Josué Batista Matos Deschamps de Melo
Leandro Guimarães Miranda
Marcos Vinicius Santos Souza
Vagner Ferreira Santos

Projeto Desafio - Análise de descrição de vagas de emprego

São Paulo

2023

1 Introdução

O problema é apresentado a partir da dificuldade de empregabilidade da cidade de Los Angeles, onde haverá muitas aposentadorias em um curto período.

Para isso visa compreender quais impactos (positivos e negativos) um informativo de vaga de trabalho pode causar para os interesses de candidatos.

Logo, é necessário encorajar candidatos com melhores descrições dos requisitos e clareza nas descrições de atividades a serem desenvolvidas. Outros fatores importantes como a descrição salarial podem influenciar.

Portanto, para solucionar esta demanda, serão realizadas: coleta, organização, manipulação e análise dos dados para realizar a interpretação e predição das informações.

2 Solução Proposta

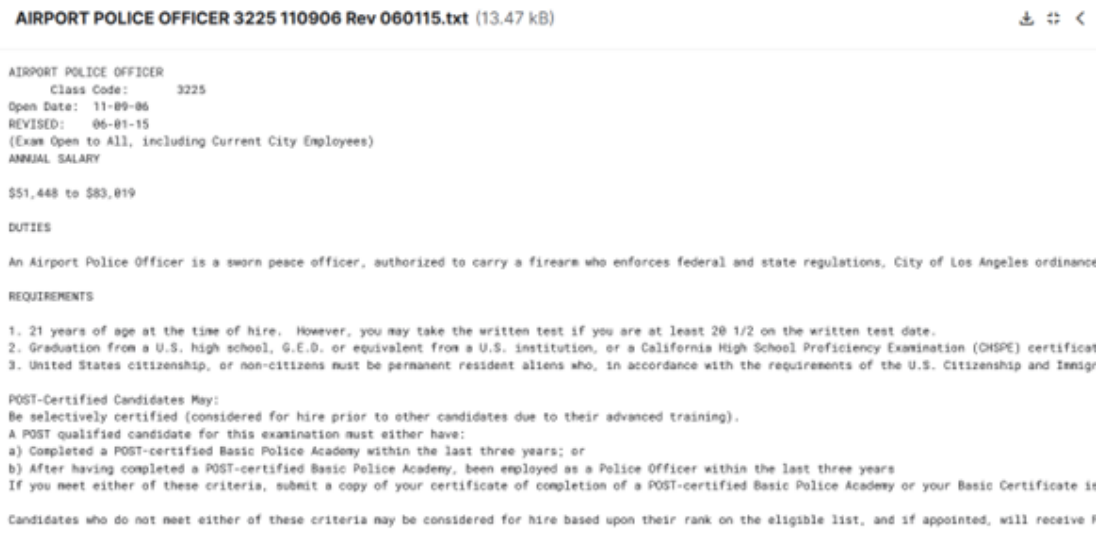
- Análise de Diversidade (masculino, feminimo, etc)
- Salários por cargos e por requerimentos.
- Categorização de vagas similares (ex: vagas de TI, etc)
- Modelo de Machine Learning para viés de gênero das vagas

3 Análise e extração de insights

3.1 Coleta de dados

Através do site Kaggle foram coletados os dados do desafio “Data Science for Good: City of Los Angeles”, onde possui uma proposta similar a deste projeto. Estes dados se tratam de boletins de vagas de emprego que totalizam 660 arquivos.

Os dados são armazenados em arquivos de texto (.txt) e cada arquivo descreve um tipo de vaga de emprego diferente. A descrição de cada vaga pode variar em conteúdo e em estrutura.



AIRPORT POLICE OFFICER 3225 110906 Rev 060115.txt (13.47 kB)

```
AIRPORT POLICE OFFICER
      Class Code:      3225
Open Date:  11-09-06
REVISED:    06-01-15
(Exam Open to All, including Current City Employees)
ANNUAL SALARY

$51,448 to $83,019

DUTIES

An Airport Police Officer is a sworn peace officer, authorized to carry a firearm who enforces federal and state regulations, City of Los Angeles ordinance

REQUIREMENTS

1. 21 years of age at the time of hire.  However, you may take the written test if you are at least 20 1/2 on the written test date.
2. Graduation from a U.S. high school, G.E.D. or equivalent from a U.S. institution, or a California High School Proficiency Examination (CHSPE) certificat
3. United States citizenship, or non-citizens must be permanent resident aliens who, in accordance with the requirements of the U.S. Citizenship and Immigr

POST-Certified Candidates May:
Be selectively certified (considered for hire prior to other candidates due to their advanced training).
A POST qualified candidate for this examination must either have:
a) Completed a POST-certified Basic Police Academy within the last three years; or
b) After having completed a POST-certified Basic Police Academy, been employed as a Police Officer within the last three years
If you meet either of these criteria, submit a copy of your certificate of completion of a POST-certified Basic Police Academy or your Basic Certificate is

Candidates who do not meet either of these criteria may be considered for hire based upon their rank on the eligible list, and if appointed, will receive F
```

Exemplo de um arquivo de texto para vaga de Policial de Aeroporto

A manipulação e a análise, mesmo com as variações encontradas em cada arquivo de texto, podem ser resolvidas através de técnicas de Processamento de Linguagem Natural.

3.2 Organização de dados

Neste processo também, foi usado como referência o notebook do usuário Shahules chamado *“Discovering opportunities at LA”* também encontrado no site *Kaggle.*

Um bom tratamento de dados é essencial para realizar a extração de informação embasado nos dados coletados. Como as descrições de vagas de emprego são campos

abertos de digitação do usuário, foi necessário um trabalho massivo de tratamento para transformar os dados em informações de valor.

