Zondo Commission on State Capture:
What we missed

INTRODUCTION

- The transcripts from the Zondo Commission still contain a wealth of potential information that has not yet been tapped into.
- This information contains names of officials, organisations, and individuals and the interconnectedness between them that were not part of the original work and reports that were released on the Zondo Commission.
- The aim is to marry the worlds of data science and journalism in order to draw new insights that could generate new leads for journalists to pursue.

MOTIVATION

- Journalists do not have sufficient time or manpower to manually go through three years' of testimony in the Zondo Commission on State Capture transcripts to make new connections and uncover new leads.
- Data science is required to categorise trends, map out connections, and point out outliers from the Zondo Commission transcripts.

METHODS

1. Exploratory Data Analysis (EDA)

- EDA was performed on the Zondo Commission court transcripts text documents.
- The data was cleaned, new attributes created to enhance the data, and stored in a csv file for later use.
- The EDA process includes word frequency analysis, bi-gram and tri-gram analysis, and initial clustering that was visualised using PCA and t-SNE.

2. Modelling

- **Topic Modelling**
  - Latent Dirichlet Allocation (LDA) was used. Part of speech (POS) filtering, lemmatisation, and removal of special characters, punctuation, and stopwords were done as pre-processing steps for LDA.
  - Main themes from the State Capture documents were identified. The topic labels were derived from the top 20 words per topic.
- **Named Entity Recognition**
  - Person names, Organisations, FAC's, etc. were extracted from the data using the SpaCy en_core_web_md model.
  - Occurrence count of entities in the whole dataset is used to classify top and bottom entities.
- **Network Analysis**
  - Relationships between people testifying and those that they implicated were defined and a Network highlighting most influential people was constructed using Gelphi. The Networkx library was used to construct the Network Graph.

