

# 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

- Dataset was collected by SDG hub from various institutions and partners.
- 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

Teboho Maloka, Thato Rachamose, Ntsikayezwe Faku

### RESULTS

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

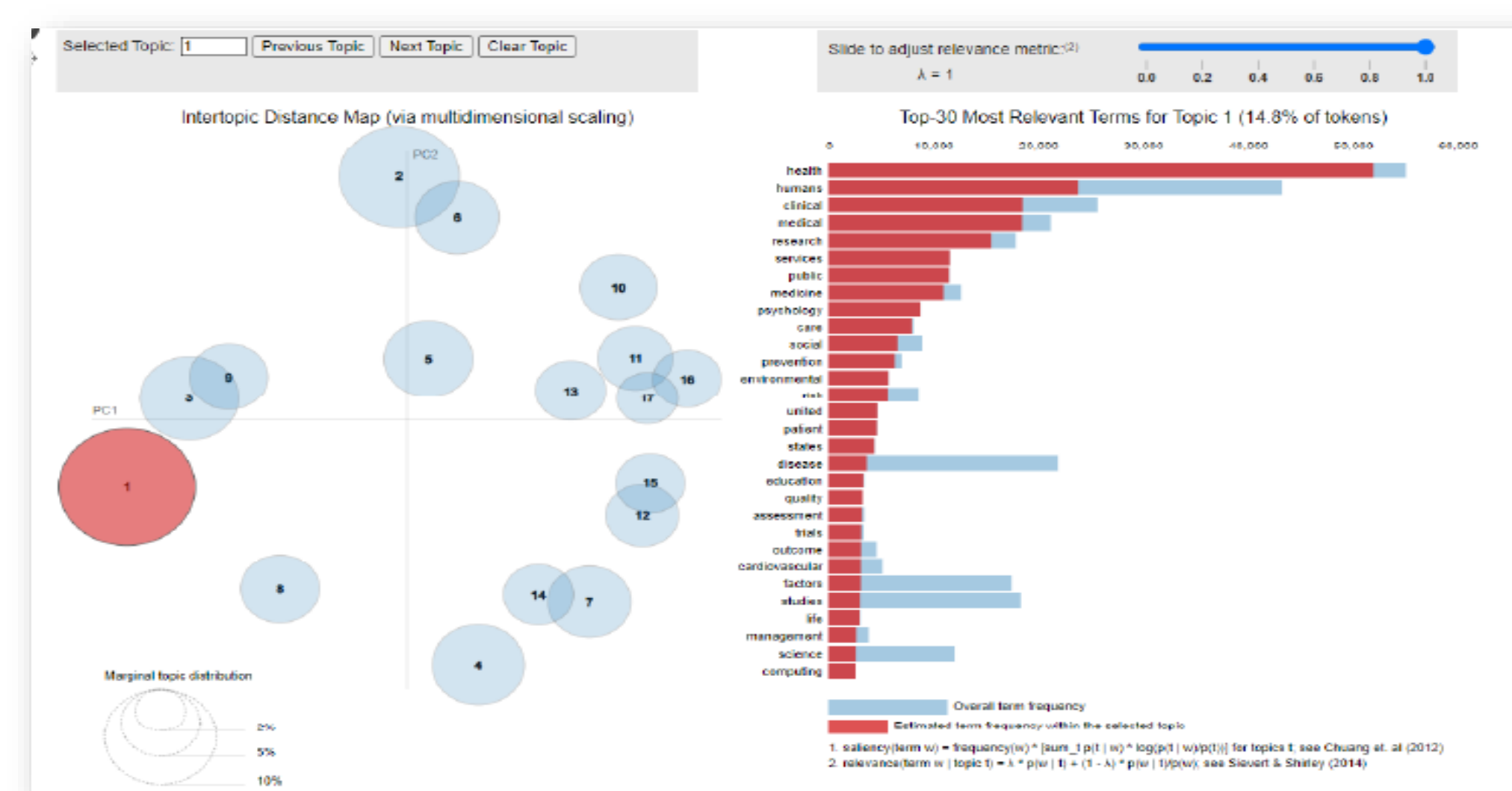
#Find top similar words to the word passed as a parameter
doc_occurrence_most_similar('health', X_SVD_similarity, word2id, id2word, n=3)