File Name	Position	salary_start	salary_end	opendate	requirements	duties	deadline	selection	EXPERIENCE_LENGTH
WATER SERVICE REPRESENTATIVE 1693 111717.txt	water service representative	70,177	$87,194	2017-11-17	Three years of full-time paid experience in a ...	A Water Service Representative makes field inv...	NOVEMBER 30, 2017	['Test']	Three
HARBOR PLANNING AND ECONOMIC ANALYST 9224 1118...	harbor planning and economic analyst	70,908	$103,648	2016-11-18	1. Graduation from an accredited four-year col...	A Harbor Planning and Economic Analyst perform...	DECEMBER 8, 2016	['Questionnaire']	Two
MANAGING WATER UTILITY ENGINEER 9406 032417 RE...	managing water utility engineer	148,561	$184,579	2017-03-24	1. Two years of full-time paid experience at t...	A Managing Water Utility Engineer may serve as...	APRIL 13, 2017	['Essay', 'Interview']	Two

Exemplo de tabela criada a partir dos arquivos obtidos

Com base nas descrições de vagas foi realizada a “limpeza” dos dados, a tokenização dos textos, para fossem removidos sinais de pontuação e *stopwords* (itens dos quais não contribuem para a análise) e *Stem* para exibir palavras mais frequentes.

Utilizamos bibliotecas do python (regex, nltk, pandas e unicodedata) para realizar padronizações e conversões de *datatypes*, sendo consideradas como as principais deste trabalho. Cada biblioteca é caracterizada por:

- Pandas: usada para extrair e mostrar as informações em formato de tabelas;
- Matplot: usada para visualizar as informações em gráficos com o objetivo de realizar análises estatísticas;
- NLTK: Sigla para Natural Language Toolkit, usada para auxiliar na classificação e tokenização dos insights;
- Unicodedata: usada para auxiliar na padronização de codificação de caracteres usados em diferentes idiomas;
- Regex: usada para encontrar trechos específicos e padronizar a coleta destas informações.

3.3 Análise de dados

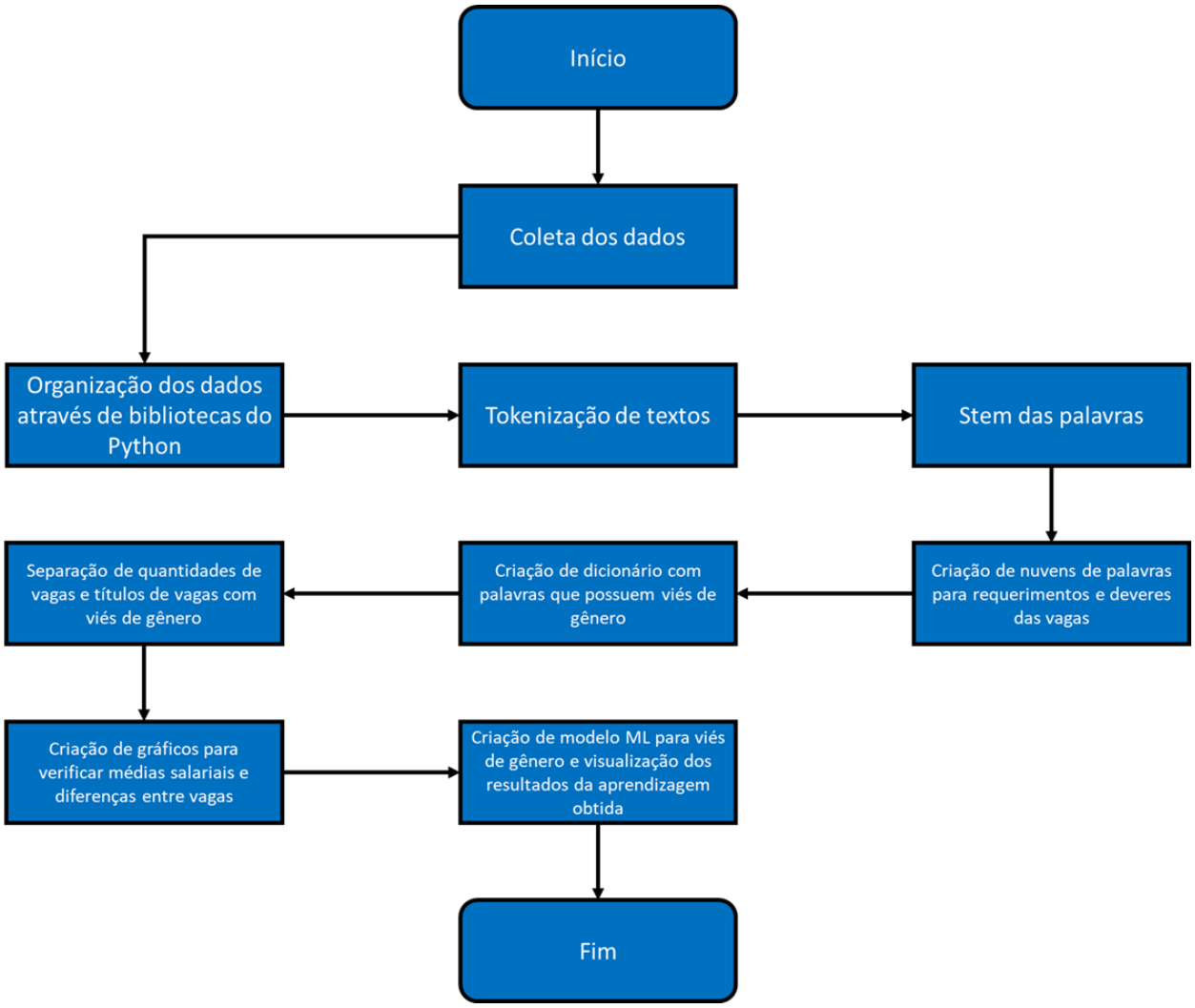
Após ser realizada toda a organização dos dados, foram gerados gráficos para revelar pontos positivos e negativos de cada descrição de vaga.

Com isso, foi possível identificar que existe um viés de gênero nas palavras usadas nos anúncios de emprego. Com base em um estudo sobre o viés de palavras, são

apresentadas palavras que são geralmente carregadas de viés inconsciente de gênero. Essas palavras podem ser usadas para identificar a presença de viés de gênero nos anúncios de emprego e ajudar a garantir que os anúncios sejam mais inclusivos.

