WORD EMBEDDINGS EM PORTUGUÊS PARA O DOMÍNIO ESPECÍFICO DE ÓLEO E GÁS

Diogo S. M. Gomes[1], Fábio C. Cordeiro[2], Alexandre G. Evsukoff[3]

Resumo

Word embeddings compõem uma das unidades fundamentais dos algoritmos de processamento de linguagem natural e são utilizados para modelar matematicamente a representação de palavras considerando suas relações de similaridade semântica e sintática no contexto em que ocorrem. Este trabalho descreve o processo de geração e disponibilização do primeiro conjunto público de modelos de *word embeddings* em português para o domínio específico de óleo e gás. Para sua geração, um corpus textual foi composto a partir de diversas fontes de dados publicadas por instituições de referência nesta área de conhecimento. Os modelos são qualitativamente analisados no aspecto de sua capacidade de representação de termos técnicos na área de O&G. São descritos os passos utilizados no pré-processamento, no treinamento dos modelos e os resultados obtidos na análise qualitativa. Por fim, os scripts, o corpus e os algoritmos utilizados no estudo, assim como os modelos gerados, são disponibilizados para uso público.
Palavras-chave: Word Embeddings, O&G, IA, PLN

Abstract

Word embeddings are some of the fundamental units of natural language processing algorithms, used to represent words mathematically by considering semantic and syntactic similarities in the context in which they occur. This paper describes the process of generating the first set of word embeddings models in portuguese for the specific domain of oil and gas. A textual corpus was composed from several data sources published by reference institutions in this field. The generated models are qualitatively evaluated in their ability to represent technical terms in the O&G domain. We describe each step, since pre-processing, training and the results obtained in the qualitative analysis. Finally, the scripts, corpus and algorithms used in the study, as well as the generated models, are made available for public use.
Keywords: Word Embeddings, O&G, AI, NLP

[1] Mestre, Analista de Sistemas – PETROBRAS
[2] Engenheiro de Produção – PETROBRAS
[3] Doutor, Professor Associado – COPPE/UFRJ

1. Introdução

Em tempos da chamada transformação digital (WORLD ECONOMIC FORUM, 2017), essencialmente apoiada por uma ampla disponibilidade de técnicas de ciência de dados e inteligência artificial, a indústria de óleo e gás (O&G) tem sido continuamente desafiada a extrair mais conhecimento a partir de suas bases de dados atualmente disponíveis, reforçando seus investimentos na adoção de tecnologias digitais no intuito de obter um melhor aproveitamento dessas informações. Estima-se que uma fração de 80% desses dados sejam armazenados em formatos não estruturados (BLINSTON e BLONDELLE, 2017), pressupondo a existência de valiosas informações possivelmente dispersas em diversas bases não utilizadas em seu completo potencial, ocultas em documentos tais como artigos e estudos técnicos, análises laboratoriais, *logs* de operação, publicações, periódicos e relatórios diversos.

Para viabilizar o acesso a essas informações, diversos algoritmos de processamento de linguagem natural (PLN) e aprendizado profundo (*deep learning*) (LECUN et al., 2015; GOODFELLOW et al., 2016) têm sido utilizados para lidar com o desafio de extrair conhecimento a partir desse vasto conteúdo em formato textual, representando um papel cada vez mais significativo no processo decisório das corporações. Como algoritmos desse tipo consistem no uso de palavras como dados de entrada, é necessário utilizar-se de abstrações matemáticas para promover uma representação adequada para cada termo do vocabulário, preferencialmente considerando o seu significado dentro de um determinado contexto. Estudos recentes demonstram que algoritmos de vetorização de palavras (*word embeddings*) podem prover tais representações de forma eficiente, sendo capazes de capturar relações de similaridade sintática e semântica entre as palavras em função do contexto em que ocorrem (MIKOLOV et al., 2013a, 2013b; HARTMAN et al., 2017).

O processo de vetorização de palavras demanda a utilização de um grande volume de dados no formato texto (corpus), porém muitas vezes esses dados não estão disponíveis para utilização no escopo dos projetos de PLN em desenvolvimento. Uma técnica comumente aplicável é denominada transferência de aprendizado (*transfer learning*), que consiste em utilizar nos algoritmos de PLN modelos de *embeddings* pré-treinados a partir de um corpus de contexto geral (CER et al., 2018). Algumas das principais implementações disponibilizam para uso público os seus modelos de referência, como Word2vec (MIKOLOV et al., 2013a), FastText (BOJANOWSKI et al., 2017) e GloVe (PENNINGTON et al., 2014). Esses modelos públicos de contexto geral são comumente denominados *embeddings* globais.

