







DAT500 - PRESENTATION

Audun and Emil

Dataset

- Abstracts of 10,000 Covid Research Papers
- Collected from PubMed
- <https://www.kaggle.com/anandhuh/covid-abstracts>

About this file		
Title, Abstract and URL of 10,000 Covid Research Papers		
 title 	 abstract 	 url 
Title of research paper	Abstract of the research paper	URL of the research paper
10000 unique values	10000 unique values	10000 unique values
Real-World Experience with COVID-19 Including Direct COVID-19 Antigen Testing and Monoclonal-Antibo...	This article summarizes the experiences of COVID-19 patients diagnosed and treated at Faulkton Area ...	https://pubmed.ncbi.nlm.nih.gov/35008137

USE CASE

- Many research papers
- Find out if they are similar
- Motivation:
 - Could some papers be grouped?
 - Unnecessary papers?
 - Can we use this to look for plagiarism?

Related work

- Same dataset
 - Compare readability
 - Generate text using Markov Chains
 - <https://www.kaggle.com/roshanr11/interesting-eda-covid-abstracts-dataset>
- Similar dataset and analysis
 - Googles BERT
 - Analyse Covid related twitter posts; negative/neutral/positive
 - <https://www.kaggle.com/ludovicocuoghi/twitter-sentiment-analysis-with-bert-roberta/>

Algorithms

- Locality-sensitive hashing
- Encoding
 - One-Hot Encoding
 - MinHash
- Jaccard similarity
- MRJob for preprocessing