A partir disso foi possível realizar um modelo com técnicas de Machine Learning para análise de viés de gênero.

4. Fluxograma

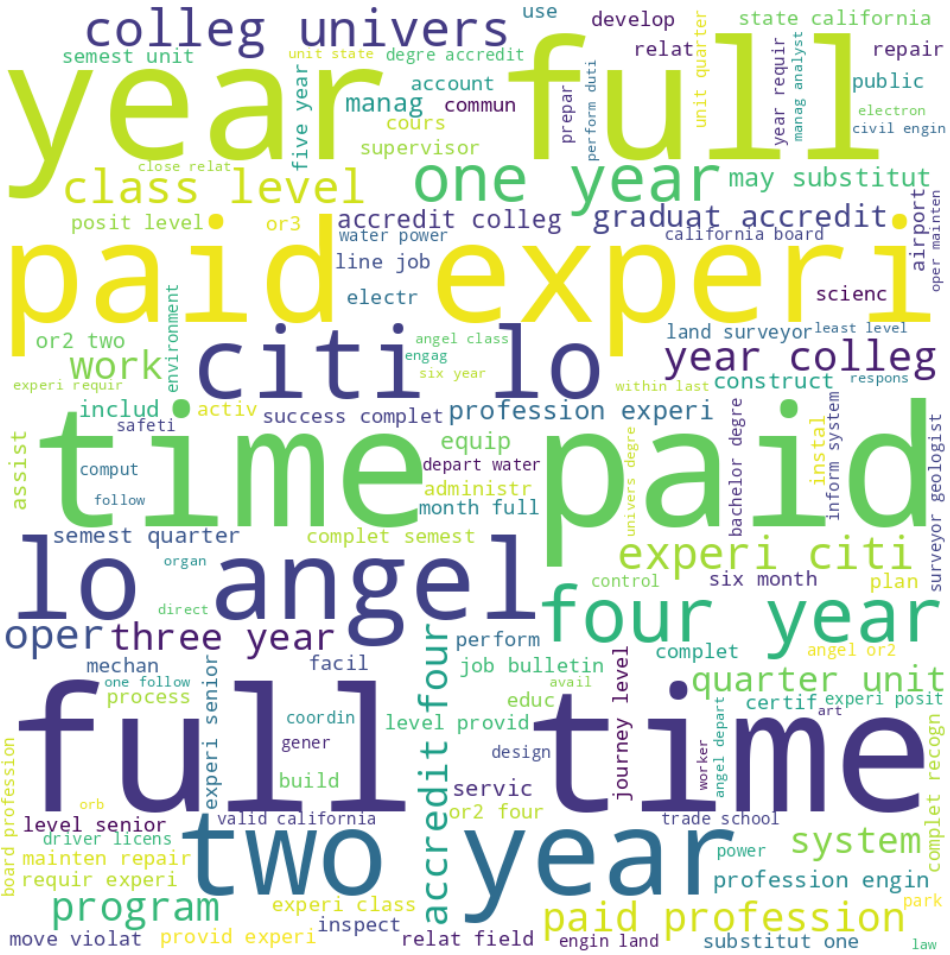


5. Resultados

5.1 Nuvem de palavras

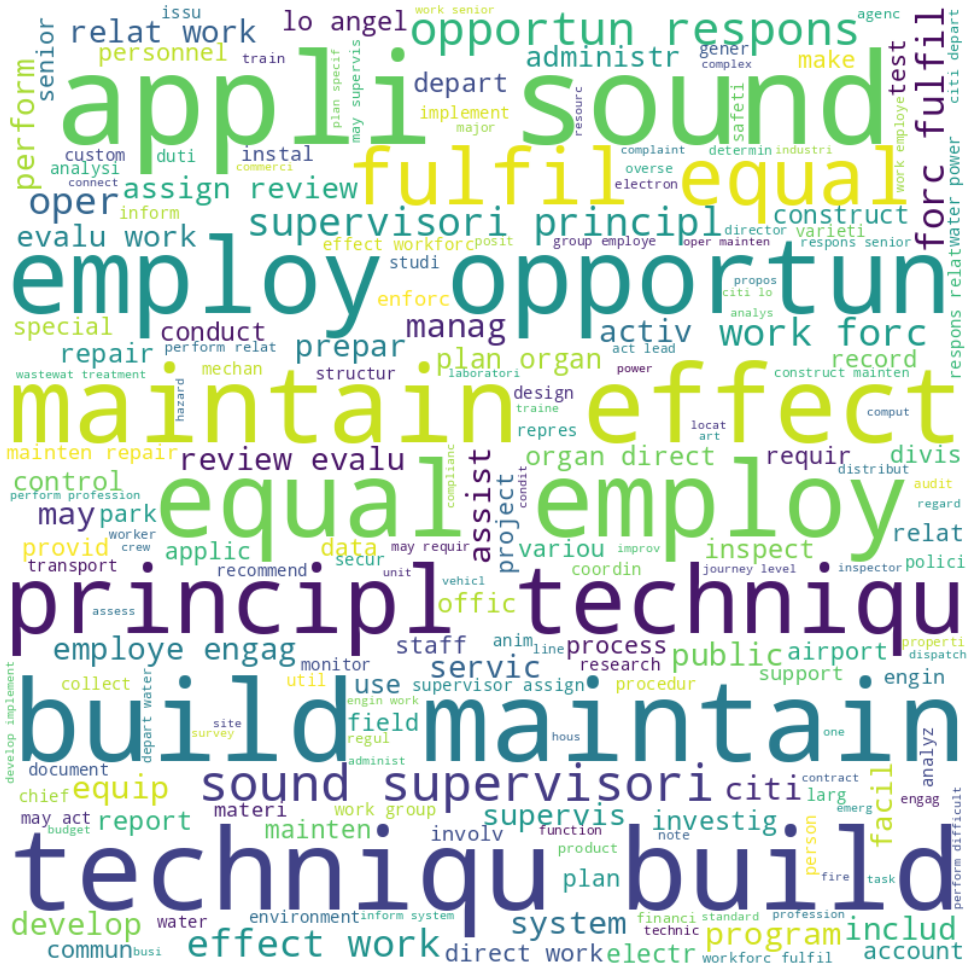
Uma nuvem de palavras pode mostrar de forma simples a quantidade de vezes em que uma palavra pode aparecer nas análises.

