Zondo Commission on State Capture – what we missed?

Using NLP to unfold Zondo Commission transcripts

INTRODUCTION

The Zondo commission has been an ongoing inquiry into the State Capture since 2018 focused on fraud and corruption cases in South Africa. Over 400 hearings of evidence presented and over 300 witness testimonies have been documented. Journalists face a laborious and time-consuming challenge of manually analyzing this enormous pool of information for leads going through thousands of transcripts – this gives way to automated methods using NLP to unearth key information of the State Capture.

Research questions aimed to answer are as follow:

1. Who are the major role players (entities) in the State Capture?
2. What are the entity relationships and communities / clusters that exist is the State Capture?
3. What are the predominant themes that can be uncovered concerning the major communities of the State Capture?
4. What are the missed terms of reference or topics in mainstream media relating to entities in the State Capture?

METHODOLOGY

1. Pre-processed and cleaned 386 text files provided by the Mail & Guardian detailing the testimonies during Zondo Commission of Inquiry
2. Used *Named Entity Recognition (NER)* to identify categories of entities – people and organisations
3. Initial investigation of relationships involving top person and organisation entities using cosine similarities and clustered the *Gensim Word2Vec* representations using *t-SNE* model
4. *Network analysis* on all Zondo Commission transcript data to gain insight on influential entities in the State Capture, visualized interesting inter-cluster links that exist and extracted influential communities using *Louvain Community Detection*
5. Extracted themes using *Top Modelling* from influential entities and communities within the network
6. Scrapped and pre-processing new articles from a News Tool and used *Gensim's Doc2Vec* model to extract vectors and compared similarities with transcripts to surface what was missed in online media publication

EXPERIMENTS

NER

- Comparing accuracy of entity categorization between *spaCy* and *NLTK* NER models.

Clustering entities

- *Skipgram* word2vec model vs *CBOW* word2vec

Network Analysis

Creating edges between nodes experiments:

- *NER* per row (text processed per speaker) vs *Search Algorithm* for every sentence to find centre entity and creates edge to context entity within n = 5 left and right text words window (experiments with window size)

Outlier detections

Cross references with scraped media articles:

- *Topic Analysis* to compare topics in transcripts and articles vs *Similarity Detection* to determine which spaces in the transcripts were not covered by the articles (experiment runs for hyper-parameter optimization)

RESULTS

NER

- Top person entities: *Gupta, Matshela Koko, Jacob Zuma, Siyabonga Gama.*
- Top organisational entities: *Eskom, Transnet, SAA, Prasa*

Clustering entities (High cosine similarities)

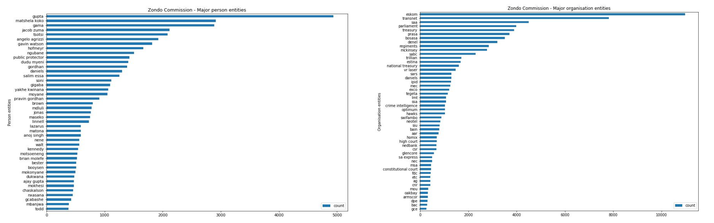
- Gupta cluster: Former finance ministers Malusi Gigaba and Mcebisi Jonas
- Eskom cluster: Optimum Coal Mine, Exxaro, Gupta-owned Oakbay and Tegeta and private companies (Mckinsey, Trillian, Regiments)
- Transnet cluster: Private companies as Eskom above and others - PWC, Deloitte, Nedbank, Neotel and Telkom

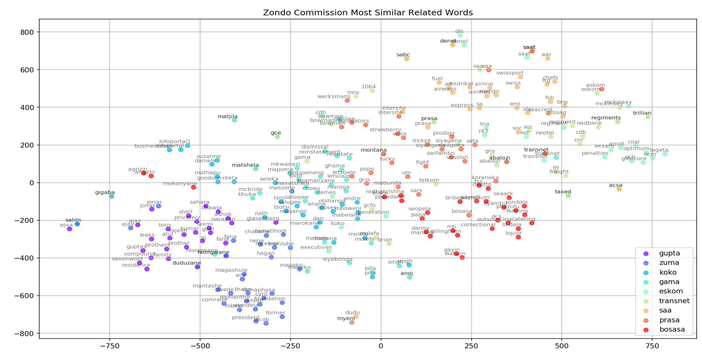
Network Analysis

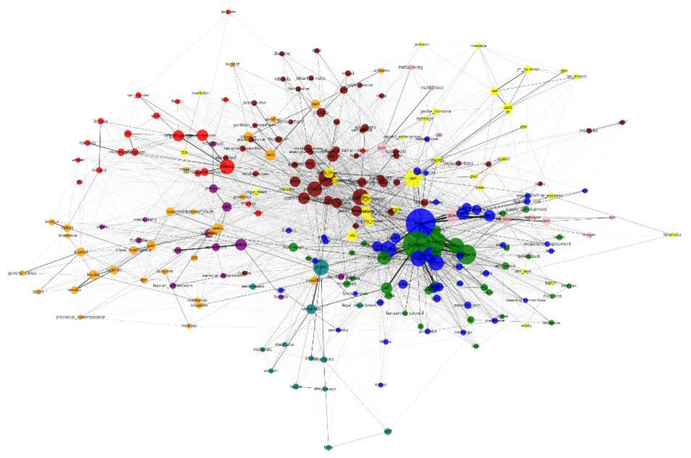
- Top 8 communities of influence in order of node degree: Eskom, Transnet, Prasa, Bosasa, SAA, Denel, Crime Intelligence, Parliament.
- Gupta cluster with most inter-cluster connections
  - Gupta family and business partner (Salim Essa) connected to major entities through associated connections: Jet Airways (SAA), VR Laser (Denel), Tegeta (Eskom), Transnet (Liebherr)

Outlier detection

- Low similarity on inter-cluster between the major SOE entities, involvement of private companies, connections of CR17 to OCH and Tegeta, involvement of other government officials not emphasized (Fikile Mbalula, Gwede Mantashe, Ace Magashule), connections to Lynne Brown and Van Rooyen to National Treasury (where mainstream media focused this attempted capture on Jacob Zuma)







K. Mongoai, M. Ledwaba

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