Bridging the Gap for Interdisciplinary Research Through Feeding Extracted Metadata into a Text-based Recommendation model with accurate results.

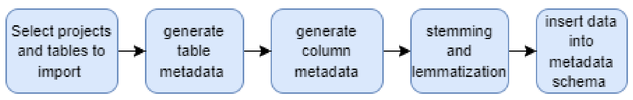
An Automated Metadata Mining and Dataset Recommendation System

INTRO

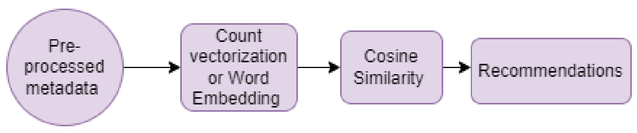
- Innovation Africa @UP is a collaborative research initiative that aims to promote sustainable pan-African development and economic growth.
- One way they have managed to reach this goal is through the Information Hub which is a research repository platform where its users can upload and share datasets, images, etc.
- The goal of this project is to build upon the search functionality of Information Hub by automatically identifying and recommending possibly relevant datasets to the users of the Information Hub.

METHODOLOGY

1. **Metadata Extraction:** Given that *any* tabular dataset of varying columns and rows could be imported into the Information Hub, the metadata extraction needed to make provision for datasets of varying dimensions.
   - Descriptive metadata is extracted that describes the project on dataset and column levels.
   - Stemming is used to derive tags for the tables and column names which is later used on the dataset recommendation model.

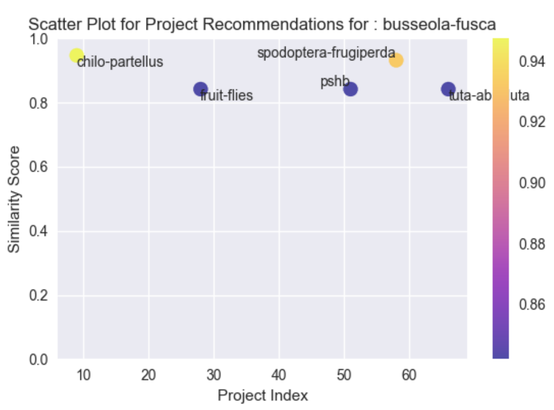


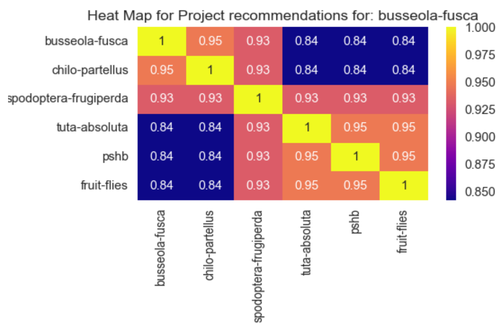
2. **Dataset Recommendation:** Given a project name as a search input, this model aims to make a recommendation of other relevant projects based on project metadata. This is achieved by:
   - Using the processed project text-based metadata to generate either a vector of terms or a word embedding.
   - Using the newly generated features to compute the cosine similarity between the input project and all the other available project.



RESULTS

- With no empirical way to measure the performance of the recommendation model, the results are manually assessed from a subset of randomly selected project names as inputs to the model.





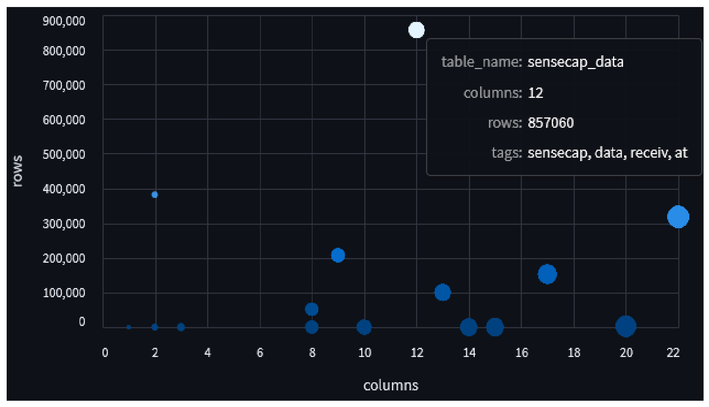
DISCUSSION

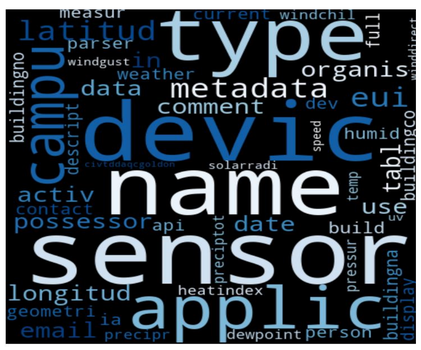
- Metadata extraction was achieved for tabular datasets of varying shapes and sizes.
- Applying stemming to table and column names proved to be an effective approach to providing data for a recommendation model based on text searches.
- For the feature engineering methods, both word embedding and vectorization of the metadata produce very similar recommendations, with minor discrepancies.

FUTURE WORK

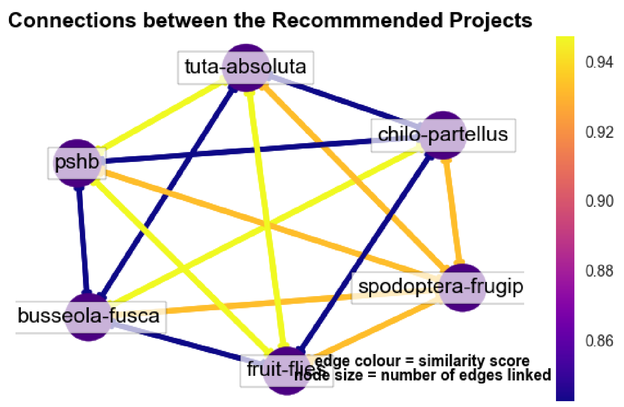
- Similar automated tagging approaches can be performed on non-tabular datasets.
- Adding an evaluation system for the recommendations produced, where the users give feedback through a rating. The feedback can then be used to improve recommendations.

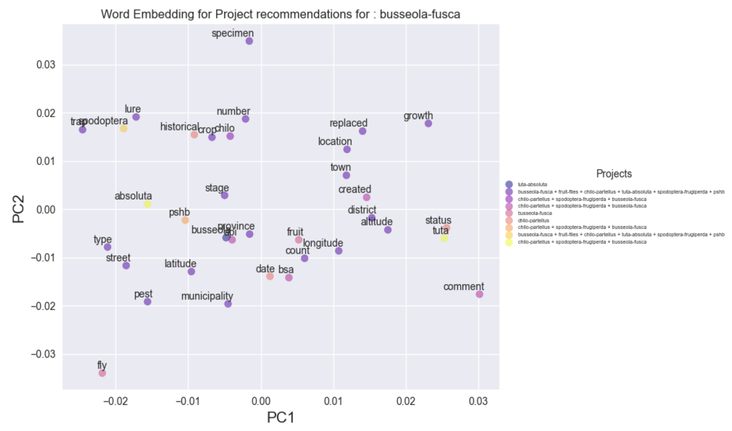
Metadata Extraction Visualization





Recommendation Model Result Visualization





Fiskani Banda, Kris Hamersma

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

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

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