XLI CILAMCE
XLI Ibero-Latin American Congress on Computational Methods in Engineering
16-19 November, 2020 FOZ DO IGUAÇU PARANÁ - BRAZIL

Application of Natural Language Processing to Search for Business Opportunities for the raw Glycerol in Biodiesel Production

Andressa Nery Lopes[1], Alexandre D'Elia[1], Beatriz Souza Leite Pires de Lima[1], Nelson F.F. Ebecken[1]

[1] *Programa de Engenharia Civil, COPPE/ Universidade Federal do Rio de Janeiro*
CEP: 21941-909 Rio de Janeiro – RJ- Brasil
lopes.an@live.com, delialexandre97@gmail.com, bia@coc.ufrj.br, nelson@ntt.ufrj.

Abstract.

Recently, in Brazil, the level of biodiesel added to oil diesel fuel has expanded because of compulsory rules. The increase of biodiesel production has provided an excess of crude glycerol which is its primary co-product. This paper intends to search business opportunities for the excess of crude glycerol by a Natural Language Processing Approach. This process of exploring and analyzing unstructured data was performed by text mining thousands of recovered abstracts from the scientific literature available on online research platforms over ten years. The methodology involved the information retrieval and compilation of a database for the textual content analysis and the use of word embedding concepts including Word2vec and Bert. The results indicated the tendencies for producing value-added substances from crude glycerol over ten years of research.

Keywords: biodiesel, glycerol, text analysis, Natural Language Processing.

1 Introdução

A GLICERINA é o principal coproduto da produção de biodiesel e a sua transformação em matéria-prima tem sido registrada amplamente na literatura científica Bagheri [1], Beltram [2], Garlapati [3] e Skrzyriska [4]. Além disso, pode ser convertida em centenas de substâncias químicas com aplicações específicas, que vão desde a indústria alimentícia até a construção civil.

O registro destes inúmeros usos que a glicerina recebeu como excedente da produção mundial de biodiesel representa um aporte de informações, as quais estão publicadas em artigos de caráter científico e disponíveis em plataformas de pesquisa online. Estas informações podem ser transformadas em um grande banco de dados para análise através de ferramentas de mineração de textos, o que não seria possível fazer por um único indivíduo.

A mineração de dados textuais através de uma quantidade massiva de resumos de artigos científicos que relatam usos e aplicações sobre a glicerina resultante da produção de biodiesel pode revelar padrões de informações que representam usos diretos (sem a necessidade de processo de refinamento), produtos de valor agregado e principalmente novas áreas de aplicação. Estes usos e aplicações surgiram ao longo dos últimos 10 anos em função da quantidade demasiada disponível de glicerina oriunda da indústria de bioenergia. O biodiesel compõe a matriz energética brasileira desde 2004, quando foi implantado o Programa Nacional de Uso e Produção do Biodiesel – PNPB.

Em 1917 existiam 51 usinas de biodiesel capazes de produzir um volume de 26602,26 m³/dia e a porcentagem obrigatória de biodiesel adicionada ao diesel é de 8% (cujo percentual poderá ser aumentado para 9%, até 2018; e 10%, até 2019) de acordo com a lei nº 13.263 (23 de março de 2016).

A transesterificação é o processo mais utilizado nas plantas produtoras de biodiesel e consiste na obtenção de um éster (biodiesel) a partir de outro éster (triglicerídeos de origem vegetal ou animal) reagindo com um álcool que possui a função de solvente, com a presença de um catalisador, geralmente, um ácido ou base forte. Além do biodiesel, um biocombustível resultante, gera-se a glicerina como principal coproduto, também conhecida como um álcool trivalente.

