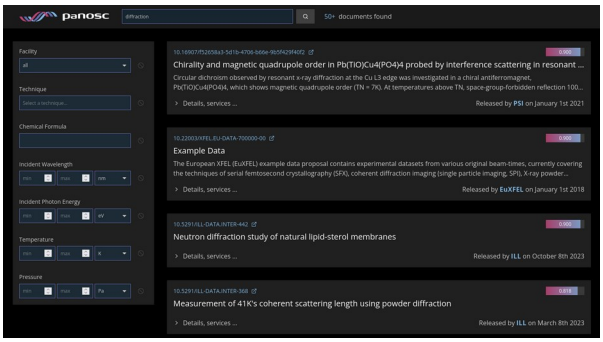
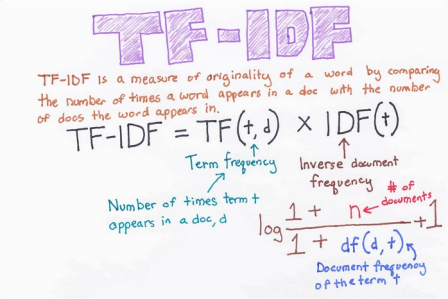
Student project
European Spallation Source
Data Management and Software Centre

Improving at scale information retrieval relevancy score service for distributed search





https://towardsdatascience.com/tf-term-frequency-idf-inverse-document-frequency-from-scratch-in-python-6c2b61b78558

Supervisor	?	Desired coding skills	★★★★☆
Co supervisor	Max Novelli	Desired data science skills	★★☆☆☆

DESCRIPTION

The European Spallation Source (ESS) has embraced the FAIR data principles. Its data policy mandate that all publicly founded research data must be made public after three years. In this context, ESS has participated in the european PaNOSC project and the creation of the PaNOSC data portal. The data portal allows users to search public data available at multiple research facilities across Europe. A key component for this infrastructure is the scoring service which assigns the relevancy score, the index between zero and one that indicates how relevant is the result in the context of the query submitted by the user. The current implementation of the scoring services needs to be improved and thoroughly tested, especially regarding the big data aspect when scaling up the number of documents.
In this project, we will first review the code of the scoring service, decide the best testing strategy, implement the testing strategy and integrate it in the continuous integration pipeline. We will analysis the testing data and provide performance statistics. We will continue by refactoring the code in order to improve performance and response time, while scaling up the number of data points

REQUIREMENTS

Experience with Python, micro services, docker and automation.
Willing to learn about: data management, information retrieval (IR), natural language processing (NLP) and machine learning techniques.

LINKS

https://www.panosc.eu
https://data.panosc.eu
https://en.wikipedia.org/wiki/Information_retrieval
https://en.wikipedia.org/wiki/Tf%E2%80%93idf

Contact Info

Max Novelli (max.novelli@ess.eu)