Recuperação Inteligente de Informação – Relatório final do projeto

Edinadja Mayara de Macedo

Agrupamento da base de projetos de extensão do Instituto Federal do Rio Grande do Norte (IFRN)

1. Introdução (breve resumo do trabalho - o que você fez e como)

Inicialmente, busquei fazer um reconhecimento do dataset, de forma a conhecer os atributos e seus respectivos formatos. Em seguida, selecionei o atributo resumo para construir o nosso corpus de documentos, uma vez que este cumpria os requisitos desejados para tal.

Feito isso, passei à etapa de pré processamento, onde foram definidas funções para aplicar algumas das técnicas de pré-processamento vistas anteriormente em sala: lowercase e remoção de stopwords, optei por não aplicar o stemming, pois a visualização dos termos após a clusterização ficou estranha.

Estas funções previamente definidas foram chamadas dentro da função principal: `def processCorpus(corpus, language):`, onde também foram chamadas algumas regex para remover quebras de linha, trechos com códigos, termos irregulares ou inesperados etc. Na função principal também foi feita a normalização do texto, removendo letras com acento e ç.

Em seguida, foi feita a ponderação estatística das palavras no corpus, isto é, foi aplicada a função TF-IDF, uma estatística numérica que pretende refletir a importância de uma palavra para um documento em um corpus, dando a cada palavra em um documento uma pontuação que varia de 0 a 1.

O passo seguinte foi realizar a clusterização dos documentos pré-processados. Para isso, foi utilizado o algoritmo K-means, um algoritmo de aprendizado não supervisionado que avalia e clusteriza os dados de acordo com suas características. Sendo assim, foi criada uma função para aplicar o k-means no nosso corpus de documentos já pré-processado.

No momento em que a função que executa o algoritmo de clusterização é acionada, também é chamada

Para verificar e selecionar o melhor valor para k, isto é, quantos pontos aleatórios do *dataset* são escolhidos para serem as coordenadas dos centroides iniciais, foi utilizada a função *silhouette score,* uma medida de quão semelhante um objeto é ao seu próprio cluster (coesão) em comparação com outros clusters (separação). A *silhouette score* mede o quão bem um ponto se encaixa em um cluster.

Com isso, notou-se que o "melhor" valor para k está entre 2 e 3, embora os valores não tenham mostrado números muito bons sobre nenhum dos cenários apresentados.

Por fim, os clusters gerados são exibidos em forma de nuvem de palavras para melhor visualização e cada documento foi rotulado de acordo com o cluster ao qual pertence.

2. Descrição do sistema - overview

2.1 Breve descrição do sistema, incluindo seus objetivos

O sistema trata-se do resultado do processo de clusterização utilizando o algoritmo K-means na base de dados de projetos de extensão do Instituto Federal do Rio Grande do Norte (IFRN).

O objetivo do sistema foi tentar identificar padrões nos projetos de extensão, e verificar a relação existente entre estes padrões e a área de conhecimento à qual os projetos de extensão pertencem.

2.2 Base de documentos/itens utilizada no projeto. A base pode ter sido montada por você ou pode ter sido coletada de algum site.

- **Aquisição da base - como foi realizada a aquisição do corpus? De onde vieram esses documentos (da Web?). O sistema fez aquisição automática ou os documentos foram manualmente selecionados?**

O corpus foi feito com base no dataset acerca dos projetos de extensão do IFRN, disponíveis em https://dados.ifrn.edu.br/dataset/projetos-de-extensao, selecionados manualmente.

- **Descrição da base - Como são os documentos/itens utilizados pelo sistema? Dê um ou mais exemplos.**

A base contém a relação dos projetos de extensão do Instituto Federal do Rio Grande do Norte, e apresenta atributos referentes à:

- O resumo da proposta do projeto de extensão.
- A data final da execução.
- A área de conhecimento na qual o projeto está inserido.
- A justificativa.
- O coordenador do projeto.
- O identificador (id).
- A data de início da execução do projeto.
- O foco tecnológico.
- A equipe do projeto.
- O título do projeto.
- E o campus ao qual está vinculado.

Para a criação do corpus foi utilizado o atributo "resumo" do dataset, o qual se refere ao resumo de cada projeto de extensão presente na base.