Contudo, realizar a vetorização de palavras utilizando um corpus do domínio específico no qual o algoritmo de PLN será aplicado pode melhorar significativamente a qualidade dos vetores e, consequentemente, o desempenho dos algoritmos que atuem nesse domínio (LAI et al., 2016; DIAZ et al., 2016). Esses vetores são treinados a partir de elementos textuais mais representativos do domínio e, portanto, conseguem representar mais adequadamente os termos técnicos específicos de seu vocabulário. Em um determinado contexto técnico, uma palavra pode assumir um significado completamente distinto, demandando assim representações adequadas a esta especialidade. Esses *embeddings* são chamados de domínio específico, ou locais.

Apenas recentemente alguns estudos se propuseram a disponibilizar modelos de *word embeddings* globais para o idioma português (BOJANOWSKI et al., 2017; HARTMAN et al., 2017; RODRIGUES et al., 2016). Especificamente na área de conhecimento de óleo e gás, até o momento da elaboração deste trabalho, não encontramos nenhum estudo publicamente disponível que se proponha a oferecer modelos de vetorização de palavras treinados em um corpus significativamente representativo desse domínio em português.

Este estudo se propõe, portanto, a descrever o processo de geração de modelos de vetorização de palavra (*word embeddings*), treinados em um corpus em português da área de O&G composto de milhares de documentos públicos nesse domínio de conhecimento, tais como boletins,

artigos e estudos técnicos, publicações, teses e dissertações, disponibilizados por instituições de referência nessa área de conhecimento. Os modelos são gerados utilizando três dos principais algoritmos disponíveis: Word2vec, FastText e GloVe. A principal contribuição deste trabalho está em disponibilizar para uso público o corpus utilizado, os modelos gerados, assim como o conjunto de scripts e algoritmos implementados neste processo, acessíveis a partir do repositório: https://github.com/diogosmg/wordEmbeddingsOG. Espera-se que diversas iniciativas em andamento na área O&G com uso de PLN, observadas tanto na indústria quanto na comunidade acadêmica, possam ser beneficiadas com a disponibilização destes modelos.

2. O corpus

O primeiro passo deste trabalho, portanto, trata da composição de um corpus significativo utilizando bases textuais do domínio de O&G, obtidas a partir de fontes públicas disponibilizadas por instituições de referência nesta área de conhecimento, especificamente Petrobras e Agência Nacional do Petróleo, Gás e Biocombustíveis (ANP). Para este fim, reuniu-se um conjunto de 202 boletins técnicos publicados pela Petrobras, contendo ao todo cerca de 2000 artigos na área de Petróleo. Adicionalmente, a fim de se obter uma melhor representação do vocabulário técnico para composição do corpus, obteve-se um conjunto de 625 documentos publicados pela ANP, correspondendo a 316 documentos acadêmicos (teses, dissertações e monografias na área de petróleo, elaboradas no contexto de seu Programa de Recursos Humanos), e mais 309 publicações periódicas, notas e estudos técnicos. Para enriquecimento do vocabulário, foram incluídos dois glossários (contendo aproximadamente 560 termos técnicos) e um dicionário de siglas (com 1255 termos), também publicados pela Petrobras e pela ANP. Ao todo, o conjunto de dados brutos inicialmente utilizados no corpus conta com cerca de 20,2 milhões de *tokens* sendo analisados. *Token* é a unidade mínima para avaliarmos o tamanho de um corpus e representa uma palavra após o preprocessamento do documento. A composição do corpus é um trabalho contínuo e progressivo, com amplo potencial para adição de novas bases.

Os arquivos obtidos no formato PDF foram convertidos para o formato texto simples, através de um script desenvolvido com a ferramenta Apache Tika (MATTMANN e ZITTING, 2011) para extrair os elementos textuais a partir dos documentos originais. A realização de experimentações adicionais sobre melhorias de extração com técnicas de reconhecimento de caracteres (OCR - *optical character recognition*) estão fora do escopo deste estudo, embora possam contribuir para uma melhor qualidade do texto extraído.

Uma vez convertidos para o formato texto, os arquivos foram então submetidos a uma série de pré-processamentos para sua correta adequação aos propósitos de treinamento, observando-se as recomendações propostas por Hartman et al. (2017) e Rodrigues et al. (2016). Neste contexto, diversas técnicas foram aplicadas de forma gradativa, observando o efeito gerado na distribuição das palavras mais comuns e na composição total do vocabulário. Primeiramente, todos os caracteres foram convertidos para sua representação minúscula, e foram eliminados os caracteres de pontuação e números, não aplicáveis aos propósitos deste estudo. Observou-se a princípio uma ampla divergência no formato de palavras acentuadas, o que motivou a uniformização desta grafia substituindo todos os caracteres acentuados por sua forma equivalente não-acentuada. No decorrer do tratamento do corpus, cabe ressaltar a observação da ocorrência de muitas palavras raras (com frequência igual a 1), potencialmente associadas a erros de grafia resultantes de problemas na conversão dos arquivos em PDF para o formato texto. Estes termos foram eliminados, porém melhores técnicas de conversão dos documentos podem ser aplicáveis, buscando aumentar a qualidade final do corpus. Palavras muito comuns (*stopwords*), que para os propósitos deste trabalho não contribuem com informação útil, também foram eliminadas. Para esta finalidade, foi utilizada a biblioteca NLTK (LOPER e BIRD, 2002). Técnicas adicionais disponíveis na literatura

