

AS02: Representação Textual


Iniciar tarefa

- Vencimento Domingo por 23:59
- Pontos 15
- Enviando um URL de site
- Disponível 23 fev em 23:59 - 16 mar em 23:59

TAREFA

Esta é a tarefa **AS02: Representação Textual**, uma atividade prática que estimula o aluno a absorver **conceitos básicos de mineração e análise de texto**.

Problema


Gerar Embeddings Textuais em *Python* para cada termo do vocabulário da coleção [20 News Group Dataset](https://scikit-learn.org/0.19/datasets/twenty_newsgroups.html)  (https://scikit-learn.org/0.19/datasets/twenty_newsgroups.html), utilizando as seguintes abordagens de representação textual:

1. One-Hot Encoding
2. Count Vectors
3. TF-IDF
4. n-grams (2-grams)
5. Co-occurrence Vectors (Context Window = 1)
6. Word2Vec

Gerar o arquivo de saída `20News_XX.txt`, onde `XX` é o número da abordagem. Por exemplo, o arquivo `20News_01.txt` é a saída da


abordagem **One-Hot Encoding**.

Produto

O aluno deve entregar o link para um *notebook* em *Python* no ambiente do [Google Colaboratory](https://colab.research.google.com/notebooks)  (<https://colab.research.google.com/notebooks>) contendo a solução para o problema descrito anteriormente. O Notebook entregue DEVE RODAR INTEIRAMENTE em ambiente Google Colaboratory, não acessando arquivos locais fora do ambiente. Notebooks que não rodarem inteiramente no ambiente Google Colaboratory serão considerados inválidos.

Recursos

Para a execução da tarefa o aluno deve consultar as referências bibliográficas especificadas no [Programa do Curso \(https://pucminas.instructure.com/courses/157410/pages/programa\)](https://pucminas.instructure.com/courses/157410/pages/programa). A seguir encontram-se indicados alguns recursos e materiais de apoio para a execução da tarefa. Outras referências bibliográficas podem ser utilizadas, desde que devidamente citadas no produto.

- Livros e Materiais de Apoio:
 - Manning, Raghavan & Schütze, 2008
 - [Text Data Representation in Practice](https://towardsdatascience.com/text-data-representation-with-one-hot-encoding-tf-idf-count-vectors-co-occurrence-vectors-and-f1bccbd98bef)  (<https://towardsdatascience.com/text-data-representation-with-one-hot-encoding-tf-idf-count-vectors-co-occurrence-vectors-and-f1bccbd98bef>)
- Slides e Vídeos:
 - [LC02: Google Colaboratory](https://pucminas.instructure.com/courses/232662/pages/lc02-google-colaboratory) (<https://pucminas.instructure.com/courses/232662/pages/lc02-google-colaboratory>)
 - [LC09: One-Hot Encoding](https://pucminas.instructure.com/courses/232662/pages/lc09-one-hot-encoding) (<https://pucminas.instructure.com/courses/232662/pages/lc09-one-hot-encoding>)
 - [LC10: Count Vectors](https://pucminas.instructure.com/courses/232662/pages/lc10-count-vectors) (<https://pucminas.instructure.com/courses/232662/pages/lc10-count-vectors>)
 - [LC11: TF-IDF Vectors](https://pucminas.instructure.com/courses/232662/pages/lc11-tf-idf-vectors) (<https://pucminas.instructure.com/courses/232662/pages/lc11-tf-idf-vectors>)
 - [LC12: Co-Occurrence Vectors](https://pucminas.instructure.com/courses/232662/pages/lc12-co-occurrence-vectors) (<https://pucminas.instructure.com/courses/232662/pages/lc12-co-occurrence-vectors>)

- [LC13: N-Grams](https://pucminas.instructure.com/courses/232662/pages/lc13-n-grams) (<https://pucminas.instructure.com/courses/232662/pages/lc13-n-grams>)
- [LC14: Neural Word Embeddings](https://pucminas.instructure.com/courses/232662/pages/lc14-neural-word-embeddings) (<https://pucminas.instructure.com/courses/232662/pages/lc14-neural-word-embeddings>)