O campo “requerimento”, onde o contratante descreve suas exigências para aprovação da vaga, podemos encontrar palavras em comum na maioria das vagas, como por exemplo “college university”, remetendo ao nível de escolaridade.



Exemplo mostrando palavras com maior ocorrência no campo “requerimento” de cada vaga

O campo “deveres”, onde o contratante descreve as atribuições que o funcionário irá desempenhar, podemos encontrar palavras em comum na maioria das vagas, como por exemplo "assist", remetendo ao conceito de “dar assistência”.



Exemplo mostrando palavras com maior ocorrência no campo “deveres” de cada vaga

5.2 Número de vagas com viés de gênero

Após obter os requerimentos foi possível obter palavras com viés voltados para homens ou para mulheres.

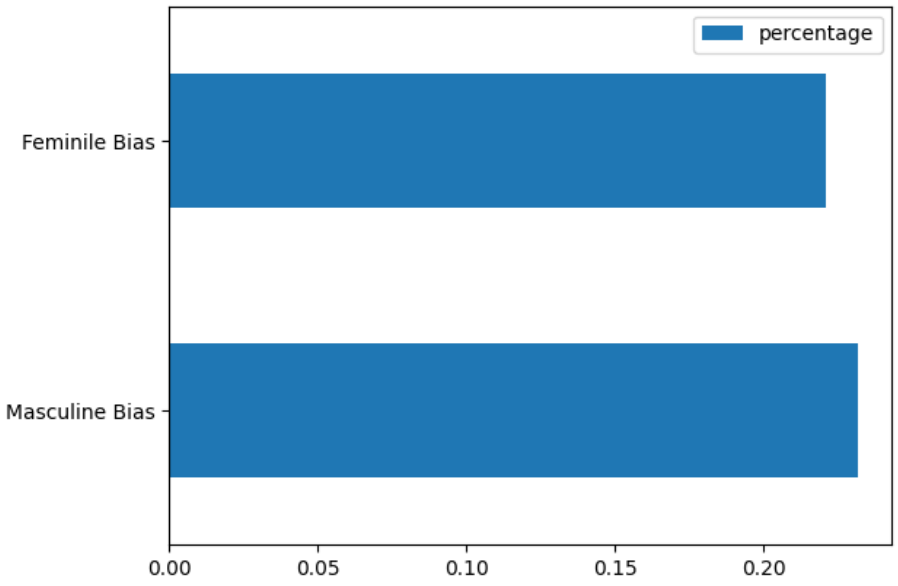
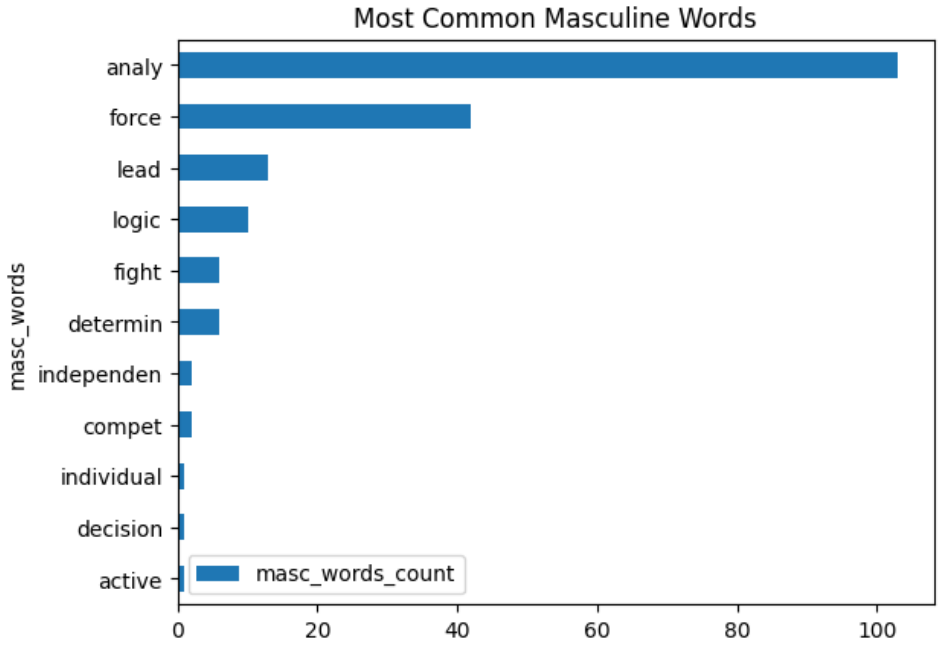
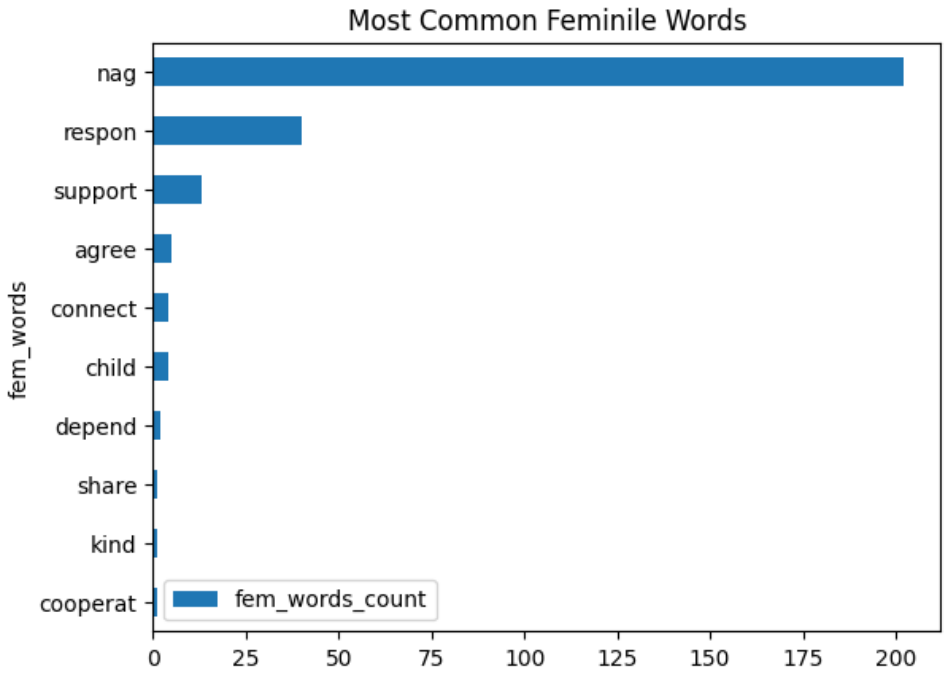