A indústria de biodiesel no Brasil produziu 390 mil m³ de glicerina bruta em 2015, sendo que 63% deste

volume (246 mil m³) foram exportados. A China importou 80% deste volume. Esta alta demanda chinesa pode ser justificada pelo consumo de produtos de cuidado pessoal, como os cosméticos. Entretanto, tal excedente produzido de glicerina bruta no país ao invés de ser vendida, poderia ser transformada em matéria-prima ou ainda, convertida em produtos de alto valor agregado, fornecendo lucro às usinas. Do mesmo modo, estas substâncias podem representar uma oportunidade de negócio para os produtores de biodiesel.

2 Processamento de Linguagem Natural

A mineração de dados consiste na aplicação de métodos inteligentes e de ferramentas computacionais para auxiliar especialistas na análise e descoberta de conhecimento de grandes volumes de dados Thomas [5]. Esta área, também conhecida como *Data Mining* (ou ainda *Knowledge Discovery from Data* – KDD), ganhou uma rápida expansão nas últimas décadas e possui relevante atuação em campos interdisciplinares.

A mineração de textos (*Text Mining* ou *Knowledge Discovery from Text* – KDT) é uma subárea da mineração de dados e, de forma análoga faz parte do processo de descoberta de conhecimentos através de dados textuais. Ela possui como objetivo extrair padrões ou informações relevantes a partir de documentos de texto.

Text Mining ou Processamento de Linguagem Natural pode ser definido ainda como processamento intensivo de conhecimento em que o usuário interage com uma coleção de documentos várias vezes usando uma suíte de ferramentas de análises.

Neste trabalho, utilizamos como coleção de documentos os resumos de artigos científicos. Apesar do corpo do texto de um artigo científico conter a maior parte da informação, o resumo sintetiza o conteúdo do artigo, reunindo uma melhor variação de palavras-chave, além de estar quase sempre disponível para download nas plataformas online de pesquisa, mesmo quando o artigo completo não é de livre acesso.

Alguns artigos científicos em formato PDF encontram-se bloqueados como imagem ou por autores, e isto dificulta a leitura automática por meio de algoritmos do seu conteúdo textual. Por isso, outra vantagem para a utilização de resumos é que eles possuem poucos erros de leitura de OCR, tecnologia que converte imagem em texto.

- Este trabalho consiste na utilização de resumos (*abstracts*) de artigos científicos como alvo para realizar uma análise semi-supervisionada por meio de técnicas de mineração de texto. A metodologia envolveu agrupamento de dados textuais por data de publicação, classificação usando o algoritmo SVM (*Support Vector Machine,[9])* e quantificação das frequências por ano para estudo de tendências e oportunidade de negócio.

A glicerina tem sido mencionada na literatura científica de diferentes formas, e para tanto foi estabelecido um conjunto de termos e frases, no idioma inglês, como: glycerol; glycerin e biodiesel; glycerine; “crude glycerol”; “crude glycerin”; “crude glycerine”; “glycerol from biodiesel”; “glycerin from biodiesel” and “glycerine from biodiesel”. Desta forma, delimitou-se o *download* de resumos de artigos científicos que continham informações o mais próximo possível referente à glicerina proveniente da produção de biodiesel.

Em seguida, partiu-se para a fase de recuperação de informação. Eesta etapa refere-se ao *download* de resumos das plataformas *Science Direct* e a *Web of Knowledge* por conterem uma base de dados multidisciplinar de periódicos, livros, capítulos de livros, etc. Optou-se por delimitar o período das publicações para recuperar os resumos a partir de 2007 até 2016 (10 anos). Vale ressaltar que os campos de pesquisa utilizados para a busca de palavras-chave foram “**resumo**” ou “**título**”, sendo este último utilizado quando a base de dados não permitia busca por resumo.

Após esta pesquisa inicial, os resumos e referências foram exportados das plataformas de pesquisa online e depois organizados numa biblioteca de referências. A partir daí, eles são exportados no formato “.txt” e então convertidos para um formato mais comum que é o “.csv”.