como redução ao radical (*stemming*) e correção ortográfica não foram aplicadas, embora tenham amplo potencial de contribuir para uma melhoria na qualidade dos modelos gerados.

A Tabela 1 descreve a composição do corpus com as bases obtidas a partir de fontes públicas, com a respectiva totalização de documentos. Cabe ressaltar o efeito da aplicação do pré-processamento nos arquivos textuais, ocasionando a redução do tamanho total do corpus em decorrência da eliminação de caracteres de pontuação e numeração, termos incorretos, raros ou muito frequentes, em função de uma melhor qualidade do texto final.

Tabela 1 – Composição do corpus no domínio de óleo e gás

Instituição	**Base**	**Documentos / termos**	**Tokens (bruto / final)**	**Vocabulário (bruto / final)**
Petrobras	Boletins Técnicos: - Boletins de Geociências - Boletins Técnicos Petrobras - Boletins de Produção de Petróleo	~2000 artigos	10.229.664 / 4.962.545	829.907 / 86.862
Petrobras e ANP	Glossário Petrobras Glossário ANP Dicionário do Petróleo-Siglário	562 termos 554 termos 1255 termos	47.470 / 30.680	12.215 / 7.334
ANP	Trabalhos finais do PRH	316 documentos	6.288.925 / 3.308.466	394.771 / 52.880
ANP	Publicações	168 documentos	2.415.849 / 1.138.685	115.345 / 18.236
ANP	Notas e Estudos Técnicos	141 documentos	1.238.429 / 669.356	91.391 / 16.922
Total			**20.220.337 / 10.109.732**	**113.934**

3. Vetorização de palavras

Algoritmos de vetorização de palavras buscam atribuir uma representação matemática para cada palavra de um vocabulário, capturando relações de similaridade sintática e semântica a partir do contexto em que ocorrem em um conjunto de dados textual (*corpus*). Essas técnicas consistem em representar cada termo como um vetor contínuo n-dimensional de valores reais, de maneira que palavras relacionadas entre si sejam posicionadas em regiões próximas no espaço vetorial criado. Dessa forma, é possível inferir relações de similaridade entre duas palavras a partir do cálculo da distância entre seus vetores, utilizando por exemplo a distância cosseno (MIKOLOV et al., 2013a).

Word2vec (MIKOLOV et al., 2013a, 2013b) é um algoritmo baseado em redes neurais para composição dos vetores de palavras, e está disponível em duas arquiteturas de treinamento: *continuous bag-of-words* (CBOW), que busca predizer uma palavra central a partir de uma janela de palavras de contexto; e *skip-gram*, que busca predizer o conjunto de palavras vizinhas a partir de uma palavra de referência. FastText (BOJANOWSKI et al., 2017) é uma variação da arquitetura word2vec, em que os *embeddings* são associados a sequências contíguas de caracteres (*n-grams*), e cada palavra é representada pelo somatório das representações de seus *n-grams*. GloVe (PENNINGTON et al., 2014) é baseado na composição de uma matriz de co-ocorrência X, de forma que cada elemento da matriz X_{ij} representa o número de ocorrências em que um termo i é observado no contexto de um termo j.

4. Treinamento

O treinamento dos modelos de *word embeddings* concentrou-se nos principais algoritmos disponíveis sob avaliação deste estudo: word2vec (MIKOLOV et al., 2013a, 2013b), FastText (BOJANOWSKI et al., 2017) e GloVe (PENNINGTON et al., 2014). Para a geração dos modelos FastText e Glove utilizou-se a ferramenta de referência disponível em seus respectivos sites oficiais. O treinamento com Word2vec e a avaliação comparativa dos diferentes modelos foram realizados utilizando a biblioteca Gensim (REHUREK e SOJKA, 2011), em função de sua versatilidade ao permitir a leitura dos modelos e a execução de operações para diferentes algoritmos.

