Results obtained through searching **any phrase, project, table, or column name,** including **date**, in the **search widget** utilising **semantic search modelling**.

An Exploratory Data Analysis and Data Mining Approach to Dataset and Dataset Relationship Discovery in Production Systems

INTRO

Goal: Develop a data search model capable of retrieving related datasets from the Information Hub portal.
Purpose: Identify the relationships between datasets, make it easier to search for relevant datasets, and enrich the studies of students and researchers using the data platform.

METHODS

1. Sample size: 3,146.
2. Exploratory Data Analysis using Ydata_profiling.
3. Information Hub project dataset collected to create a metadata & corpus for the model.
4. Four(4) pre-trained Semantic Search Models which does embeddings (vectorization).
5. Final selected model based on the similarity score: multi-qa-MiniLM-L6-dot-v1.

RESULTS

- Different texts tested on the four semantic search models and the model which continuously gave high similarity score is the multi-qa-MiniLM-L6-dot-v1 as compared to other three models as shown in figure below.

```
# using general text
query = "The department of soil sciences is doing a research about the humidity in the soil"
results = sentiment_search(public_df, query)
print(results)
```

```
   Top K       Project Name              Table Name  \
0      1  Soil collections          SASSYcollection
1      2  Soil collections   Soil_sample_database
2      3  Diagnostic Clinic  Diagnostic Clinic Trial

                                        Column Names     Score
0  [GPS_latitude, GPS_longitude, Collector, Count...  0.754926
1  [Unique, Sample_nr, Collector, Project, Countr...  0.752797
2  [latitude, longitude, samples_twigs_branches_n...  0.730209
```

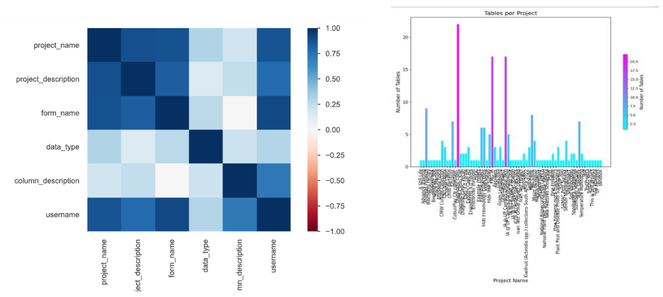
```
# using column name
query = "gtp_longitude"
results = sentiment_search(public_df, query)
print(results)
```

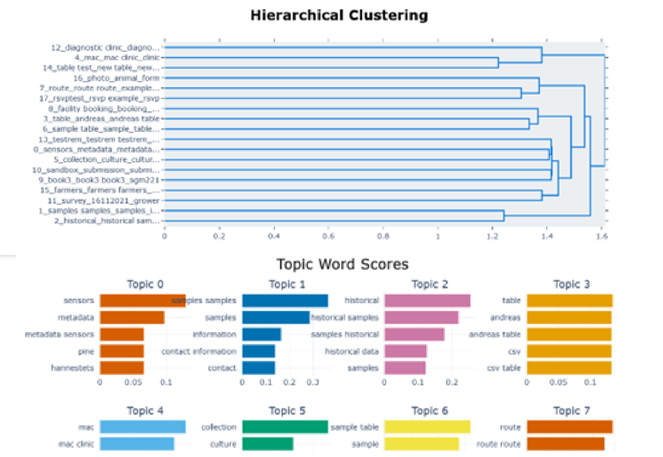
```
   Top K       Project Name   Table Name  \
0      1  Andreas Projects        book32
1      2  Andreas Projects         book3
2      3       Biodiversity  submissions

                                        Column Names     Score
0  [GPS_latitude, GPS_longitude, Year, Month, Day...  0.736144
1  [GPS_latitude, GPS_longitude, Year, Month, Day...  0.736144
2  [longitude, latitude, latitude, longitude, lon...  0.709641
```

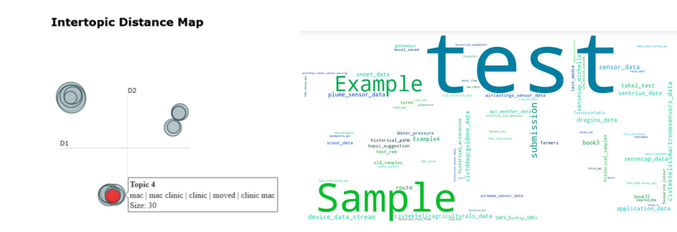
DISCUSSION

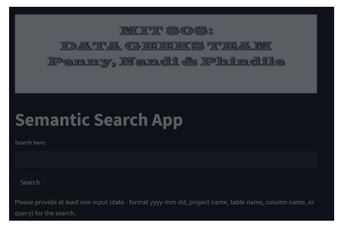
- In accordance with the user's search query, the model presents the top results in a dataframe along with the top position, the project, table, and column name.
- The search term could be a phrase, a project name, a table name, a column heading, or even a date in the format yyyy-mm-dd.
- The search outcomes are limited to the current datasets in the information hub data portal.

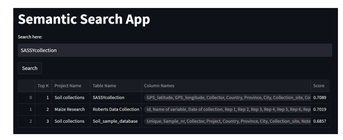
AMMO BAR











Penelope Matloga, Nandi Mnguni, Phindile Binda

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Department of Computer Science

Faculty of Engineering, Built Environment and Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en Inligtingtegnologie / Lefapha la Boetšenere, Tikologo ya Kago le Theknolotši ya Tshedimošo

Capstone Project - MIT 808

Course Coordinators:
Dr. Vukosi Marivate (vukosi.marivate@cs.up.ac.za)
Abiodun Modupe (abiodun.modupe@cs.up.ac.za)

Scan me