SDG dataset categorised into 16 Topics - based on the coherence analysis.

Visualising University Research and SDG Contribution in South Africa

INTRODUCTION

- Aim:
  - Visualising the dataset in the SDG Hub database as best as possible, using topic analysis => enables users to scan through few topics, rather than tons of documents
  - Improve dataset navigation through interactive tools
  - Overview of the research landscape
- Dataset
  - Obtained from the SA SDG Hub – abstracts were in a separate dataset.
  - 159,846 observations
  - 36 variables
  - Size = 118MB

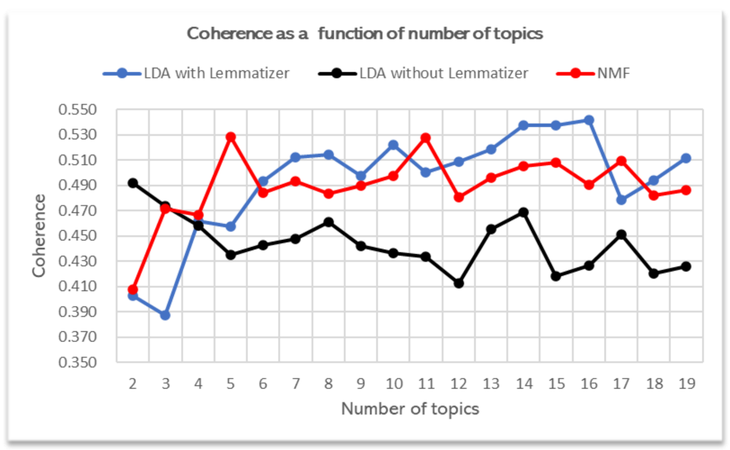
METHODS

Assess different unsupervised models for topic analysis

1. Source and join datasets
2. Explanatory Data Analysis
3. Preprocessing
4. Modeling
   - Latent Dirichlet Allocation or LDA (without lemmatizing) => **0.43**
   - Latent Dirichlet Allocation or LDA (with lemmatizing) => 0.54
   - Non-Negative Matrix Factorization or NMF (with lemmatizing) => 0.49
5. Final model selection: LDA with lemmatizing, based on coherence
6. Visualisation

RESULTS

- Modeling Results: Coherence plot analysis **(0.54)** – optimum number of topics under LDA with lemmatizing = 16

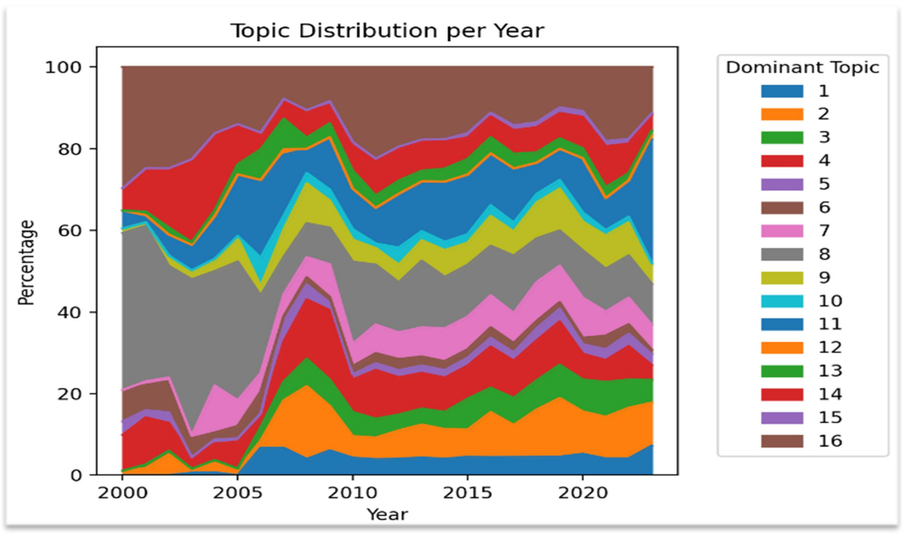


DISCUSSION

- **Intertopic Distance Map:**
  - Per topic => size of bubble rep. % of tokens in the corpus (Topic 16, 13.6% of tokens)
  - Distance between bubbles rep. how closely related their key words are
  - Bar chart: key words from selected topic(bubble). Term frequency within selected topic + overall term frequency within the corpus