Gráfico revelando a porcentagem de viés em relação ao gênero

Também foram separadas quais destas palavras e a frequência em que apareciam nos textos. Por conta da “lematização”, técnica que faz redução de uma palavra à sua forma base para criar termos mais específicos, algumas palavras são retornadas de forma “neutra” sem que haja um direção para o gênero implícito na palavra em questão:



Quantidade de palavras mais comuns para gênero masculino



Quantidade de palavras mais comuns para gênero feminino

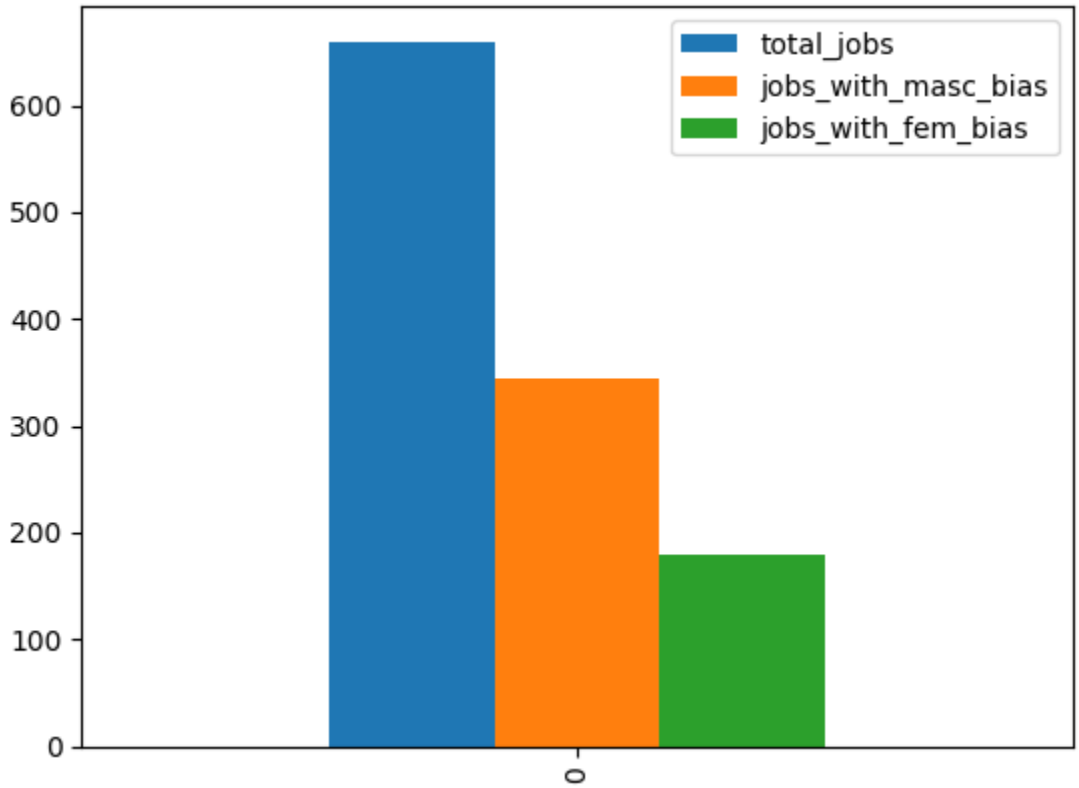
5.3 Título de vagas com viés de gênero

Outro tipo análise para determinar capturar vieses masculinos e femininos, mostrando quais vagas possuem esta característica:

Palavras que possuem um maior direcionamento ao gênero masculino

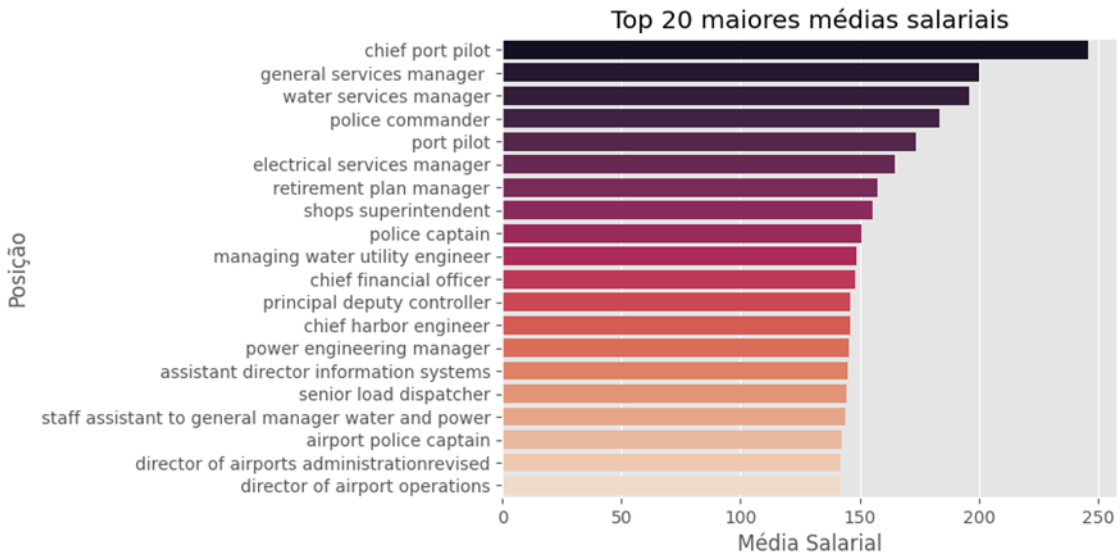
Palavras que possuem um maior direcionamento ao gênero feminino