Em função da ampla variedade de combinações possíveis para os hiperparâmetros e considerando as limitações de escopo, este estudo concentrou-se em utilizar as principais configurações para a geração dos modelos, baseadas nas recomendações descritas em estudos anteriores (MIKOLOV et al., 2013a, 2013b; HARTMAN et al., 2017; LAI et al., 2017). Os modelos foram gerados considerando os vetores com 50, 100, 200 e 300 dimensões. Segundo Lai et al. (2017), vetores de menor dimensão, como 50 ou 100 dimensões, são suficientes para fornecer desempenho satisfatório para a maior parte das aplicações de PLN. A Tabela 2 descreve os principais hiperparâmetros disponíveis para a execução do treinamento, com destaque em negrito para as opções efetivamente utilizadas.

Tabela 2 – Listagem dos principais hiperparâmetros, com destaque em negrito para as opções utilizadas

Hiperparâmetro	***Valores***
Dimensão do Vetor	**50**, **100, 200, 300**, 400, 500, 1000
Tamanho da Janela de Contexto	2, 3, **5**, 10, 15, 20
Negative Sampling	3, **5**, 10, 15
Algoritmo	cbow, **skipgram**
min. count (fastText)	2, 3, **5**, 10

5. Resultados

Os modelos gerados a partir do corpus específico foram qualitativamente analisados segundo a hipótese de que palavras relacionadas entre si em um determinado contexto terão seus respectivos vetores localizados em uma mesma região de vizinhança no espaço vetorial (MUNEEB et al., 2015). Portanto, um subconjunto do vocabulário foi selecionado, composto especificamente de termos técnicos relevantes na área de O&G. Para cada palavra, obteve-se uma listagem dos termos mais próximos, a partir do cálculo da distância cosseno entre os seus vetores (MIKOLOV et al., 2013a; SCHNABEL et al., 2015). A Tabela 3 descreve os três principais resultados para cada termo de referência, retornados em cada modelo específico nos diferentes algoritmos gerados, comparativamente com o modelo geral NILC/USP disponibilizado por Hartman et al. (2017). Observa-se que os elementos retornados pelos modelos de domínio específico estão semanticamente relacionados ao termo de referência dentro do contexto de O&G, evidenciando a capacidade do modelo em capturar estas propriedades a partir do corpus. Quando comparados com o modelo geral (NILC/USP), nota-se que este retorna elementos cujo significado denota conceitos fora do domínio de O&G, influenciado pela natureza mais abrangente da composição de seu corpus. Cabe ressaltar, em todos os casos, ocorrências ocasionais de termos com erros de grafia, potencialmente provenientes de problemas de conversão das fontes originais para o formato texto.

Tabela 3 – Resposta dos modelos a termos técnicos da área de O&G

Termo	***NILC/USP***	***Word2vecO&G***	***FastTextO&G***	***GloVeO&G***
lula	ex-presidente inácio inpcio	sapinhoa jubarte bauna	jubarte baleia trilha	sapinhoa jubarte mexilhao
campo	estádio gramado farião	reservatorio jazida potencial	oampo pampo alvo	covetorial reservatorio pocos
rocha	ribeiro alves moreira	porosidade geradora porosa	rochamatriz olheorocha rochafonte	catole idalina porosidade
falha	pane avaria desatenção	falhamento alinhamento descontinuidade	falhas falhada falhamento	falhas matacatu maragogipe
duto	tubo aquecedor orifício	oleoduto tubulacao cilindro	oleoduto gasoduto percurso	trecho diametro biapoiado
óleo	sabão alcatrão alumínio	petroleo vazar liquido	gasoleo oleosa oleoso	agua diesel volume
choke	defying remorse disciple	succao valvula bop	stemout bop pumpout	valvula estrangulador bloqueio
árvore	planta trepadeira frondosa	anm semisubmersivel submarina	arvoredenatal sbm canhoneio	natal molhada tree
fadiga	irritabilidade sonolencia salivacao	flexao desgaste tracao	trincas trincamento tenacidade	util vida extremos
bop	blues funky blues-rock	descida camisa haste	choke blowoff borbulhador	preventer gaveta blowout
calado	silvino revez quieto	costado navio conves	talado callado tdo	navio milimetros tpb
rao	meir ghozali ouyahia	pitch heave roll	arao darao crao	heave sway roll
bcs	cfd nfs nist	centrifugo equipado bombeamento	bcd iocs dcs	bombeio centrifugo submerso

Em seguida, os resultados foram analisados em diferentes visualizações gráficas, observando-se a estrutura espacial dos vetores gerados para algumas amostras de termos específicos. Inicialmente, selecionou-se um novo subconjunto de termos técnicos, representando diferentes subdomínios das áreas de geociências e O&G, composto por: *'gasolina', 'duto', 'geologia', 'sismica'* e *'paleoceno'*. Para cada termo, obtém-se os dez elementos vizinhos mais próximos, utilizando o cálculo da distância cosseno. A técnica t-SNE (VAN DER MAATEN e HINTON, 2008) foi utilizada para diminuir a dimensionalidade desses vetores de maneira a adequá-los a um gráfico bidimensional, permitindo preservar os agrupamentos de palavras ao projetá-los no subespaço vetorial. O conjunto de vetores foi processado pelo algoritmo de clusterização K-Means,

e os resultados são apresentados na Figura 1. O gráfico confirma a correspondência dos *clusters* obtidos pelo K-Means com os vizinhos mais próximos a cada termo de referência.