A etapa seguinte trata da limpeza do banco de dados inicial. Esta etapa possui papel relevante uma vez que o banco de dados inicial contém resíduos, tais como, referências sem resumos e vice-versa, resumos publicados de forma repetida (presente em várias bases de dados), ou ainda, com data de publicação em desacordo com o período de delimitação do trabalho. A etapa da leitura reconheceu o banco de dados no formato “.csv”. E em seguida foi gerado um banco de dados bruto. A filtragem consistiu na exclusão dos resumos publicados no período anterior ao ano de 2007, uma vez que ao fazer o *download* dos resumos nem todos os sites das plataformas de pesquisa permitiam essa delimitação nos campos de busca. A eliminação dos resumos de artigos duplicados foi necessária, pois observou-se que algumas vezes um mesmo resumo era encontrado em duas ou mais plataformas de pesquisa, o que poderia interferir nos resultados. A correção ortográfica dos textos foi realizada de acordo com a língua inglesa. O dicionário padrão utilizado como referência foi o WordNet 3.0. E finalmente, o banco de dados para análise textual foi então gerado.

Depois de realizar as etapas de limpeza dos dados textuais foram criados descritores que representam substâncias químicas de valor agregado (Tab. 1). Com a inserção destes descritores executou-se uma análise semi-supervisionada, em que os resumos foram classificados parcialmente.

TABELA 1
DESCRITORES PARA BUSCA DE PALAVRAS-CHAVE

Palavras-chave	**Descritores em linguagem PDL**
Acrylic acid	"acrylic acid" or "propenoic acid"
Epichlorohydrin	epichlorohydrin
Propane	propane
Solketal	solketal or ketal
Triacetin	triacetin or "1,2,3-triacetoxypropane" or (triacetate AND (glycerin or glycerol)))
Vitamin	vitamin

A classificação através do algoritmo SVM foi aplicada isoladamente para cada palavra-chave e seus respectivos descritores para formar bancos de dados individuais. Em seguida, realizou-se uma extração de frases ao redor das palavras-chave presentes nos resumos. A extração de frases permitiu a captura de tópicos para geração de *tagcloud*. Neste caso, o algoritmo procede para identificar a ocorrência dos tópicos mais significativos. Esta ocorrência é projetada sob a forma de fontes maiores de palavras (ou cor vermelha, verde) para altas frequências e fontes menores (ou cor azul) para baixas frequências.

A última etapa deste trabalho consistiu em verificar se existe uma diferença significativa quanto à presença dos termos "glicerina purificada" ("*pure*") e "glicerina bruta" ("*crude*") nos resumos, também ao longo do período de delimitação do trabalho. Observa-se, entretanto, que as citações destes termos não garantem que de fato ocorreu o processo de refinamento da glicerina, porque nos resumos encontram-se apenas as ideias principais dos trabalhos científicos.

3 Análise dos documentos

O banco de dados compilado resultou em 11.382 resumos recuperados, de cunho científico, desde o ano de 2007 até 2016, em diferentes plataformas de pesquisa online. Após o pré-processamento dos dados textuais não-estruturados e da etapa de limpeza, o banco de dados foi reduzido a 8.724 resumos.

A descoberta automática de tópicos permitiu a geração de seis tagclouds, uma para cada substância: a) acrylic acid, b) epichlorohydrin, c) propane, d) solketal, e) triacetin, f) vitamin.

A maioria das palavras projetadas ao redor das palavras-chave equivale a características físico-químicas da substância, do processo de reação química, ou ainda, sinônimos destes termos que são as nomenclaturas de acordo com a IUPAC – União internacional da Química Pura e Aplicada.

Entretanto, as palavras destacadas, na maior parte adjetivos, representaram altas frequências e evidenciaram a valorização destas substâncias que podem ser produzidas a partir da glicerina do biodiesel, a exemplo: "*high-value*", "*value-added*" e "*valorization*".