Para cada gênero é possível verificar a quantidade de viés em relação ao total de vagas, onde pode-se verificar que há um viés maior para o gênero masculino:



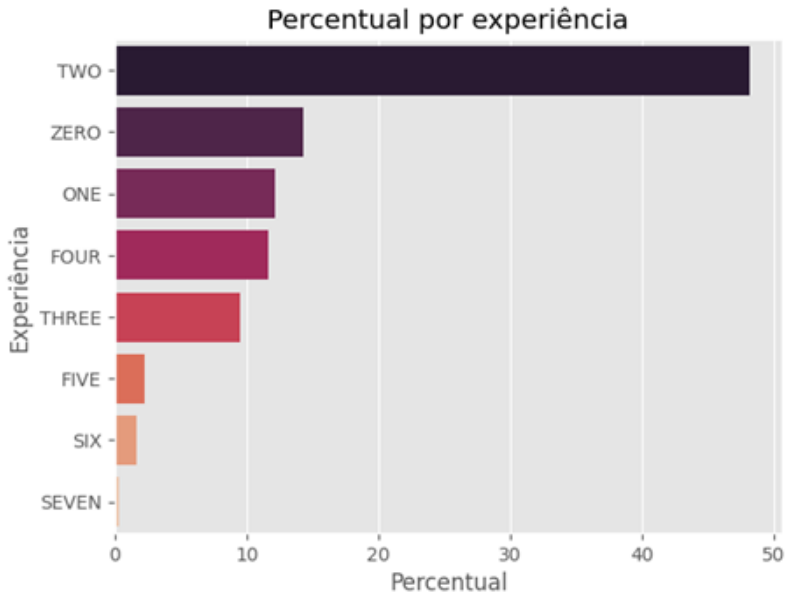
5.4 Indicadores – Requerimentos

O requerimento é uma das partes mais importantes que compõem a descrição da vaga de emprego, pois lista todos os requisitos básicos que o candidato deve atender para se candidatar ao emprego, em teoria as vagas que solicitam mais requisitos deveriam proporcionar uma melhor oferta salarial. Podemos observar por exemplo as vagas com as melhores ofertas salariais:



20 maiores vagas mostrando posição (cargo) e suas respectivas médias salariais

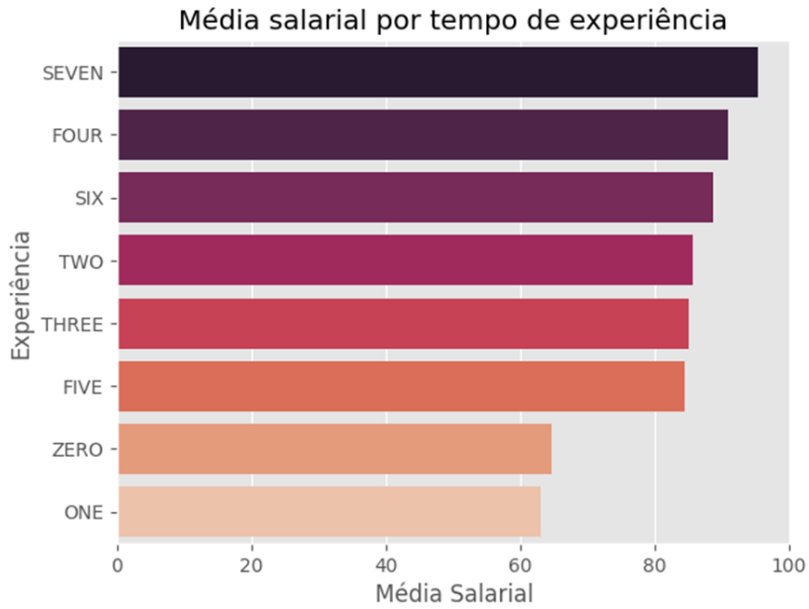
Ao observar o requisito tempo de experiência, percebemos que 70% das vagas solicitam pelo menos 2 anos de experiência na área, sendo que 22% exigem de 3 a 7 anos de experiência:



20 maiores vagas mostrando posição (cargo) e suas respectivas médias salariais

Quando analisamos a média salarial por tempo de experiência requerido, percebemos que não há uma discrepância muito grande de oferta de salário, principalmente ao analisar entre 2 e 6 anos de experiência. Isso levanta a questão se é realmente necessário cobrar tanto tempo de experiência em determinadas ofertas de

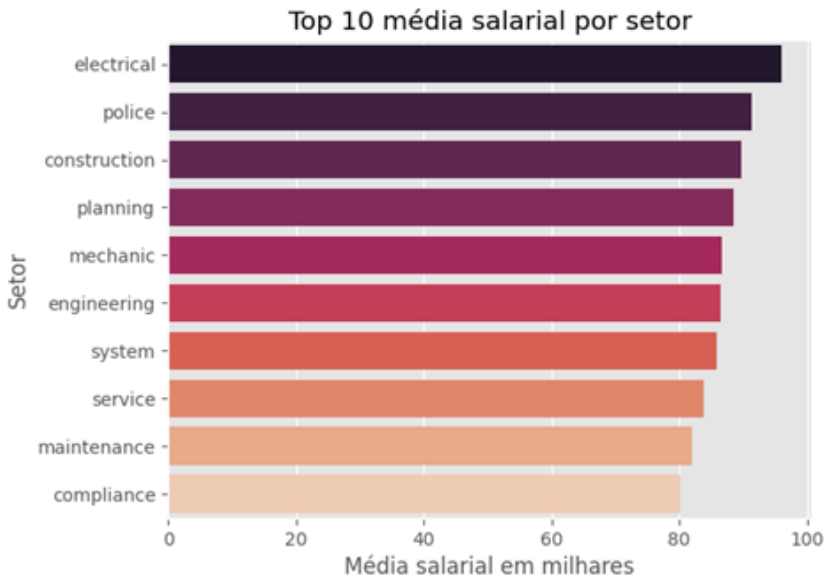
trabalho. Diminuir o requisito de experiência profissional, pode aumentar a quantidade de pessoas que se candidatam.

