é a criação de arquivos sintéticos para ampliar a base de documentos. Com um corpus maior pode-se ter um resultado melhor para o algoritmo de classificação além de permitir gerar métricas quantitativas para a tarefa de identificação de similaridade entre documentos.

Bibliografia

Mikolov, Tomas; Chen, Kai; Corrado, Greg; Dean, Jeffrey. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013. Em: https://arxiv.org/pdf/1301.3781.pdf

Le, Quoc; Mikolov, Tomas. Distributed Representations of Sentences and Documents. arXiv preprint arXiv:1405.4053v2, 2014. Em: https://arxiv.org/pdf/1405.4053.pdf

Dai, Andrew; Olah, Christopher; Le, Quoc. Document Embedding with Paragraph Vectors. arXiv preprint arXiv:1507.07998v1, 2015. Em: https://arxiv.org/pdf/1507.07998.pdf

Nooralahzadeh, Farhad; Øvrelid, Lilja; Lønning, Jan Tore. Evaluation of Domain-specific Word Embeddings using Knowledge Resources, 2018. Em: http://www.lrec-conf.org/proceedings/lrec2018/pdf/268.pdf

Gomes, Diogo; Cordeiro, Fábio; Evsukoff, Alexandre. Word Embeddings em Portugês para o Domínio Específico de Óleo e Gás. Rio Oil & Gas, 2018.

Yeergin, Daniel; A busca: energia, segurança e reconstrução do mundo moderno / Daniel Yergin; tradução Ana Beatriz Rodrigues. - 1. ed. - Rio de Janeiro: Intrínseca, 2014.