O objetivo do sistema seria tentar prever a área de conhecimento dos projetos de extensão com base no resumo do mesmo.

3. Implementação do sistema/protótipo

3.1 Representação dos documentos/itens

- **Como são representados os documentos/itens dentro da base? Dê um exemplo, mostrando a representação interna do documento apresentado como exemplo na seção 2.2.**
- **Que etapas de preparação (pré-processamento) do documento textual foram realizadas? (stoplist, stemming, n-grams, etc...).**

Como mencionado anteriormente, foram aplicadas a remoção de stopwords, lowercase, tokenização, normalização/decodificação com substituição de letras com acentos por suas respectivas versões sem acento, além de algumas limpezas utilizando expressões regulares como: remover quebras de linha, trechos com códigos, termos irregulares ou inesperados etc.

3.2 Arquitetura do sistema:

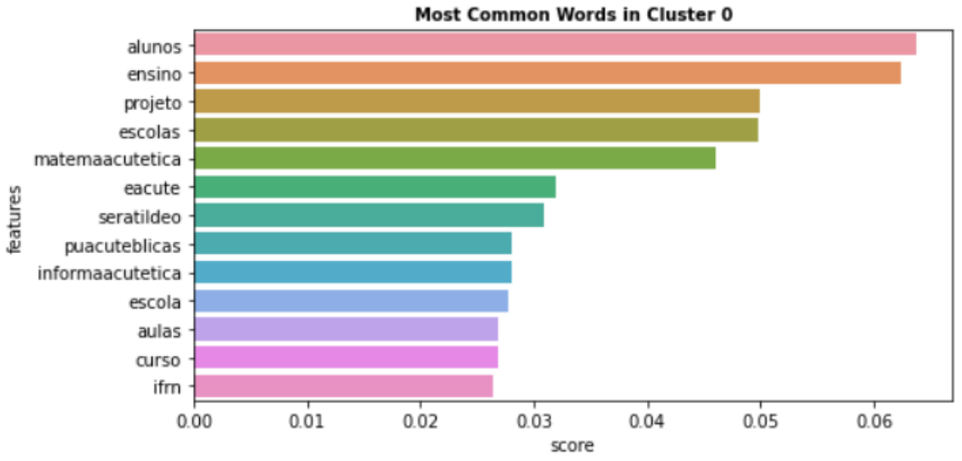
- **Desenhe a arquitetura do sistema e dê uma breve descrição de seus módulos (diga como se dá o fluxo de informações dentro do sistema).**
- **Que técnica foi utilizada para resolver o problema?**

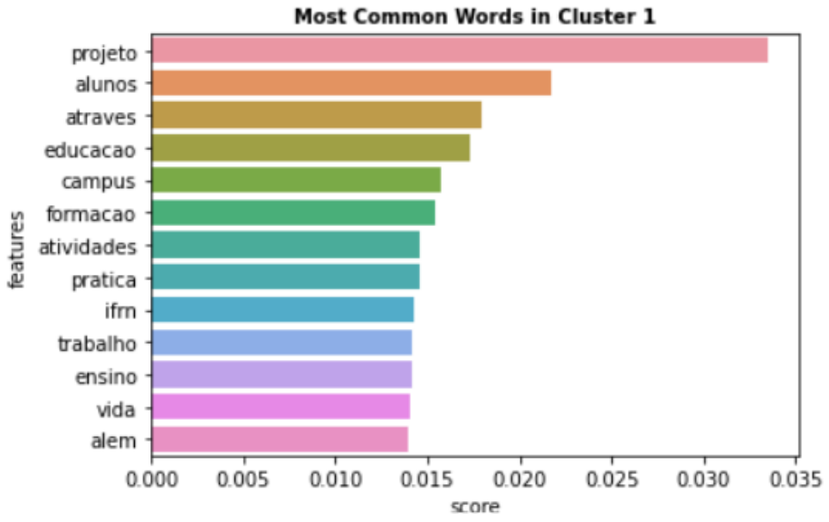
A técnica escolhida foi clusterização, utilizando o algoritmo K-means.