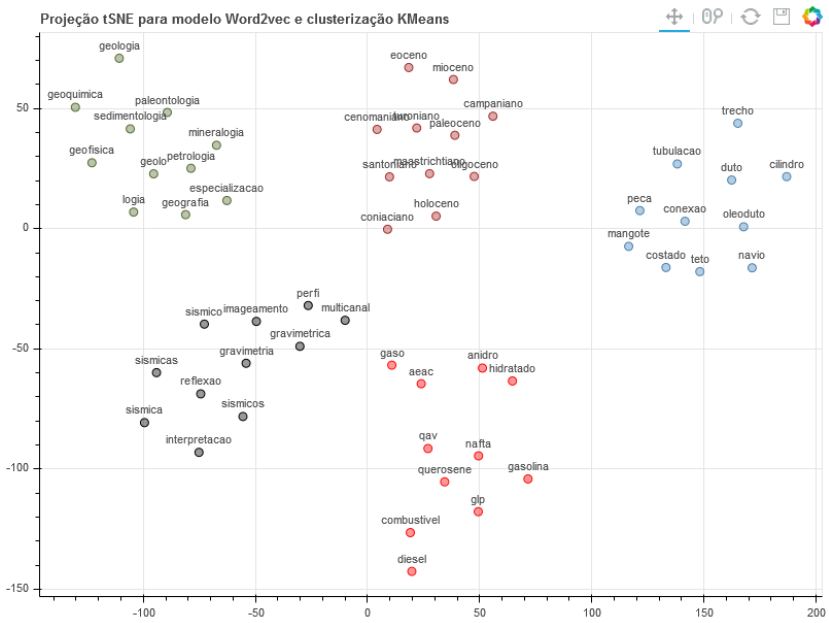

Figura 1. Projeção t-SNE para os vizinhos mais próximos aos cinco termos de referência. Nota-se que os termos mais próximos a cada palavra formam agrupamentos, conforme obtido pelo algoritmo K-Means.

Adicionalmente, para uma visualização mais abrangente da representação do vocabulário, utilizou-se t-SNE para projetar em um gráfico o subconjunto dos três mil termos mais frequentes do corpus, conforme apresentado na Figura 2. Nesse gráfico, é possível observar a formação de diversos agrupamentos de palavras relacionadas no contexto de O&G, com destaque, por exemplo, a uma região em que se observa um agrupamento contendo termos relacionados ao processo de refino (Figura 2a), e outra região contendo termos relacionados ao conceito de períodos geológicos (Figura 2b).

Os gráficos sugerem a capacidade do modelo em capturar relações de similaridade entre diversos termos técnicos presentes no corpus, considerando seus aspectos de significado particular em função do contexto em que ocorrem, posicionando-os próximos entre si no espaço vetorial e formando agrupamentos de palavras similares.