[('mental', 0.9841663769140622),
 ('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
```

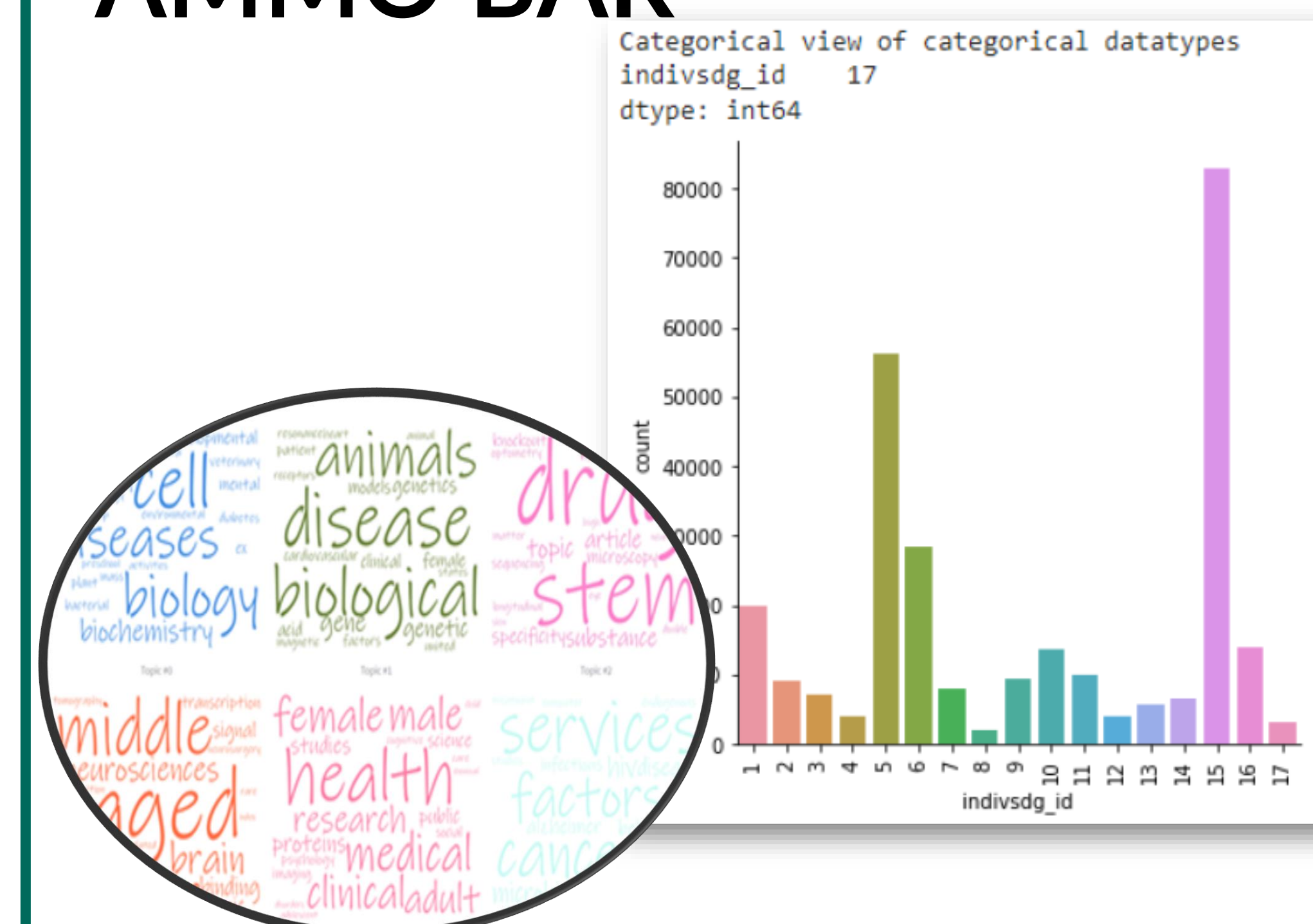
articletype_ptr_id	full_title	authors
609	1321 Community and health facility influences on co...	['Stephenson, Rob', 'Tshibangu, Delphin']



### 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



### Model Results

Topic Word Weighted Summaries

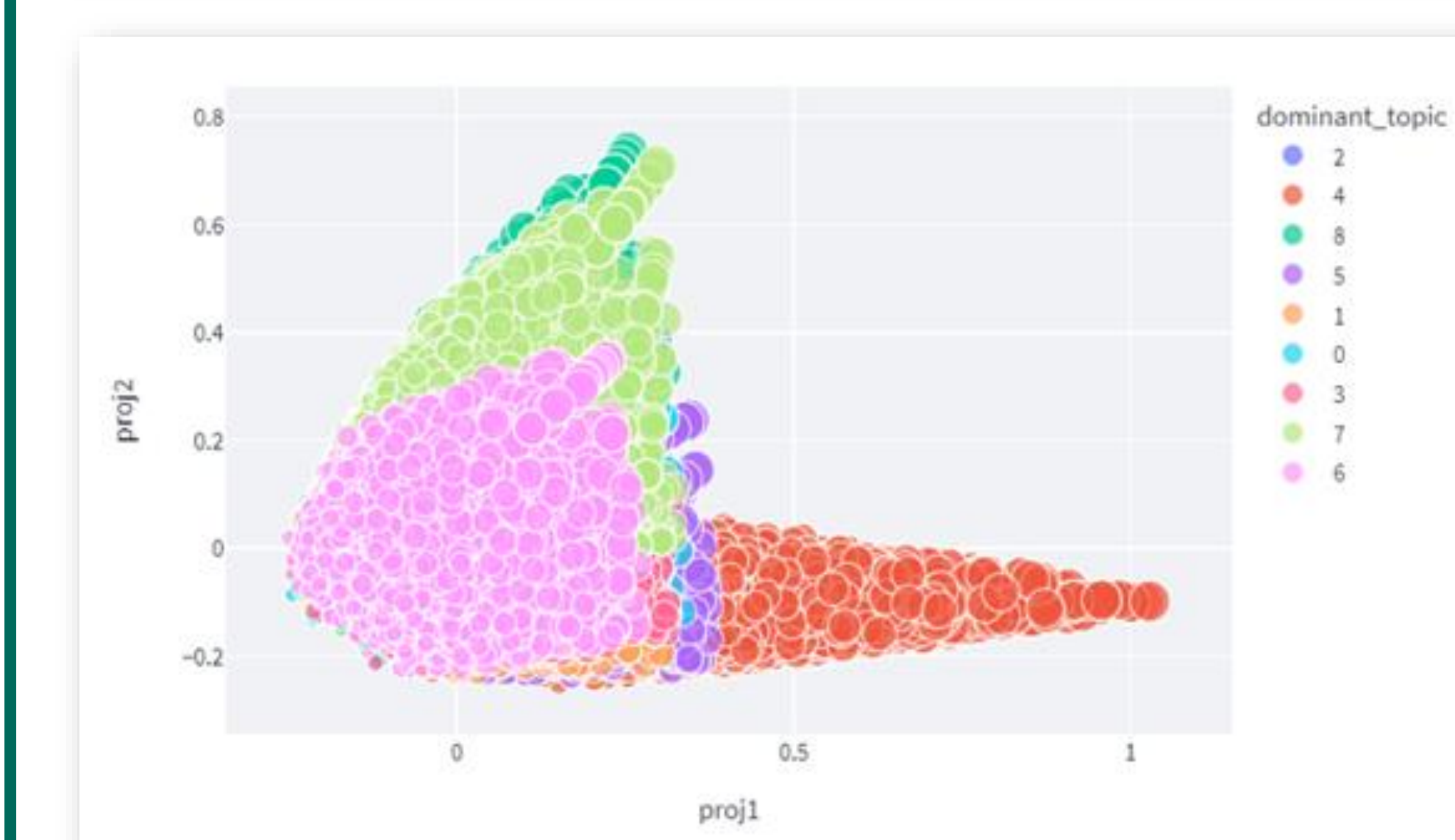
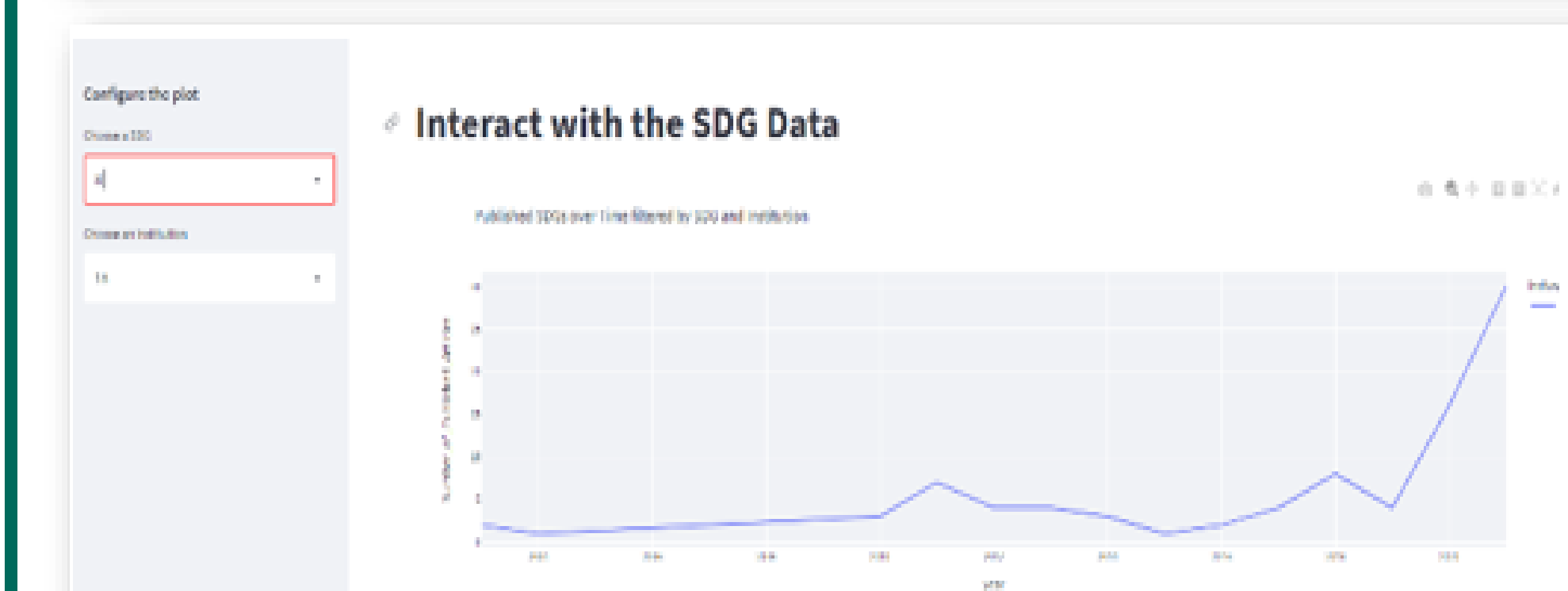
