Unsupervised learning model proposal

project 4 of bootcamp data science T5

Covid 19 Literature Clustering.

Goal:

Cluster similar research articles about COVID-19 together to make it easier for health professionals to find relevant research articles. Clustering can be used to create a tool to identify similar articles, given a target article

Tools:

1. Unsupervised Learning task because we don't have labels for the articles
2. Clustering and Dimensionality Reduction task
3. See how well labels from K-Means
4. Use Vectorization TF-IDF.
5. Use K-Means for clustering
6. Use t-SNE for dimensionality reduction
7. Use PCA for dimensionality reduction.
8. Classify.

Dataset Description

Cite: COVID-19 Open Research Dataset Challenge (CORD-19) | Kaggle

Kaggle Submission: COVID-19 Literature Clustering | Kaggle

Data size 256684 articles.

We manly focused on articles that written in English language (600 sample articles).