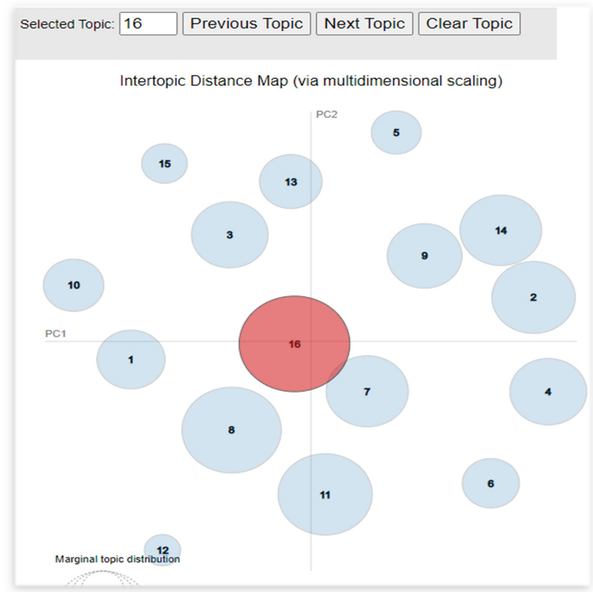
- **Wordclouds**:
  - Prominent words from selected topic (bubble) – model, method, data, etc

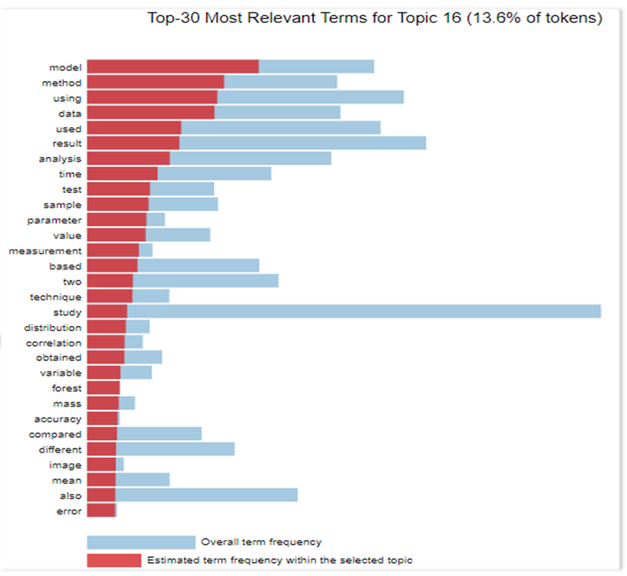
- **Other Charts**
  - **Topic Distribution Chart**– shows prevalence of different topics over time
  - SDG classifications across topics **(heat map)** – Topic 1 and SDG 3 have the highest frequency
  - Number of publications per institute **(bar graph)** – MIT, Michigan (around 40 000), Univ. of Pretoria (above 30 000)

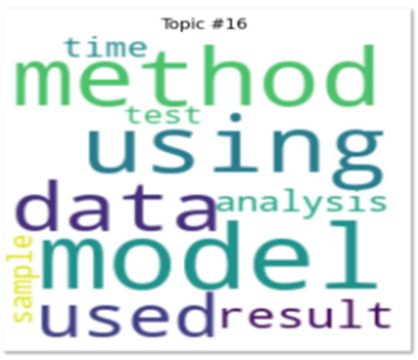


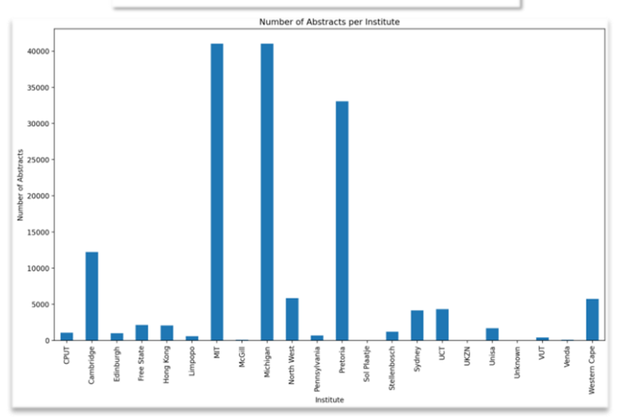
FUTURE WORK

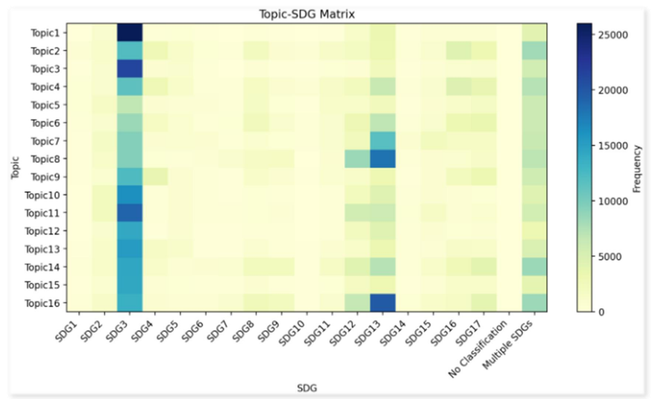
- Alternative topic analysis models e.g. Hierarchical Dirichlet Processes (HDP)
- User feedback incorporation to establish needs from the community
- Enhanced visualisation techniques e.g. network graphs











Lister Kom, Hlonela Mntonintshi

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)