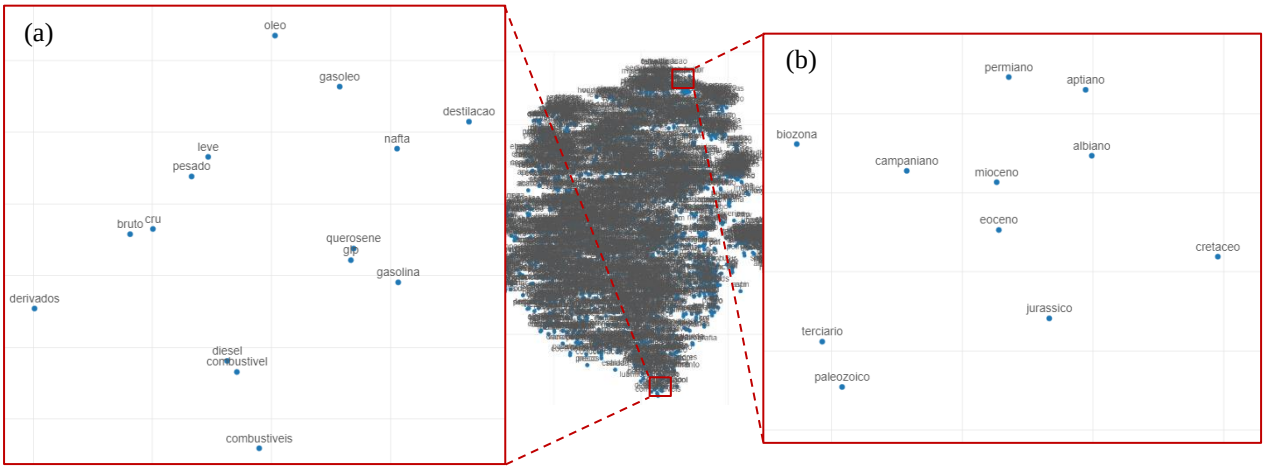

Figura 2. Principais termos do vocabulário obtidos pelo modelo word2vec com 100 dimensões, com destaque para agrupamentos de termos relacionados a: (a) refino, e (b) conceito de períodos geológicos

Por fim, selecionamos alguns conjuntos distintos de termos em subdomínios específicos da área de O&G, de forma que suas matrizes de similaridade foram graficamente representadas em uma visualização do tipo *heatmap*. Os resultados são apresentados na Figura 3, onde se observam altos índices de similaridade, conforme o esperado para os termos relacionados em cada subdomínio.

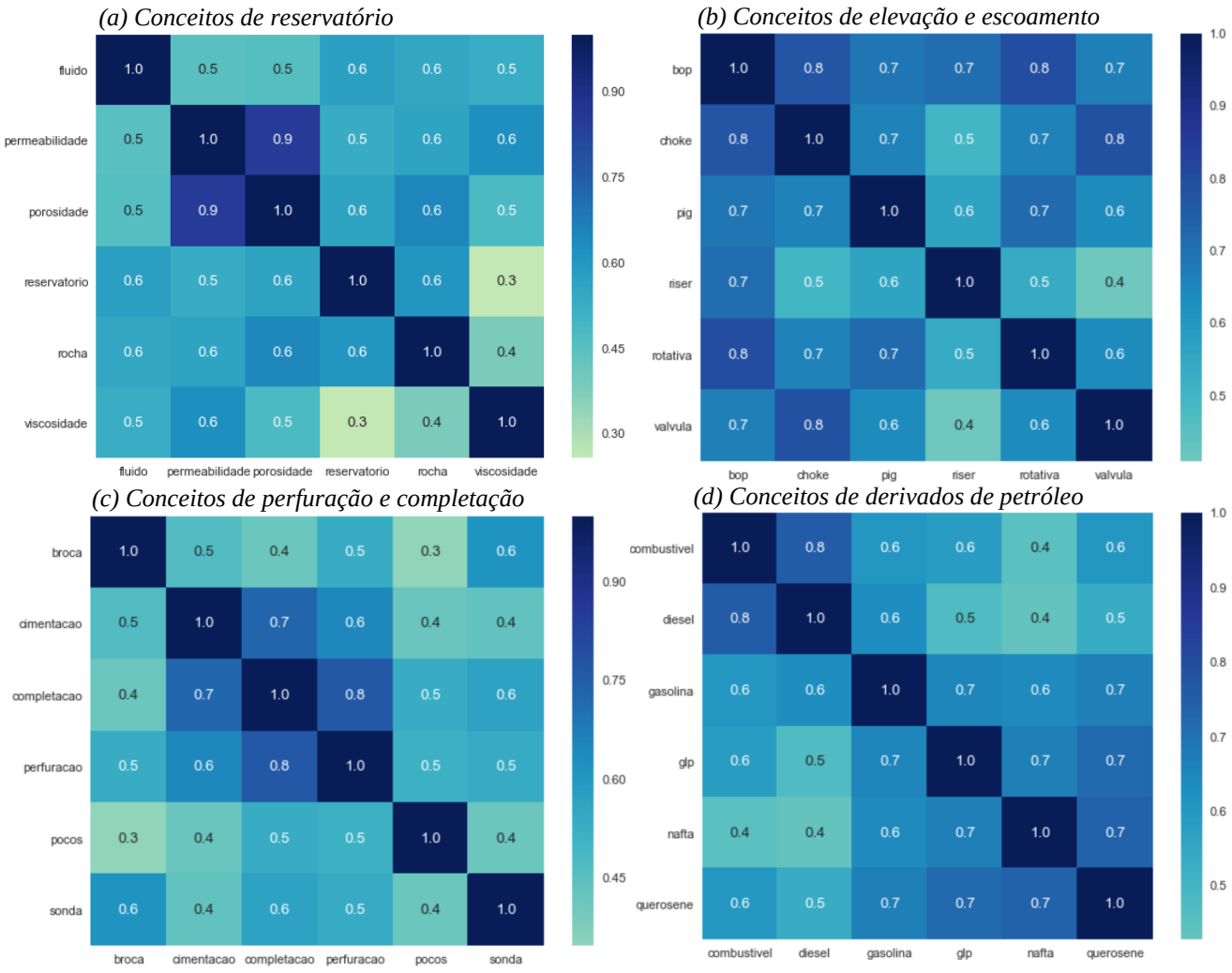

Figura 3 – Gráfico *heatmap* com matriz de similaridade entre termos de subdomínios de O&G