A primeira delas, o ácido acrílico, é um produto químico amplamente utilizado na indústria de polímeros. Seu processo de produção pode ser a base de petróleo, como também a partir da glicerina reaproveitada do processo de produção do biodiesel, a um custo que varia de US $1.09/0,45Kg [15] a US $2200 (Ton) cogitado no *e-commerce* (Tab. 2).

Tabela 2
VARIAÇÃO DE PREÇOS DAS SUBSTÂNCIAS

Palavra-chave	Preço (mín-máx)*
Acrylic acid	US $500-2200/Tonelada
Epichlorohydrin	US $10-1300/Tonelada
Propane	US $10-30/Parte
Solketal	US $1-999/Quilograma
Triacetin	US $850-1400/Tonelada
Vitamin	Vitamina B12: US $1-5000/Quilograma

*Pesquisa realizada em 12 jul. 2017. Fonte: Grupo Alibaba (*alibaba.com*). OBS: Os valores apresentados não levam em consideração o reaproveitamento da glicerina.

A epicloridrina é conhecida como um composto utilizado para produção de resinas epóxi, plásticos e adesivos, além de deter um alto poder carcinogênico. Alguns autores realizaram a purificação da glicerina do biodiesel por meio de reação com HCl em ácido hexanóico e depois a neutralização através da reação final com NaOH para produzir epicloridrina, a um custo de US $1,51/Kg. O preço no *e-commerce* pode ser encontrado a partir de US $10/Ton.

O propano (C_3H_8) é comumente reconhecido como gás GLP (gás liquefeito de petróleo) e vendido a (US $10-30/Parte) como combustível para uso doméstico. Mas publicações recentes demonstram que a sua geração pode ser realizada através da glicerina do biodiesel, que por ser de fonte renovável, transforma-o em "bio-propano" reduzindo as emissões nocivas à atmosfera. Por sua vez, o solketal vem sendo mencionado na literatura como um aditivo que pode ser potencialmente utilizado na melhoria de desempenho de combustíveis, como a gasolina, o diesel e ainda para o biodiesel, a partir do processo de acetalização do glicerol. É encontrado a um valor que varia entre US $1-999/Kg. A triacetina é amplamente usada na indústria de cosméticos, de tabaco e, recentemente, como aditivo para o biodiesel. De acordo com a tabela de preços, é a segunda substância de maior valor agregado, ficando atrás apenas do ácido acrílico a um preço de até US $1400/Ton.

No *tagcloud* , as vitaminas que mais se destacaram, e portanto, de maiores freqüências foram: a vitamina B12 (cobalamina) conhecida por manter o bom funcionamento do sistema nervoso, a vitamina E usualmente utilizada para rejuvenescimento da pele, e a vitamina D3 que ajuda na manutenção de cálcio no sangue. Porém, a vitamina B12 foi a mais citada e também a que apresentou o maior valor de mercado dentre todas as outras mencionadas, a um valor máximo de até US $5000/Kg no *e-commerce*.

[Kośmider et al, 2012] registraram que uma bactéria cultivada a partir da glicerina bruta favoreceu a um aumento de 93% da concentração final de vitamina B12, o que representa a descoberta de uma rota de reaproveitamento da glicerina ainda mais econômica, pois não houve a necessidade do processo de purificação, que geralmente demanda altos custos. A projeção da frequência por ano dos descritores selecionados ao longo de 10 anos de publicações, favoreceu a descoberta de tendências em relação aos produtos de valor agregado derivados da glicerina. Como pode ser visto nos gráficos a seguir, todas as palavras-chave apresentaram frequência de termos variada, sendo a maior quantidade de citações identificada para a substância triacetina. Esta substância possui frequência constante ao longo de 10 anos, período de delimitação do trabalho, mostrando uma tendência predominante de reaproveitamento da glicerina para sua geração. O ácido acrílico também apresentou altas frequências seguidamente ao longo do período, porém somente a partir de 2010. Também foram encontradas citações nos resumos para epicloridrina (exceto em 2009 e 2013) e o propano (exceto em 2013). O solketal só não apresentou citação no ano de 2007, mas esteve presente nos resumos em todos os anos seguintes.