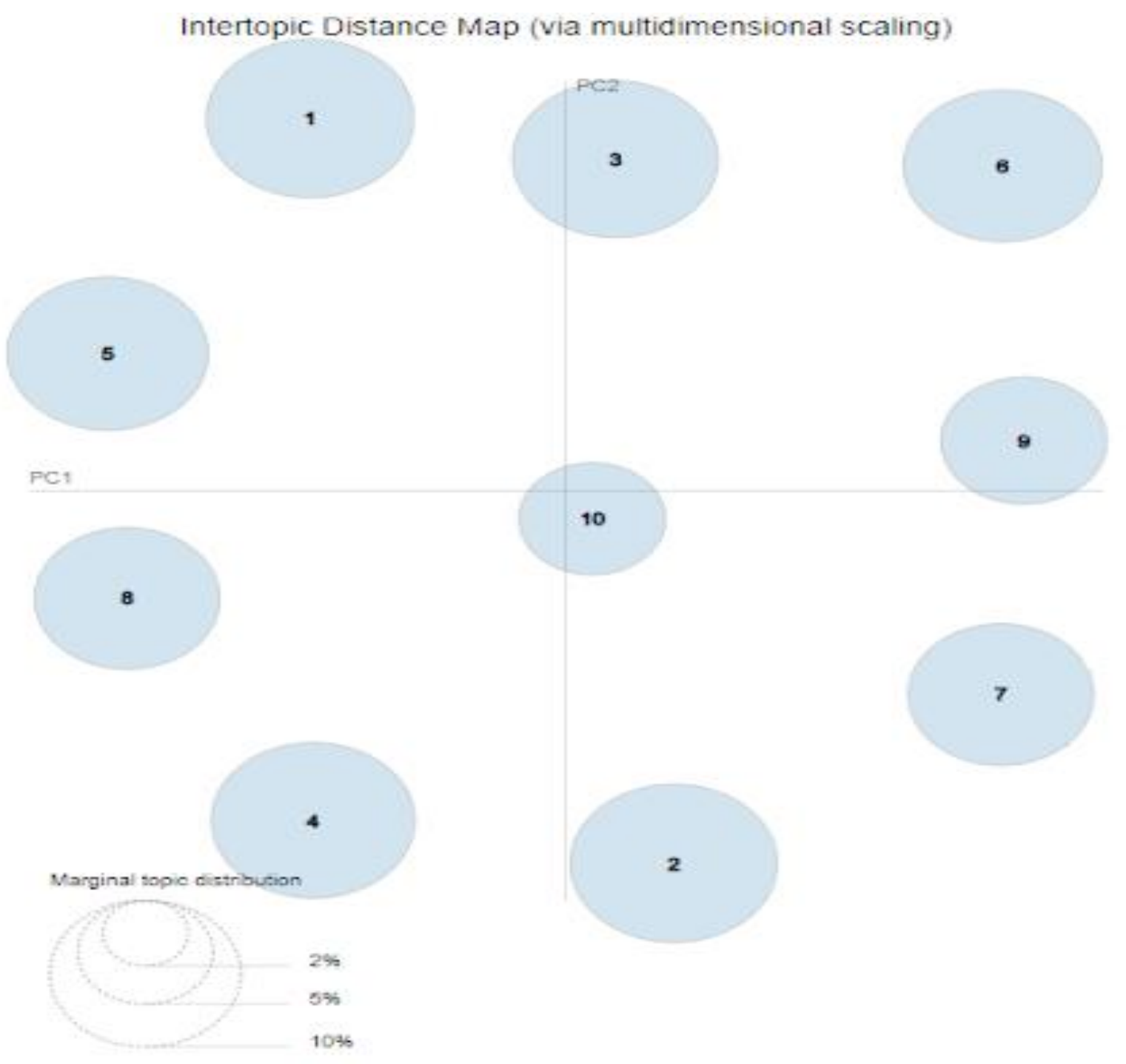
EXPERIMENTS

- Pronouns were included in the LDA model. The initial model with only verbs, adverbs, nouns, and adjectives did not give interpretable topics.
- The fuzzywuzzy library was used to remove duplicate names in the network. This improved the results and readability of the network graph.
- The SpaCy NER medium model was used. Results from the large model were similar, however the medium model was better at labelling entities than the small (trained on less data).