A Figura 4 ilustra graficamente a matriz completa de relação entre os termos dos diferentes subdomínios, onde conseguimos notar a baixa similaridade entre termos de subdomínios distintos, explicitando a formação de agrupamentos distintos e coesos.

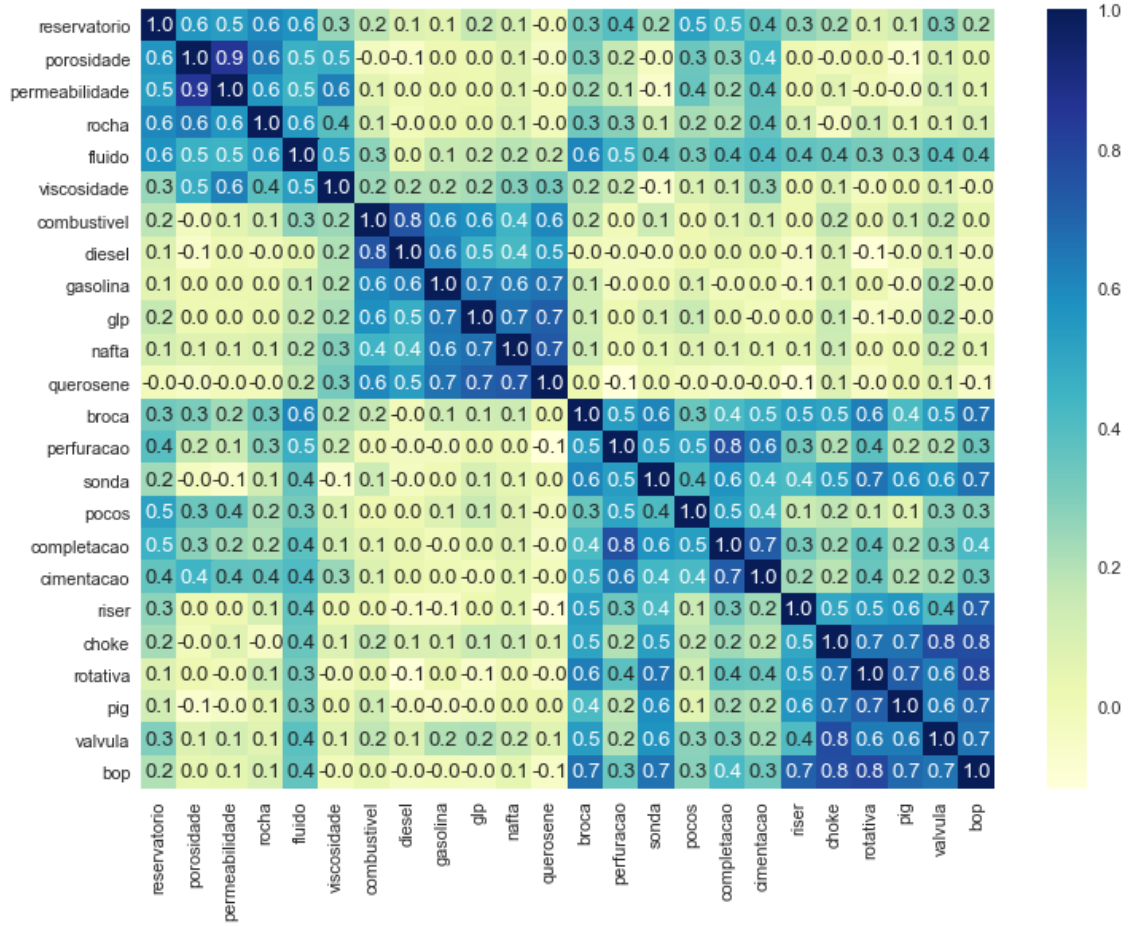

Figura 4 – Gráfico *heatmap* contendo as relações de similaridade entre os diferentes subdomínios de O&G

6. Conclusões

Este trabalho descreveu o processo de desenvolvimento de vetorização de palavras (*word embeddings*) treinados especificamente para o domínio de óleo e gás em português. Esses modelos vetoriais representam um aspecto-chave para o bom funcionamento dos principais algoritmos na área de PLN. Objetiva-se que este trabalho possa contribuir com diversas iniciativas de PLN na área de O&G, tanto na indústria como na comunidade acadêmica, que não disponham de volumes de dados suficientemente adequados para a geração de modelos próprios. Todo o material produzido será disponibilizado para uso público e poderá ser consultado no repositório destinado a este projeto em https://github.com/diogosmg/wordEmbeddingsOG.