4 Análise usando *Word Embeddings*

As duas maneiras mais usuais de se representar palavras como vetores são a representação *one-hot encoding* (vetores esparsos) e a representação *word embeddings* (vetores densos). Esse tipo de vetorização considera um vocabulário de tamanho *N* e atribui um índice *i* para cada um dos tokens do vocabulário. Assim, cada token é representado por um vetor de tamanho *N*, em que todas as posições desse vetor são preenchidas com zeros, exceto a posição *i* (correspondente a posição do token em questão) que será preenchida com o valor 1. Logo, cada token será representado por um vetor de tamanho fixo (tamanho do vocabulário).

Um problema para esse tipo de representação é o espaço necessário para se armazenar cada palavra, pois, geralmente os vocabulários costumam ter centenas de milhares de palavras e seria inviável que as palavras fossem representadas por vetores de tão alta dimensionalidade.

Outro problema desse tipo de representação é que ela não é capaz de refletir o relacionamento entre as palavras, pois, como as palavras são representadas por vetores ortogonais, todas as palavras sempre estarão igualmente distantes. Ou seja, a distância entre as palavras "hotel" e "motel", que possuem significados próximos, seria a mesma do que a de qualquer outro par de palavras. Esse tipo de representação costuma ser utilizado para representar as diferentes classes em um processo de classificação, uma vez que em tal processo as classes realmente costumam ser igualmente distantes, e o número de classes distintas é limitado.

Word embedding é uma forma de representação textual que utiliza vetores densos, de baixa dimensionalidade, cujos valores são aprendidos a partir do próprio texto. A *word embedding* utiliza a vizinhança de cada uma das palavras do texto para formular a representação das palavras. Isso permite a criação de um vetor denso que representa a projeção de cada palavra. Dessa forma, a *word embedding* representa as coordenadas da palavra no espaço vetorial que foi aprendido a partir do texto. Sendo assim, os relacionamentos geométricos entre os vetores de palavras devem refletir os relacionamentos semânticos entre essas palavras. Utilizando-se essa abordagem, pode-se comparar a relação entre duas palavras quaisquer a partir da comparação dos vetores que as representam.

Esse tipo de representação surgiu da necessidade de se expressar as características de similaridade entre as palavras, de forma que isso pudesse ser aproveitado no contexto das aplicações.

Sendo assim, passou-se a explorar o conceito de modelagem estatística da linguagem. Esse modelo permite a previsão da palavra seguinte, levando-se em consideração as anteriores. Essa ideia é uma extensão dos modelos utilizados no tratamento de séries temporais, pois da mesma forma que em sistemas lineares, o estado seguinte pode ser determinado pela combinação dos estados anteriores, pode-se prever a palavra seguinte de um determinado texto, com base nas palavras que ocorreram anteriormente nesse mesmo texto.

Essa constatação já havia sido feita por linguistas, que afirmavam que as palavras vizinhas estavam relacionadas semanticamente, logo, palavras semelhantes ocorreriam em contextos semelhantes. Nesse mesmo sentido, pode-se afirmar que é possível conhecer uma palavra através das palavras que a acompanham *("You shall know a word by the company it keeps!"*). Dessa forma, conclui-se que as palavras não ocorrem em contextos independentes, uma palavra sempre está relacionada com as palavras que vêm antes e depois.

Sendo assim, pode-se destacar duas características importantes do modelo de linguagem. A probabilidade de se encontrar uma palavra em um determinado texto é função da ocorrência de todas as palavras anteriores. Uma palavra sempre está relacionada com seus vizinhos. Logo, a partir dessas premissas, formulou-se a seguinte expressão para se representar a probabilidade de ocorrência de uma determinada palavra em uma sequência de dados textuais.

$$p\left(w^{(t)}\right) = \prod_{k=1}^{k=t-1} p\left(w^{(k)} | \{w^{(1)}, \dots, w^{(k-1)}\}\right) \tag{1}$$

Ou seja, a probabilidade de ocorrência de uma determinada palavra em um texto será igual a produto das probabilidades de ocorrência das palavras que ocorreram antes dela nesse mesmo texto.