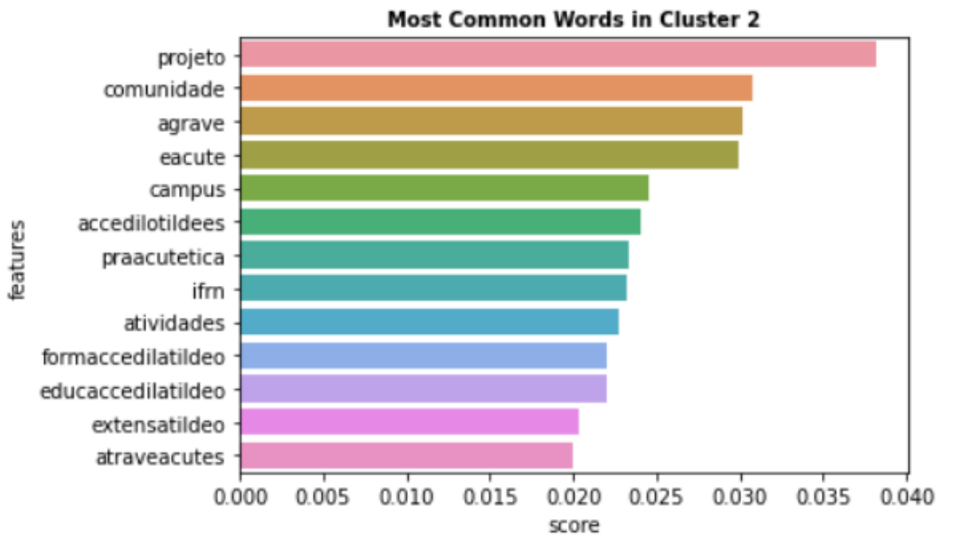
3.3 Exemplo de uso do sistema

- **Se for o caso, dê um exemplo do sistema em uso (input/output).**

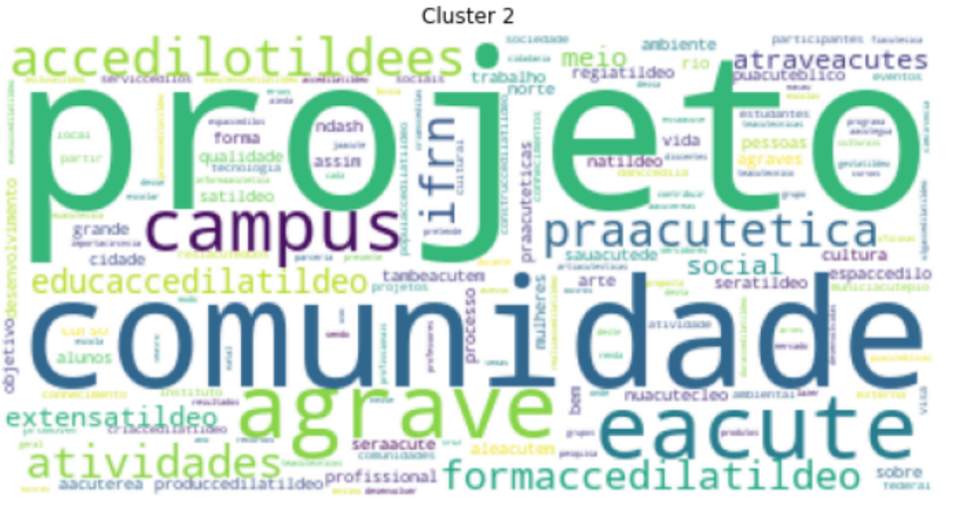
Como mencionado anteriormente, optou-se por utilizar 3 clusters, os quais apresentaram interseções entre si, com palavras que se repetem, por vezes, nos 3 clusters gerados, como pode ser observado nos gráficos abaixo.







Os inputs foram os resumos dos projetos de extensão, que foram processados pelo algoritmo k-means e o output apresentado foi a clusterização dos dados, que pode ser visualizada na forma de nuvens de palavras abaixo.



Análise dos clusters:

Apesar das homogeneidadades entre os clusters, é possível notar certas características únicas em cada um deles.

