Improved **document search engine** that enhances **user experience** by finding a correlation between **topics** and **SDGs** using **machine learning**

Topic modelling of the SDG articles using machine learning techniques

INTRO

- The provided SDG dataset is classified according to various SDGs. However, the document search functionality yields poor results.
- The aim of the project is to provide an intuitive document search engine to maximize user experience through usability.
- We explore the improvement of these results through visualisations and topic models

METHODS

1. Dataset was collected by SDG hub from various institutions and partners.
2. The data had 342k obs. and reduced to 284k obs. after pre-processing and augmentation.

Topic models

- SVD - Word similarity
- Latent Dirichlet Allocation (LDA)

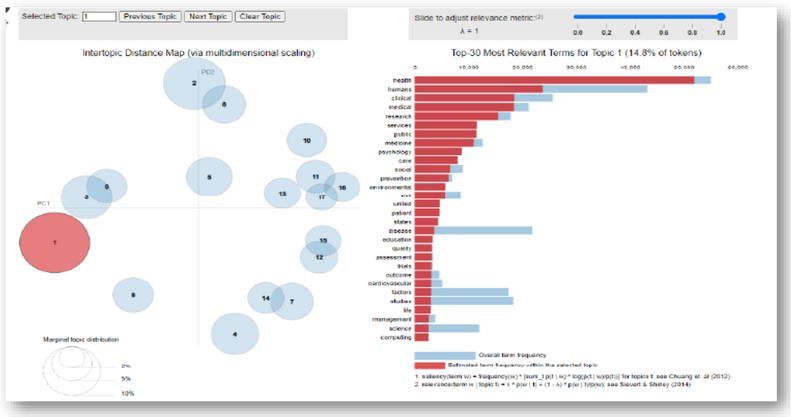
Visualization

- Plots (bar, line, word cloud...)
- PCA & LDA

RESULTS

```
[13] #Calculate and construct a similarity vector
     X_SVD_similarity = calculate_word_doc_similarity(X_SVD_emb)

     #Find 3 top similar words to the word passed as a parameter
     doc_occurence_most_similar('health',X_SVD_similarity,word2id,id2word,n=3)

     [('mental', 0.9841663760140622),
      ('professionals', 0.971887491461629),
      ('providers', 0.9649918897350027)]

[19] # Find records that contain the word searched for
     df = dataframe[dataframe['full_title'].str.contains('health')]
     #df

[20] df[df['articletype_ptr_id'] == 1321] #search for health
```

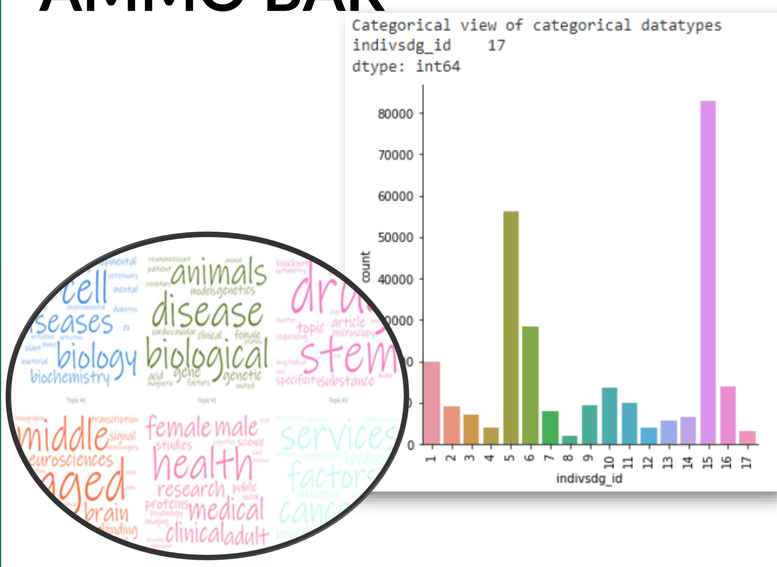
	articletype_ptr_id	full_title	authors
609	1321	Community and health facility influences on co...	['Stephenson, Rob', 'Tshibangu, Delphin']

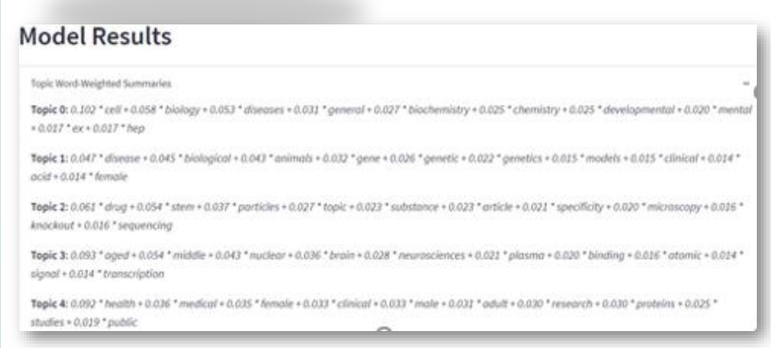


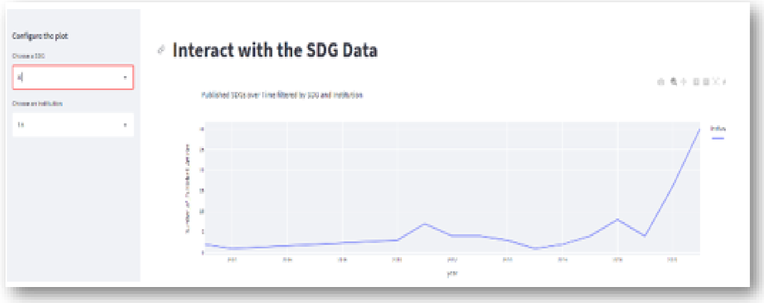
DISCUSSION

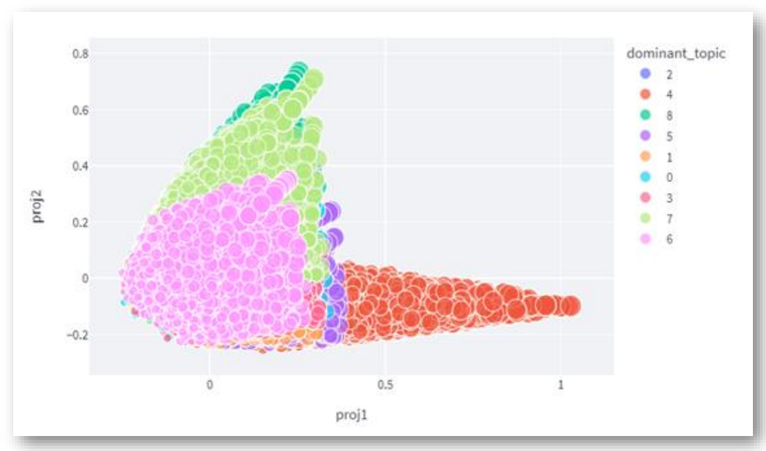
- The word similarity model has successfully been used as an ordering system to retrieve relevant results.
- The LDA model accurately allocates keywords to topics allowing for relevant results when using keywords

AMMO BAR









Teboho **Maloka**, Thato **Rachamose**, Ntsikayezwe **Faku**

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