Os resultados obtidos por avaliações qualitativas sugerem a capacidade de aprendizado do algoritmo para capturar relações de similaridade entre termos específicos da área de O&G presentes no corpus utilizado. É possível observar que, para termos técnicos do vocabulário, os modelos específicos apresentam regiões de vizinhança do espaço vetorial semanticamente mais significativas, quando considerados em relação do modelo de contexto geral.

7. Referências

BLINSTON, K., BLONDELLE, H. Machine learning systems open up access to large volumes of valuable information lying dormant in unstructured documents. *The Leading Edge*, 2017.

BOJANOWSKI, P., GRAVE, E., JOOULIN, A. MIKOLOV, T. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*. 2017.

CER, D., YANG, Y., KONG, S.-Y., HUA, N., KIMTIACO, N., JOHN, R. S., CONSTANT, N., GUAJARDO-CESPEDES, M., YUAN, S., TAR, C., SUNG, Y.-H., STROPE, B., KURZWEIL, R. Universal sentence encoder. *ArXiv*, 1803.11175, 2018.

DIAZ, F., MITRA, B., CRASWELL, N. Query Expansion with Locally-Trained Word Embeddings. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, p. 367–77. doi:10.18653/v1/P16-1035, Berlin, Germany, 2016

GOODFELLOW, I., BENGIO, Y., COURVILLE, A.. *Deep Learning.* MIT Press. Disponivel em: <http://www.deeplearningbook.org>, 2016.

HARTMAN, N. S., FONSECA, E., SHULBY, C., TREVISO, M., RODRIGUES, J. S., ALUISIO, S. Portuguese Word Embeddings: Evaluating on Word Analogies and Natural Language Tasks. *Proceedings of Symposium in Information and Human Language Technology*. Uberlandia, MG, Sociedade Brasileira de Computacão, 2017.

LAI, S., Liu, K., XU, L., ZHAO, J. How to Generate a Good Word Embedding. *IEEE Intelligent Systems*, vol. 31, no. 6, pp. 5-14. doi: 10.1109/MIS.2016.45. Nov.-Dec. 2016.

LECUN, Y., BENGIO, Y., HINTON, G. Deep Learning. *Nature*, maio 2015.

LOPER, E., BIRD, S. NLTK: the Natural Language Toolkit. *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics - Volume 1 (ETMTNLP '02)*, Vol. 1. Association for Computational Linguistics, Stroudsburg, PA, USA, 63-70. DOI: https://doi.org/10.3115/1118108.1118117. 2002

MATTMANN, C., ZITTING, J. *Tika in action*. Manning Publications Co., 2011

MIKOLOV, T., CHEN, K., CORRADO, G., DEAN, J. Efficient estimation of word representations in vector space. *ICLR Workshop*, 2013a.

MIKOLOV, T., SUTSKEVER, I., CHEN, K., CORRADO, G., DEAN, J. Distributed representations of words and phrases and their compositionality. *Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 2*(NIPS'13). 2013b.

MUNEEB, T. H., SAHU, S. K., ANAND, A. Evaluating distributed word representations for capturing semantics of biomedical concepts. *Workshop on Biomedical Natural Language Processing (BioNLP).* 2015.

PENNINGTON, J, SOCHER, R., MANNING, C. D. Glove: Global vectors for word representation, in *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543.http://www.aclweb.org/anthology/D14-1162, 2014.

REHUREK, R., SOJKA, P. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics*, Masaryk University, Brno, Czech Republic, 2011.

RODRIGUES, J., BRANCO, A., NEALE, S., SILVA, J. LX-DSemVectors: Distributional Semantics Models for Portuguese. *Computational Processing of the Portuguese Language: 12th International Conference (PROPOR-2016)*. Springer International Publishing, 2016.

SCHNABEL, T., LABUTOV, I., MIMNO, D., JOACHIMS, T. Evaluation methods for unsupervised word embeddings. *EMNLP*. 2015.

USP. Repositório de Word Embeddings do NILC. *NILC - Núcleo Interinstitucional de Linguística Computacional*. Disponível em: <http://www.nilc.icmc.usp.br/embeddings>

VAN DER MAATEN, L.J.P., HINTON, G. Visualizing High-Dimensional Data Using t-SNE. *Journal of Machine Learning Research 9*: pp. 2579-2605, 2008.