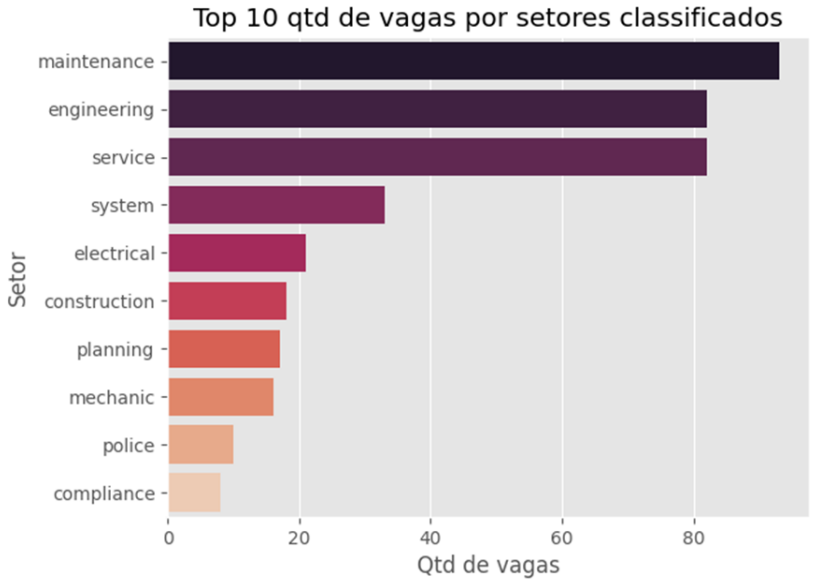


Média salarial em relação ao tempo de experiência exigido em cada vaga

5.6 Indicadores – Setor

Através de *tokenização* e *lematização* conseguimos obter palavras chaves nos títulos das vagas para realizar uma classificação por setores. Utilizamos a biblioteca unicodedata para realizar a normalização da codificação dos textos e a NLTK para realizar o processamento e extração das palavras chaves, contando a repetição dessas palavras, foi necessário realizar a exclusão de *stop words* e de palavras que repetiam mas não eram setores como *analyst* e *senior*, com isso podemos extrair indicadores de quantidade de vagas e média salarial por setor.



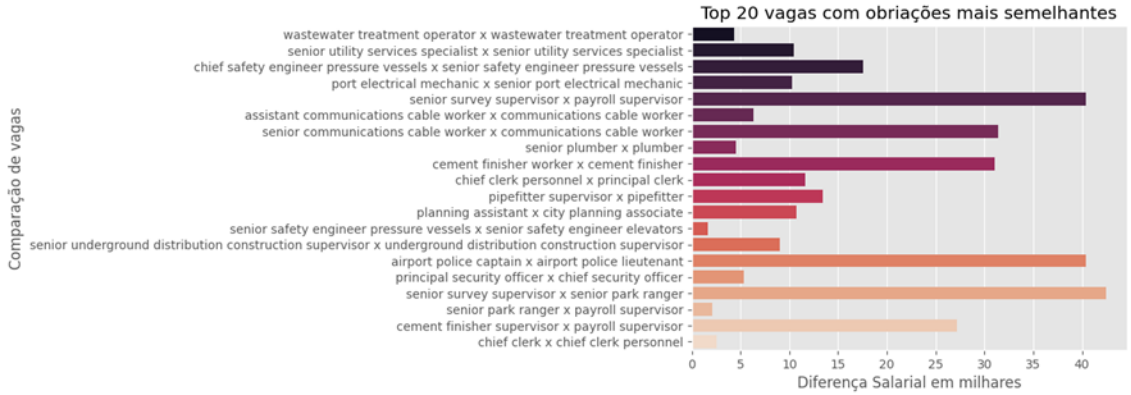


5.7 Recomendação de vagas – *Jaccard*

Utilizando o algoritmo *Jaccard* é possível calcular a similaridade entre as vagas, com isso realizar novas recomendações de vagas abertas para os candidatos. O coeficiente de similaridade de *Jaccard* é utilizado para mensurar a similaridade entre duas amostras, o retorno é um número entre 0 e 1, quanto mais próximo de 0, mais as amostras são similares. Utilizamos o algoritmo *Jaccard* implementado na biblioteca NLTK do *python*. Foram confrontadas todas as obrigações das vagas entre si para extrair quais

são as mais parecidas, o resultado foi de 148 vagas com obrigações similares de um total de 660.

Comparamos a diferença salarial entre as vagas com obrigações parecidas:



5.8 Modelo de classificação de bias (viés)

Após toda a análise de viés e exigências em cada vaga, foi possível criar um modelo.

Este modelo caracteriza o viés da vaga retornando um número:

- Se este número tende a 0, significa que a vaga possui maior viés para o gênero feminino;
- Se este número tende a 1, significa que a vaga possui maior viés para o gênero masculino.

```
0.7827056->1. Three years of full-time paid experience at the level of Printing Services Supervisor with the City of Los Angeles managing a diversified printing o
0.5142929->Two years of full-time paid experience in a class at the level of Senior Administrative Clerk performing all of the following secretarial work: routing
0.09212834->Four years of full-time paid experience in work involving all of the following areas: fabrication, welding, altering, and assembling structural steel
0.21019861->1. Current employment with the City of Los Angeles; and2. Four years of full-time paid experience at the level of Airport Manager with the City of Los
0.3067513->1. Two years of full-time paid experience with the City of Los Angeles as a Management Assistant or Management Aide interpreting and applying State and
0.86507726->Four years of full-time paid experience with the City of Los Angeles as a Steam Plant Maintenance Mechanic or in a class at that level performing main
0.3067513->1. Two years of full-time paid experience with the City of Los Angeles at the level of Housing Investigator, investigating complaints, analyzing eviden
0.44871727->1. One year of full-time paid experience as a Principal Security Officer with the City of Los Angeles; or2. Two years of full-time paid experience as
0.970265->Four years of full-time paid housing rehabilitation or production experience, at the level of Rehabilitation Project Coordinator, supervising employees
```