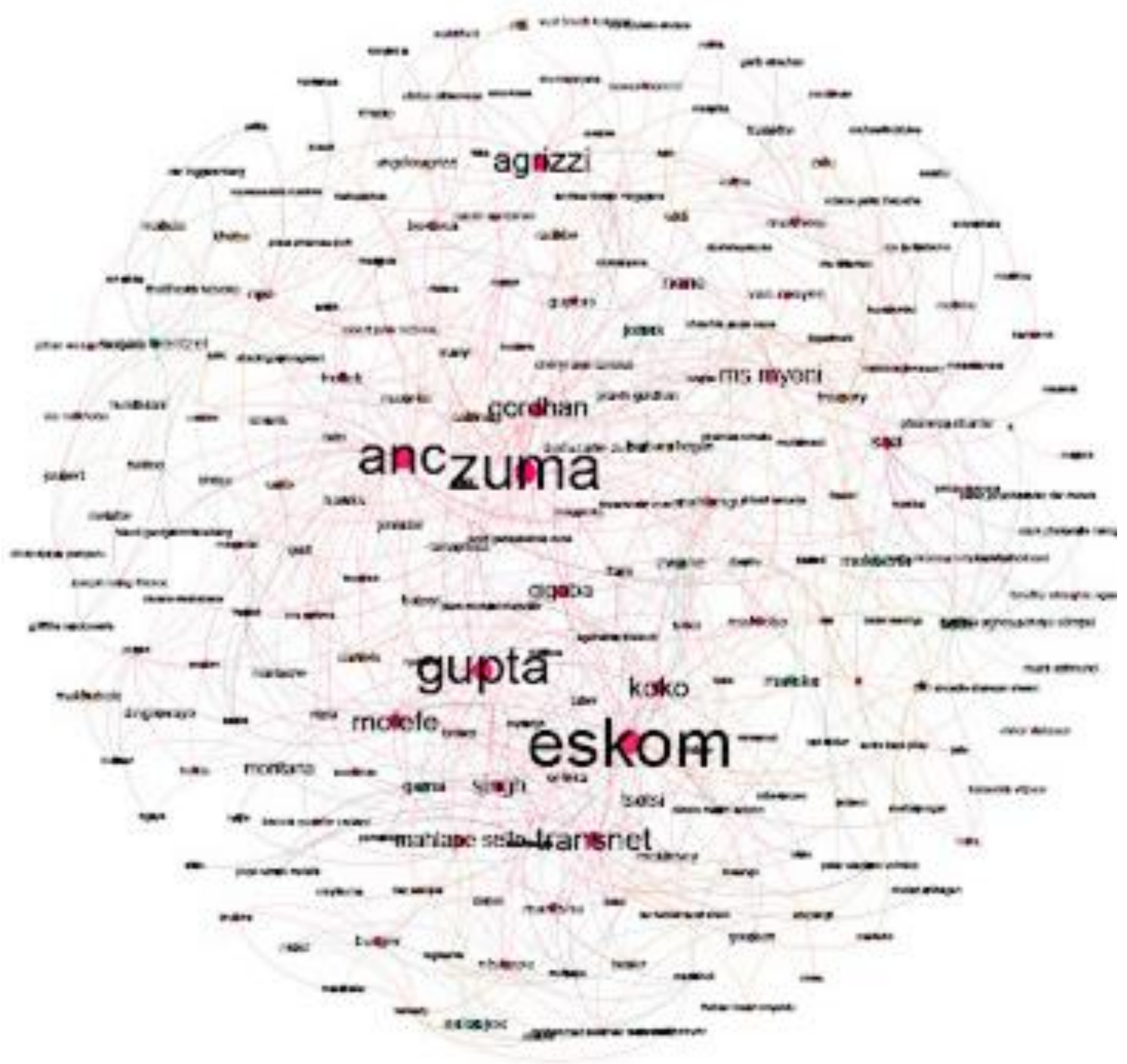
RESULTS

- The optimal number of topics for the model is 10 with a coherence score of 0.526.
- Good separation between topics are present and topics have good marginal distributions.
- Top names include Brian Molefe (SA businessman), Koko Mantshela (former Eskom exco), and Angelo Agrizzi (former Bosasa COO).
- The top organisation include Eskom, ANC, and SAA.
- Using Network Analysis the most influential entities in the Network are Zuma, ANC, Eskom, Agrizzi, Eskom, and Gupta.