No entanto, essa é apenas uma formulação teórica, pois na maioria das situações não é possível utilizar todas as palavras anteriores (desde o início do texto) para se prever a próxima. Então, para tornar essa premissa viável de ser aplicada em situações práticas, faz-se uma aproximação, e em vez de se considerar todas as palavras para o cálculo da probabilidade da palavra seguinte, realiza-se esse cálculo com base em uma janela de tamanho fixo. Ou seja, calcula-se a probabilidade de uma palavra com base nas *n* palavras anteriores (onde *n* é o tamanho da janela a ser considerada). Dessa forma, a fórmula pode ser simplificada para a seguinte expressão:

$$p\left(w^{(t)}\right) \approx \prod_{k=t-n+1}^{k=t-1} p\left(w^{(k)} | \{w^{(k-1)}, \dots, w^{(k-n)}\}\right) \tag{2}$$

Isso permite a modelagem de qualquer palavra do texto em função de um número fixo de parâmetros, o que torna tal representação muito mais concisa. A partir dessa consideração, torna-se possível desenvolver os modelos de word embeddings que são utilizados para modelagem textual. Atualmente, existem uma série de modelos de word embeddings.

Apesar de algumas tentativas anteriores, as primeiras iniciativas que apresentaram resultados satisfatórios foram as propostas em 2013 [Milokov et al, 2013]. Esses trabalhos apresentam a representação distribuída de palavras obtida a partir da utilização de uma rede neural de duas camadas. Tal modelo ficou conhecido como o Word2Vec, dividindo-se em 2 variantes: Word2Vec CBOW e Word2Vec Skip-Gran.

O modelo utiliza uma rede neural de 2 camadas, cujo objetivo principal não consiste em aprender a fazer a predição da palavra em si, mas sim em aprender a melhor representação vetorial das palavras, sendo que, essa representação é dada pelos pesos encontrados na camada de entrada da rede neural (os pesos correspondem a representação vetorial da palavra).

Neste trabalho, o arquivo com os dados textuais dos resumos científicos foi preparado para o word2vec e processado obtendo resultados semelhantes às palavras selecionadas anteriormente. Estes resultados estão mostrados na tabela 3 com as respectivas similaridades encontradas. Assim feita a análise de negócios, as mesmas oportunidades são obtidas. Oito substâncias importantes foram identificadas pelo wor2vec: biodiesel glycerol, acrylic acid, epichlorohydrin, propane, solketal, triacetin, and vitamin.

Tabela 3 - Substâncias Selecionadas e similaridades

biodiesel	
Transesterfication	0,7179
Wpco	0,7161
Macrosiphon	0,7060
Salvia	0,7052
Complying	0,7019
Deliberated	0,6939
Lunaria	0,6896
Foetida	0,6890
Nevo	0,6884
Manghas	0,6876

glycerol	
aph	0,7001
glycerin	0,6504
lsr	0,6355
unexpectedly	0,6319
stoichiometrically	0,6234
conversely	0,6222
actually	0,6199
blw	0,6175
propanediol	0,6159
coproduction	0,6144

acrylic	
glyceric	0,7570
polyglycerols	0,7545
acrylonitrile	0,7515
dcp	0,7357
oxydehydration	0,7352
acetalisation	0,7332
regioselective	0,7325
silicotungstic	0,7305
tartronic	0,7279
arenesulfonic	0,7278

epichlorohydrin	
dow	0,7984
ech	0,7709
proprietary	0,7621
solvay	0,7571
tavaux	0,7431
huntsman	0,7304
Cargill	0,7292
Corp	0,7277
Epicerol	0,7268
Rohm	0,7201