Exemplo de cada descrição de vagas com seus respectivos valores

Para o treino do modelo foram utilizadas 1000 épocas e uma função de ativação *sigmoid*. Foi obtida uma acurácia de 81%.

```
Epoch 996/1000
16/16 [==============================] - 0s 16ms/step - loss: 0.3121 - accuracy: 0.8120
Epoch 997/1000
16/16 [==============================] - 0s 14ms/step - loss: 0.3127 - accuracy: 0.8080
Epoch 998/1000
16/16 [==============================] - 0s 13ms/step - loss: 0.3123 - accuracy: 0.8080
Epoch 999/1000
16/16 [==============================] - 0s 14ms/step - loss: 0.3119 - accuracy: 0.8100
Epoch 1000/1000
16/16 [==============================] - 0s 14ms/step - loss: 0.3133 - accuracy: 0.8120
```

Captura de tela com as 5 últimas épocas e respectivas acurácias

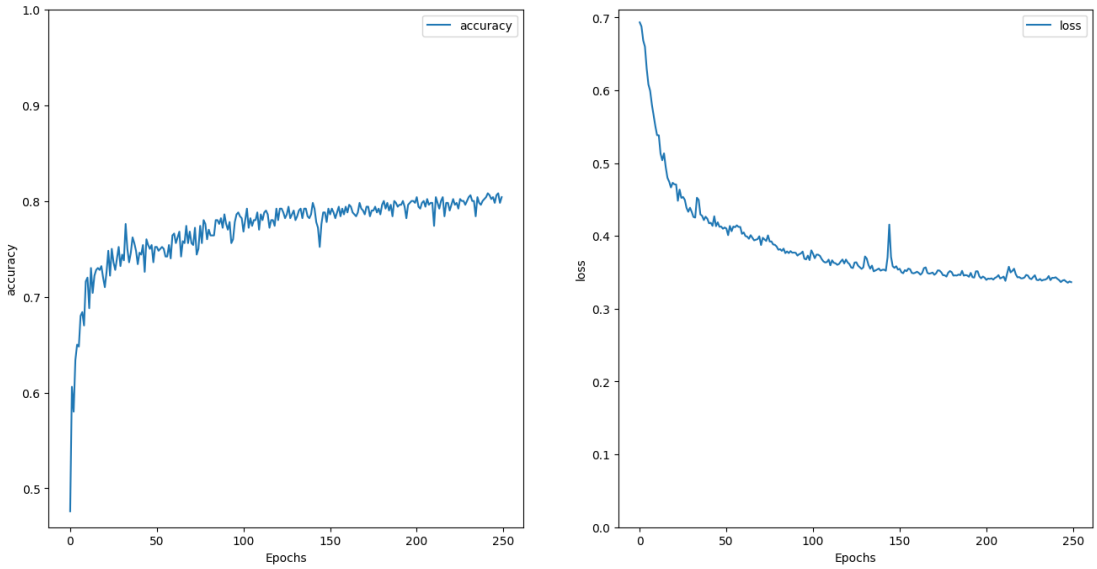


Gráfico mostrando comportamento da acurácia e perda a cada época

6. Conclusão

O projeto como um todo foi difícil, visto que exige a utilização de diversas técnicas de processamento de linguagem natural.

Dentre as principais dificuldades estão:

- Extração dos dados obtidos através da base de dados, pois em sua maioria não possuem uma padronização;
- Quantidade de dados, pois mesmo possuindo mais de 600 arquivos, cada um possui características diferentes e faz com que sejam informações "únicas", e não repetidas, o que em uma rede neural pode ser problemático;
- Processos de tokezinação, pois por se tratar remover informações do conteúdo para simplificar a análise, exige tempo para ser desenvolvido;

Mesmo diante das dificuldades houve um excelente resultado, pois o objetivo de coletar, extrair, organizar e aplicar conceitos de processamento de linguagem natural foi alcançado e, com isso, poderá ser aplicado para vagas futuras, a fim de minimizar quaisquer vieses de gênero em descrições de vagas de emprego.

O código-fonte deste trabalho encontra-se disponível no GitHub através do repositório: github.com/josueMelo/senac-rna-analise-de-vaga

7. Referências

1. Odbal, Guanhong Zhang, Sophia Ananiadou: Examining and mitigating gender bias in text emotion detection task. 2022, Volume 493, Páginas 422-434.
2. Aaron C, Kay, Danielle Gaucher and Justin, Friesen. Evidence That Gendered Wording in Job Advertisements Exists and Sustains Gender Inequality, Journal of Personality and Social Psychology. 2011, American Psychological Association Vol. 101, Páginas 109 –128.
3. Sridevi G. M. , S. Kamala Suganthi, AI based suitability measurement and prediction between job description and job seeker profiles, International Journal of Information Management Data Insights, Volume 2, Issue 2, 2022
4. Hieu Trung, Tran, Hanh Hong Phuc, Vo, Son T., Luu. Predicting Job Titles from Job Descriptions with Multi-label Text Classification. 2022
5. Al-Saiyd, Nedhal & Al-Takrouri, Amjad. (2015). PREDICTION OF IT JOBS USING NEURAL NETWORK TECHNIQUE. Ubiquitous Computing and Communication Journal. 9. 1.
6. Huynh, Tin & Nguyen, Kiet & Nguyen, Ngan & Nguyen, Anh. (2019). Job Prediction: From Deep Neural Network Models to Applications.
7. Luo, B., Feng, Y., Wang, Z., Huang, S., Yan, R., & Zhao, D. (2018). Marrying up Regular Expressions with Neural Networks: A Case Study for Spoken Language Understanding.