LDA Model Intertopic Distance Map



Network Graph Results

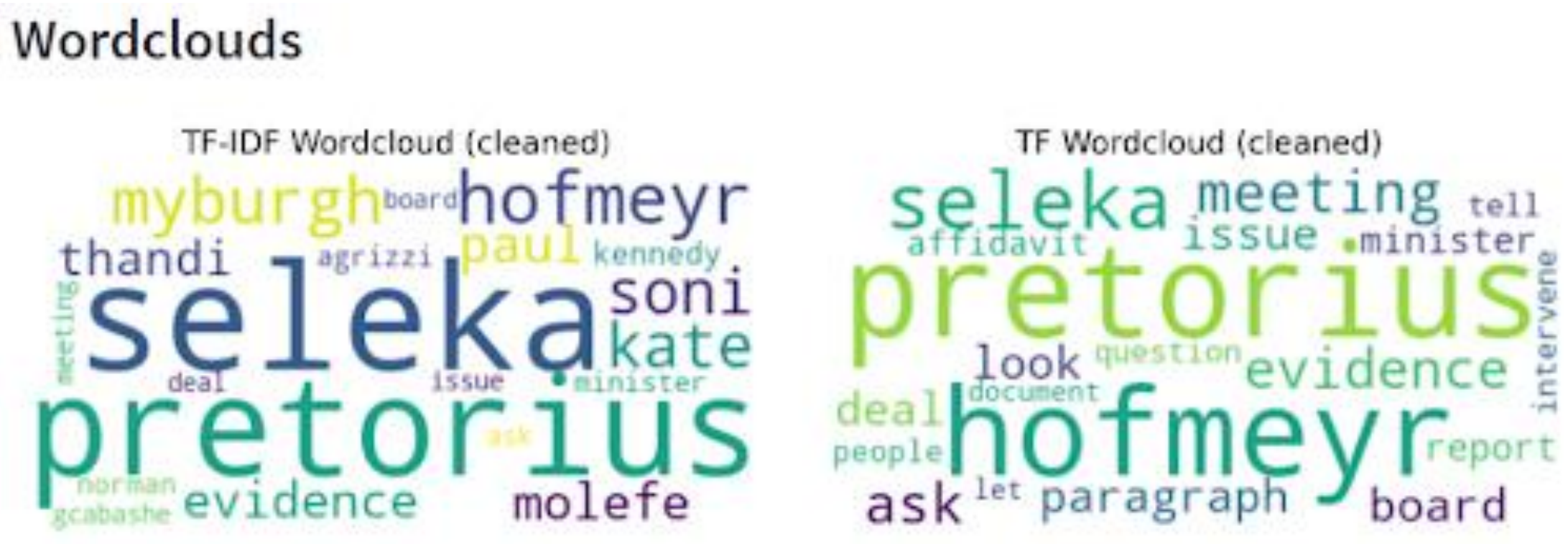


DISCUSSION

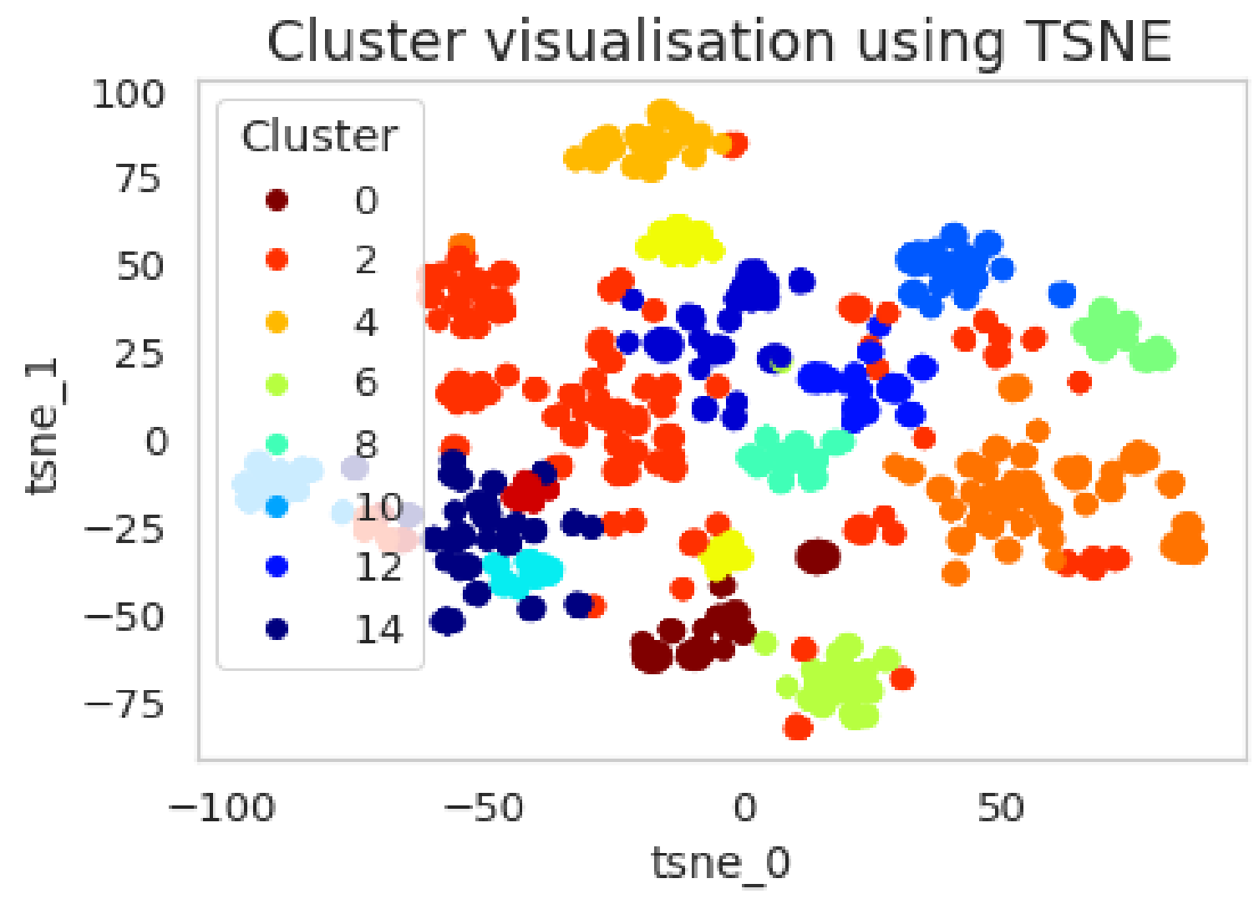
- The court transcripts hold more value as it is the raw testimonies.
- Incorrect names and organisations were highlighted as outliers due to spelling mistakes, showing how the quality of the data can significantly affect the results of ML/AI application.
- The use of raw court transcripts implicate a possible use of informal language in testimonies and thus informal and grammatically incorrect sentences. As a result the performance of NER was impacted as some entities were incorrectly labelled.
- Marrying the fields of Journalism and Data Science can provide journalist with a better means to uncover important information and leads to pursue.