propane	
tetralin	0,8021
aromatization	0,7951
thermodynamically	0,7917
ethane	0,7897
preferably	0,7731
pressurized	0,7697
azeotropic	0,7617
propene	0,7605
nucleophiles	0,7493
reformation	0,7438

solketal	
ketals	0,8444
ethers	0,8381
acetals	0,8368
acetal	0,8188
oxygenate	0,8099
ttbg	0,8068
diglycerol	0,8022
isobutylene	0,8013
isobutene	0,7990
acetalization	0,7936

triacetin	
diacetin	0,8600
monoacetin	0,8159
triacetate	0,7885
triacetylglycerol	0,7779
glyceryl	0,7740
monooleate	0,7712
triacetyl	0,7653
acetates	0,7620
acetal	0,7507
triglycerol	0,7492

vitamin	
saccharolyticum	0,7872
rhamnosus	0,7658
senegalensis	0,7593
monosodium	0,7569
cosubstrate	0,7543
ye	0,7539
tryptone	0,7530
vancomycin	0,7514
mph	0,7495
acremonium	0,7487

Os resultados mostram novamente que triacetin e vitamina B12 são excelentes oportunidades de negócio.

O BERT [Devlin et al, 2018], ao contrário do word2vec, não gera *word embeddings* únicos para cada palavra ao final de seu treinamento. O que ele faz é fornecer um modelo que, dada uma frase inteira, irá gerar um *word embedding* para cada palavra dentro do contexto da frase. A arquitetura utilizada para o BERT é construída de tal forma que, ao treinar e ao predizer os *embeddings*, tanto as palavras anteriores quanto seguintes são levadas em conta, e mecanismos de atenção são utilizados para guardar a informação das palavras mais e menos relevantes na frase.

Assim, BERT trata das diferenças entre os sentidos das palavras, mas não revela quais são os diferentes sentidos da palavra.

O treinamento de um modelo BERT envolve duas tarefas: Masked LM e Next Sentence Prediction. Na primeira tarefa (Masked LM), o objetivo é "mascarar" e prever a palavra mascarada. Assim, BERT mascara até 15% das palavras de uma sentença. Isso significa que 15% dos tokens de uma frase serão substituídos pelo símbolo [MASK]. O modelo então tenta predizer o token original, levando em conta o contexto das outras palavras não ocultadas da sequência. Há ainda algumas regras de mascaramento.

Os autores do BERT acreditam que este tipo de treinamento é o que permite maior poder de representação com relação ao treinamento de um modelo Word Embeddings tradicional. A segunda tarefa (Next Sentence Prediction), no processo de treinamento de um modelo BERT, recebe como input pares de sentenças (S1, S2), em que o modelo deve predizer se a sentença S2 é a subsequente a S1. Durante o treinamento, 50% das entradas são um par em que S2 é, de fato, a sentença subsequente, enquanto nos outros 50% uma sentença aleatória do corpus é escolhida.
Embeddings de palavras pré-treinados são essencialmente embeddings de palavras obtidos por meio do treinamento de um modelo não supervisionado em um corpus. O treinamento não supervisionado, neste caso, normalmente envolve a previsão de uma palavra com base em uma ou mais das palavras ao redor. A palavra embeddings obtida a partir de tal processo de treinamento, junto com o modelo treinado, pode então ser usada, em muitos casos, para uma tarefa supervisionada como tarefas de marcação com dados rotulados, muitas vezes muito menos do que o necessário para treinar um modelo do zero - os embeddings pré-treinados junto com o modelo são ajustados para a tarefa específica.