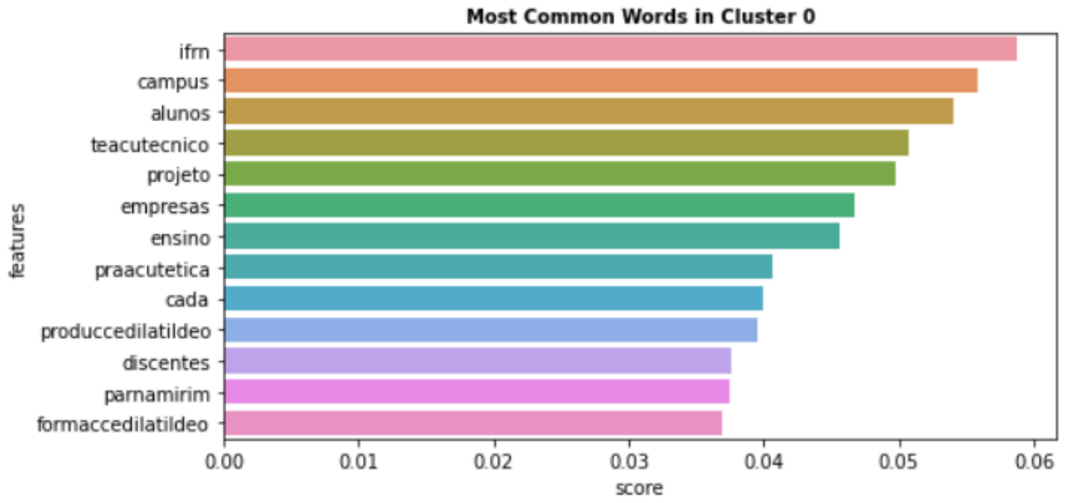
- No cluster 0 podemos observar as palavras matemática e informática (com alguns problemas de pré processamento), o que pode indicar que os projetos de extensão que contém estes termos pertencem à área de exatas.
- No cluster 1, temos os termos ensino, educação, formação, dentre outros, que indicam que os projetos que contém tais termos estão inseridos na área de educação, como pedagogia, por exemplo.
- No cluster 2, podemos ver os termos formação, comunidade, social, cultura e atividades, o que propõe que estes projetos são mais voltados à contribuir de alguma

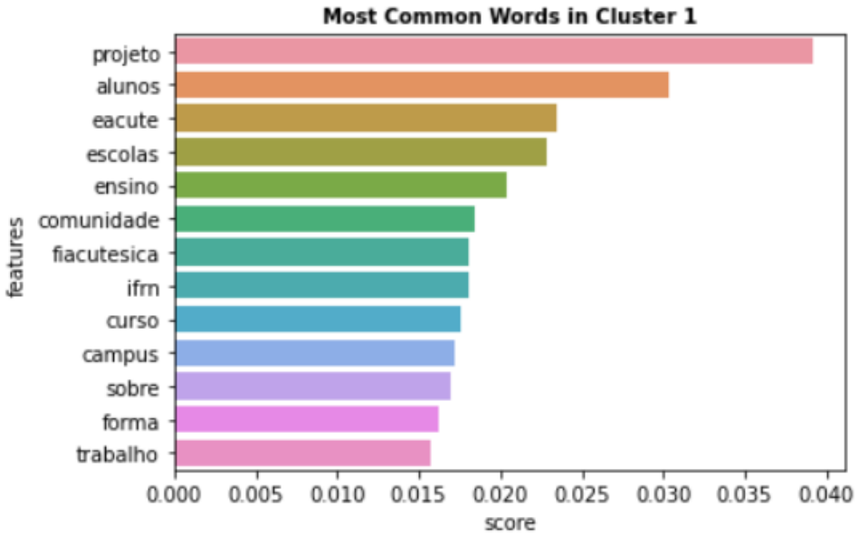
forma mais baseada nas ciências humanas, com a comunidade externa à instituição, com projetos voltados à atividades culturais e sociais.