Matimba Shingange, **Erika** Scholtz

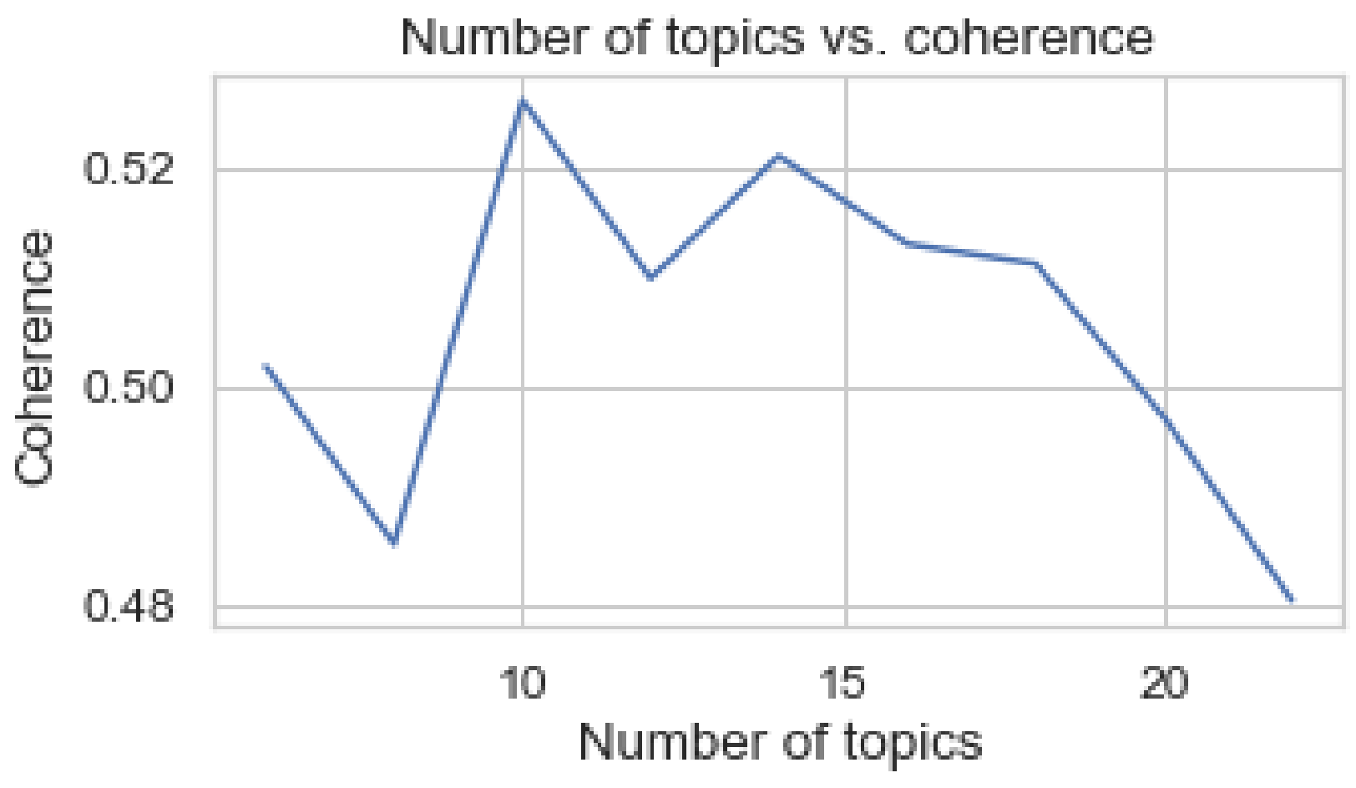
Word Clouds for TF and TF-IDF



Cluster Visualisation Using T-SNE



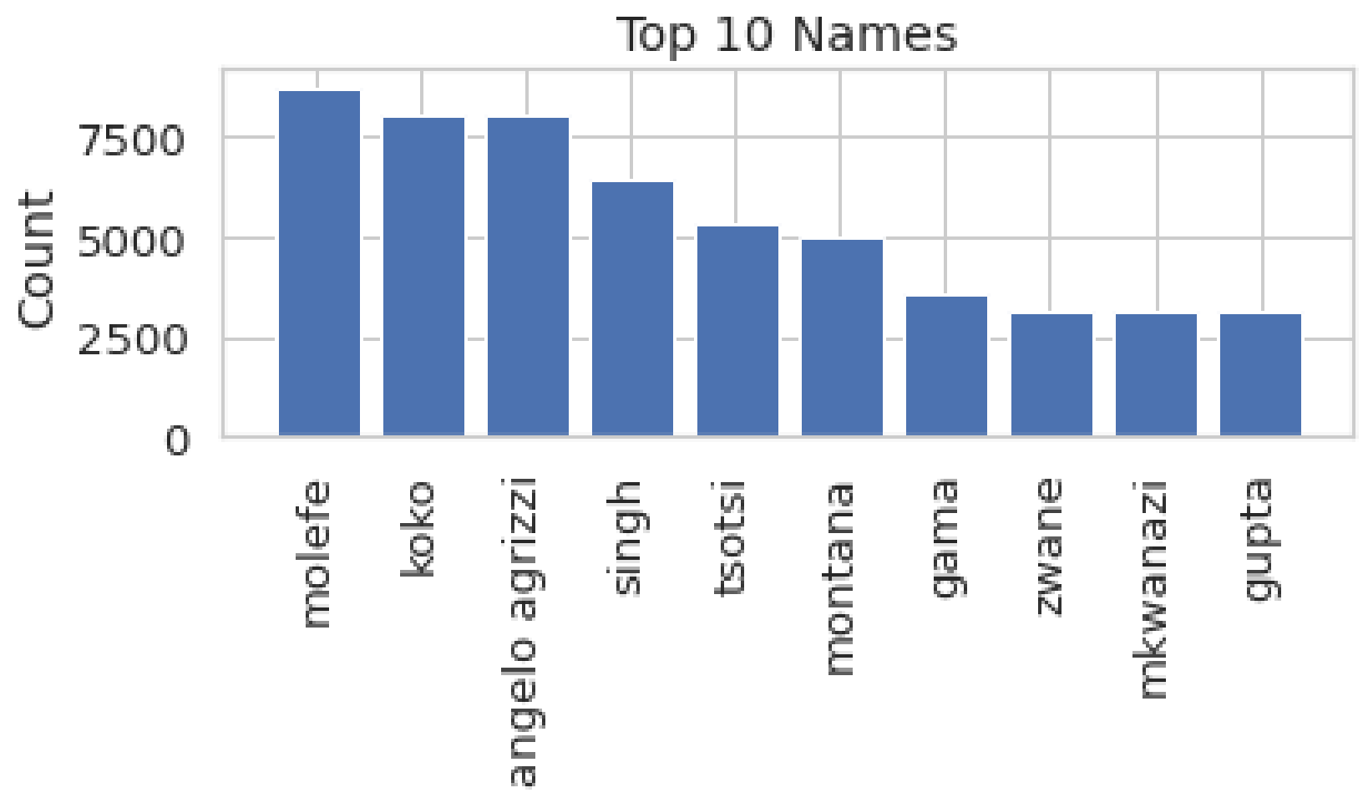
Number of Topics vs Coherence



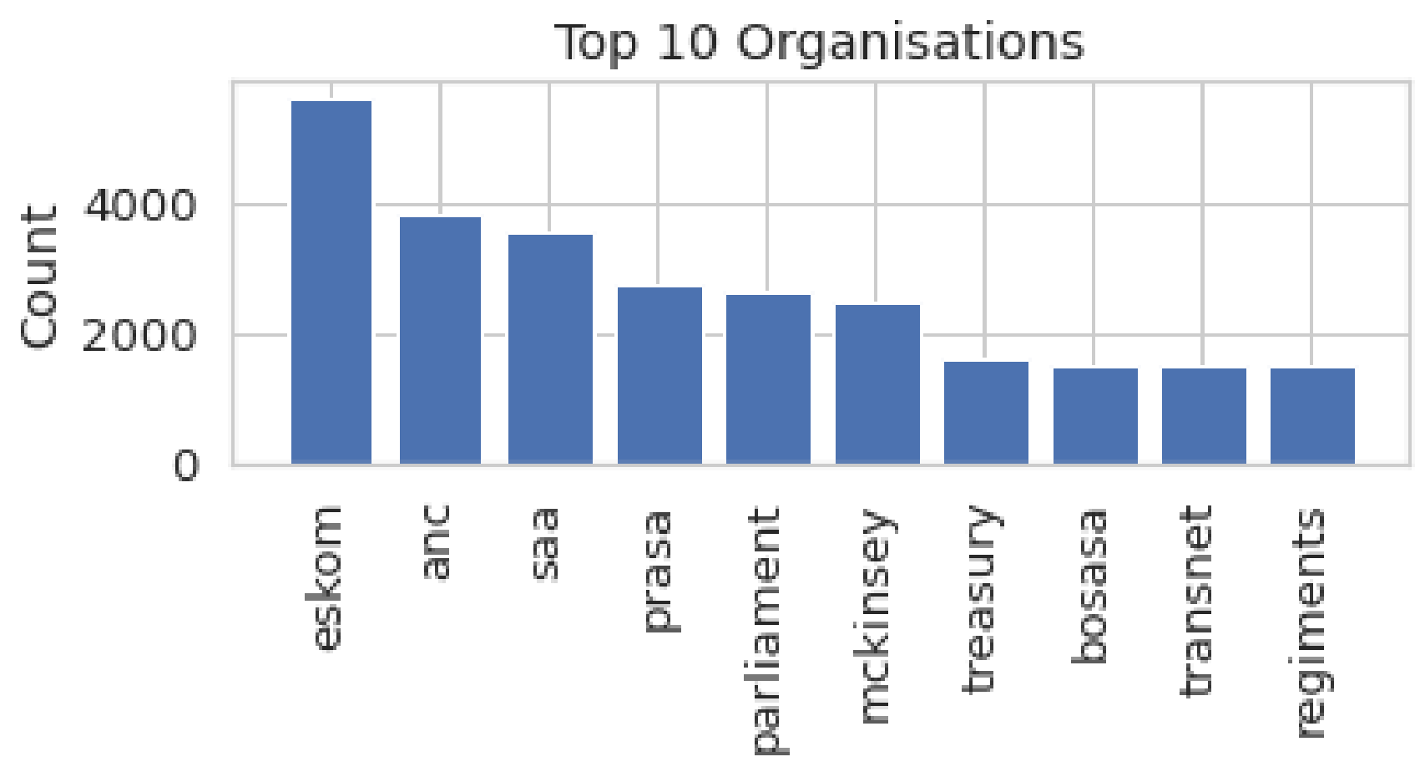
LDA Model Labelled Topics

Topic	Top words	Topic label
1	tsotsi, board, eskom, daniels, dr_ngubane, klein, burger…	Eskom and Denel State Capture
2	montana, zwane, zwane, prasa, holden, coal, transnet…	Prasa and Transnet State Capture
3	kwinana, mbana_pete_thabethe, mokhesi, sodi, dlamini, saa…	SAA and Free State State Capture
4	koko, ndzeku, naidoo, roelofse, eskom, sabc, khuba, chabi…	IPID, Hawks, Crime Intelligence, and SABC State Capture

Top 10 Names from NER



Top 10 Organisations from NER



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