Topic 0: 0.102 \* cell + 0.058 \* biology + 0.053 \* diseases + 0.031 \* general + 0.027 \* biochemistry + 0.025 \* chemistry + 0.025 \* developmental + 0.020 \* mental + 0.017 \* ex + 0.017 \* sep

Topic 1: 0.047 \* disease + 0.045 \* biological + 0.043 \* animals + 0.032 \* gene + 0.026 \* genetic + 0.022 \* genetics + 0.025 \* models + 0.015 \* clinical + 0.014 \* acid + 0.014 \* female

Topic 2: 0.061 \* drug + 0.054 \* stem + 0.037 \* particles + 0.027 \* topic + 0.023 \* substance + 0.023 \* article + 0.021 \* specificity + 0.020 \* microscopy + 0.016 \* knockout + 0.016 \* sequencing

Topic 3: 0.093 \* aged + 0.054 \* middle + 0.043 \* nuclear + 0.036 \* brain + 0.028 \* neurosciences + 0.021 \* plasma + 0.020 \* binding + 0.016 \* atomic + 0.014 \* signal + 0.014 \* transcription

Topic 4: 0.092 \* health + 0.036 \* medical + 0.035 \* female + 0.033 \* clinical + 0.033 \* male + 0.031 \* adult + 0.030 \* research + 0.030 \* proteins + 0.025 \* studies + 0.019 \* public



Scan me



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)