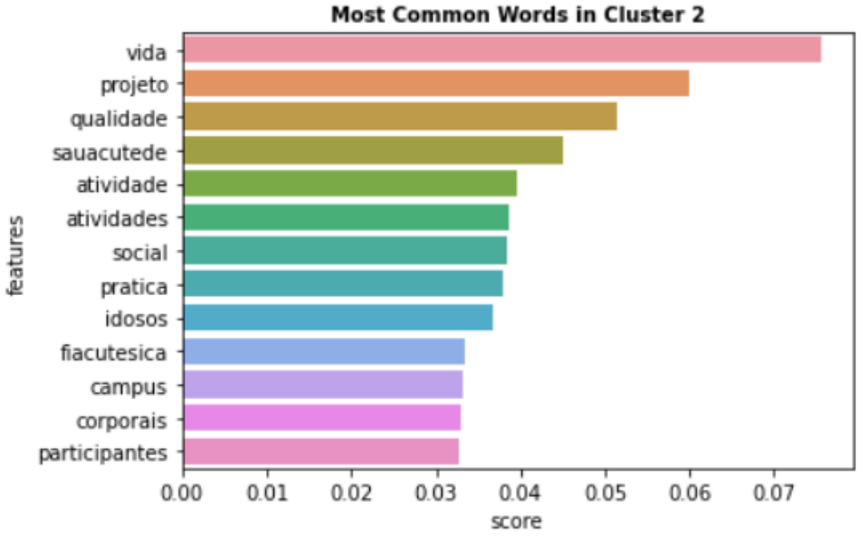
RESULTADO DOS TESTES DE CLUSTERIZAÇÃO COM ÁREAS ESPECÍFICAS PREVIAMENTE SELECIONADAS

Para tentar obter resultados mais claros com a clusterização, as áreas de conhecimento foram subdivididas em subáreas, das quais foram selecionadas 3: ciências humanas, da saúde e naturais, de forma a testar o sistema com um corpus mais reduzido, buscando melhorar os resultados da clusterização.

O corpus foi tratado/pré processado da mesma forma que o corpus integral e a clusterização retornou o seguinte resultado:

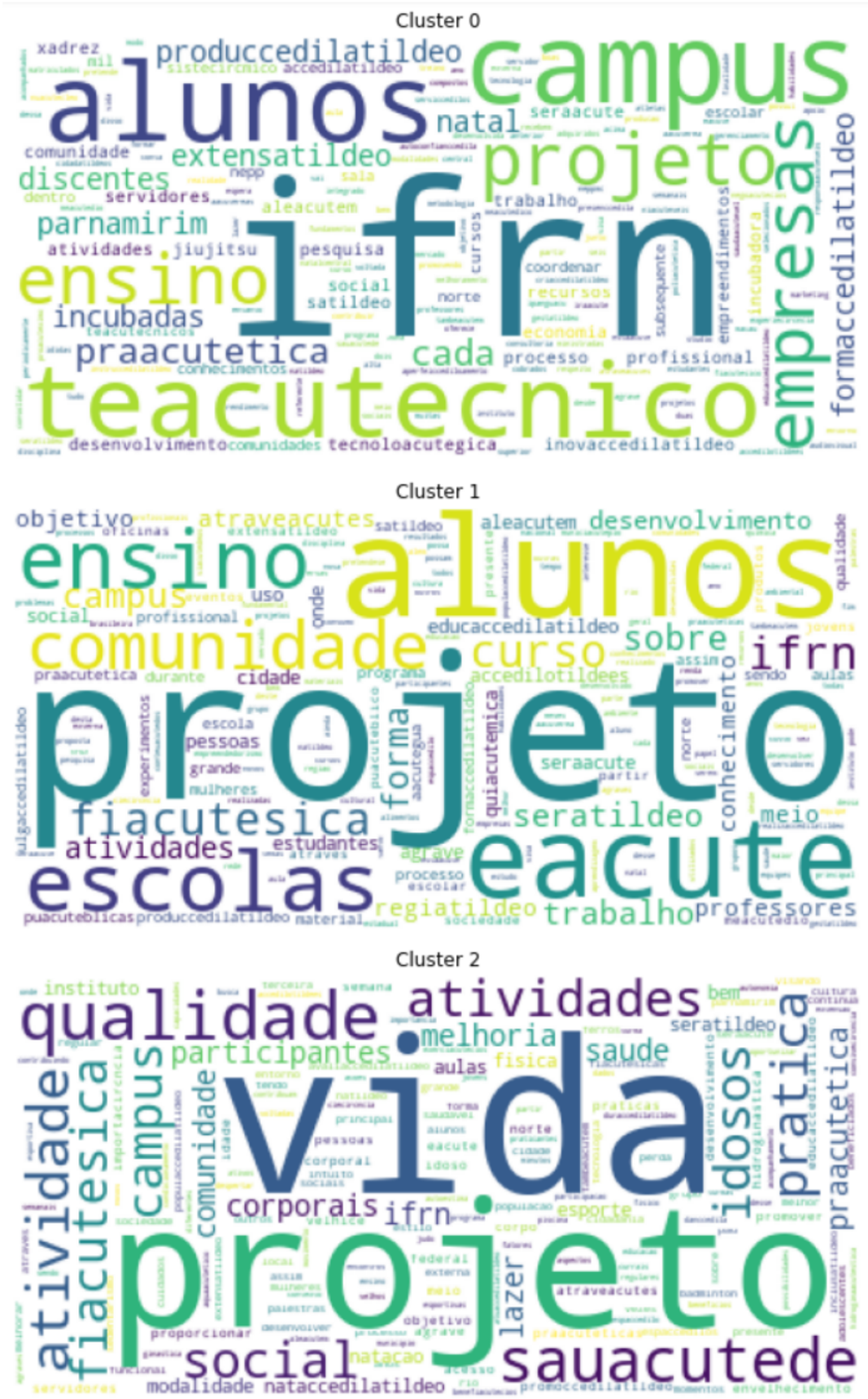






Como é possível observar, os clusters 0 e 1 ainda apresentam uma certa homogeneidade entre si, impossibilitando identificar qual a provável subárea a qual pertencem, no entanto, o cluster 2 apresenta termos que remetem muito à área da saúde, que era uma das três áreas de conhecimento específicas que haviam sido selecionadas para compor o corpus, indicando que os resumos que apresentam tais termos são, muito provavelmente, referentes a projetos de extensão dessa área.

As nuvens de palavras referentes a cada cluster podem ser observadas a seguir:



4. Avaliação do Sistema (Testes e Resultados)

4.1 Descreva os testes de desempenho do sistema:

- **Como foi criado/selecionado o corpus de testes? Quantos docs ele possui? São representativos?**

O corpus foi criado com os resumos dos projetos de extensão do IFRN. Possui cerca de 3220 documentos.

4.2 Descreva o resultado dos testes (precisão e cobertura, quando for o caso). O que pode ser feito para melhorar o desempenho do sistema?

5. Conclusão

- **Quais são os pontos fortes do seu sistema (elogios).**
- **Quais são os pontos fracos do sistema, como eles poderiam ser melhorados?**

O sistema possibilitou a visualização de padrões nos textos dos resumos dos projetos de extensão, permitindo ter uma ideia das temáticas tratadas pelos projetos e suas finalidades, por meio de uma análise dos resultados da clusterização, sem a necessidade de analisar um a um dos projetos da base de dados.

A maioria das tarefas solicitadas foram executadas, no entanto os resultados obtidos não foram tão satisfatórios, uma vez que os clusters apresentaram uma certa homogeneidade indesejada. Isto só foi percebido meio tarde para refazer o trabalho.

Talvez isto pudesse ter sido evitado se a base tivesse sido construída, com a utilização de crowlers e algoritmos de mineração de dados, produzindo um resultado final mais satisfatório.

Referências

ALVES, Gisele. Aprendizado não supervisionado com K-means. 2018. Disponível em: https://medium.com/neuronio-br/aprendizado-n%C3%A3o-supervisionado-com-k-means-f4272dee98a0. Acesso em: 26 dez. 2022.

ANASTACIO, Bruno. K-means: o que é, como funciona, aplicações e exemplo em Python. 2020. Disponível em: https://medium.com/programadores-ajudando-programadores/k-means-o-que-%C3%A9-como-funciona-aplica%C3%A7%C3%B5es-e-exemplo-em-python-6021df6e2572. Acesso em: 03 dez. 2022.

IFRN - Instituto Federal do Rio Grande do Norte. Projetos de extensão. Disponível em: https://dados.ifrn.edu.br/dataset/projetos-de-extensao. Acesso em: 27 nov. 2022.

SÁ, Lucas. Text Clustering with K-Means. 2019. Disponível em: https://github.com/lucas-de-sa/national-anthems-clustering/blob/master/Cluster_Anthems.ipynb. Acesso em: 01 dez. 2022.