Podemos usar os embeddings pré-treinados junto com o modelo para: Opcionalmente, treinamento adicional de forma não supervisionada em um corpus específico de domínio. Este treinamento não supervisionado adicional muitas vezes melhora o desempenho do modelo em tarefas downstream específicas do domínio Ajustar um modelo pré-treinado para uma tarefa específica usando dados rotulados para essa tarefa (normalmente uma camada adicional é empilhada em cima do modelo para ajuste fino - pesos do modelo são atualizados / ajustados para a tarefa específica)

5 Conclusões

A utilização da metodologia apresentada forneceu um banco de dados estruturado e livre de resíduos para posterior consulta. Foi possível notar ainda, a presença dos termos glycerol e biodiesel em todas as tagclouds

formadas reforçando o fato de que estas substâncias podem ser produzidas a partir da glicerina como matéria-prima. Em relação ao agrupamento de dados textuais por ano de publicação, foi possível perceber que nos últimos cinco anos ocorreram as maiores frequências das palavras-chave nos resumos. Este fato coincide com o aumento do volume da produção mundial de biodiesel, e consequente acúmulo da quantidade de glicerina disponível no mercado sem reutilização direta. Isto pode ter proporcionado o interesse de pesquisadores para estudo do reaproveitamento deste excedente. O acréscimo de descritores para classificar os grupos de resumos entre "pure" e "crude", acabou limitando a ação do algoritmo e diminuiu significativamente a frequência dos termos.

Ainda assim, é notório o registro de dezenas de trabalhos científicos que exploraram diferentes rotas e processos químicos quanto ao reuso da glicerina bruta ou refinada para a produção de substâncias de valor agregado. A triacetina e o ácido acrílico apresentaram as melhores oportunidades de negócio neste trabalho, pois obtiveram a contagem de frequências mais constantes por ano.

Conclui-se também que o uso do BERT ao invés do clássico word2vec ajuda a aumentar, nos casos analisados, o desempenho estatístico embora com um custo maior de tempo e memória. As duas técnicas aplicadas a diferentes conjuntos de dados e configurações do modelo apresentam não apenas mudanças em relação as métricas escolhidas, mas também diferenças substanciais relativamente ao tempo e ao quantitativo de memória necessários para o treinamento e predição dos modelos. Entretanto no caso de mineração de resumos não resultou em ganho significativo.

Acknowledgements. This work was supported by the Brazilian funding agencies CNPq and by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

Authorship statement. Os autores confirmam que são as únicas pessoas responsáveis pela autoria deste trabalho, e que todo o material aqui incluído como parte deste artigo é de propriedade (e autoria) dos autores ou tem a permissão dos proprietários a serem incluídos aqui.

References

[1] S. Bagheri, N. M. Julkapli, W. A. Yehye. Catalytic conversion of biodiesel derived raw glycerol to value added products. Renewable and Sustainable Energy Reviews, vol. 41, p. 113-127, 2015.
[2] A. R.-O. I. Beltram, J. J. D. Jaen, T. Montini, P. Fornasiero. Photocatalytic valorization of ethanol and glycerol over TiO2 polymorphs for sustainable hydrogen production. Applied Catalysis A: General, vol. 518, p. 167-175, 2016.
[3] V. K. S. U. Garlapati, A. Budhiraja. Bioconversion technologies of crude glycerol to value added industrial products. Biotechnology Reports, vol. 9, p. 9-14, 2016.
[4] E. W.-G. A. Skrzyńska, M. Capron, F. Dumeignil. Crude glycerol as a raw material for the liquid phase oxidation reaction. Applied Catalysis A: General, vol. 482, p. 245-257, 2014.
[5] A. Thomas, Natural Language Processing with Spark NLP,O'Reilly, 2020.
[6] Kośmider, A., Białas, W., Kubiak, P., Drożdżyńska, A., & Czaczyk, K. (2012). Vitamin B12 production from crude glycerol by Propionibacterium freudenreichii ssp. shermanii: Optimization of medium composition through statistical experimental designs. Bioresour Technol, 105, 128-133. https://doi.org/10.1016/j.biortech.2011.11.074
[7] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
[8] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
[9] Joachims, T., Learning To Classify Text Using Support Vector Machines, December 2001,DOI:10.1007/978-1-4615-0907-3