A Comparison of Information Retrieval Pre-processing Algorithms applied to African Historical Data

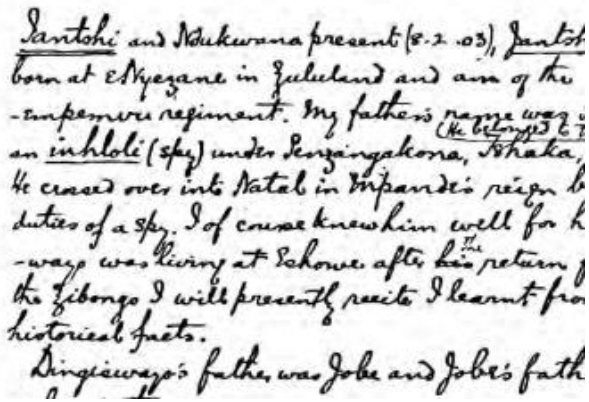
Soham Singh and Hussein Suleman
Department of Computer Science
School of IT
University of Cape Town

Last edited: 11/26/2022

Slide count:

African Historical Data

- Southern African history/archaeology has received more attention in recent years.
  - Post-colonization, post-Apartheid.
- Early records from external parties (with inconsistencies/bias).
- Many new and ongoing archaeological/historical projects, leading to new archives.


Santshé and Nkukwana present (8.2.03), Santsh
born at eNyezane in Zululand and am of the
-mpemvu regiment. My father's name was
an inhlolé (spy) under Senzangakona, Shaka,
He crossed over into Natal in Mpande's reign
duties of a spy. I of course knew him well for
-ways was living at Ekowe after his return
the Zibongo I will presently recite I learnt fro
historical facts.
Dingiswayo's father was Jobe and Jobe's fath

The Decolonization Debate (and IT)

- Ex-colonies in Africa are trying to reverse the lingering effects of colonial practices.
- Is technology neutral?

- Suppose a researcher is searching for the famous Zulu king Shaka kaSenzangakhona.
  - Many different spellings of his name exist.
  - Drawings/pictures are often inconsistent.
  - Names in isiZulu are often embedded in English texts.

Questions

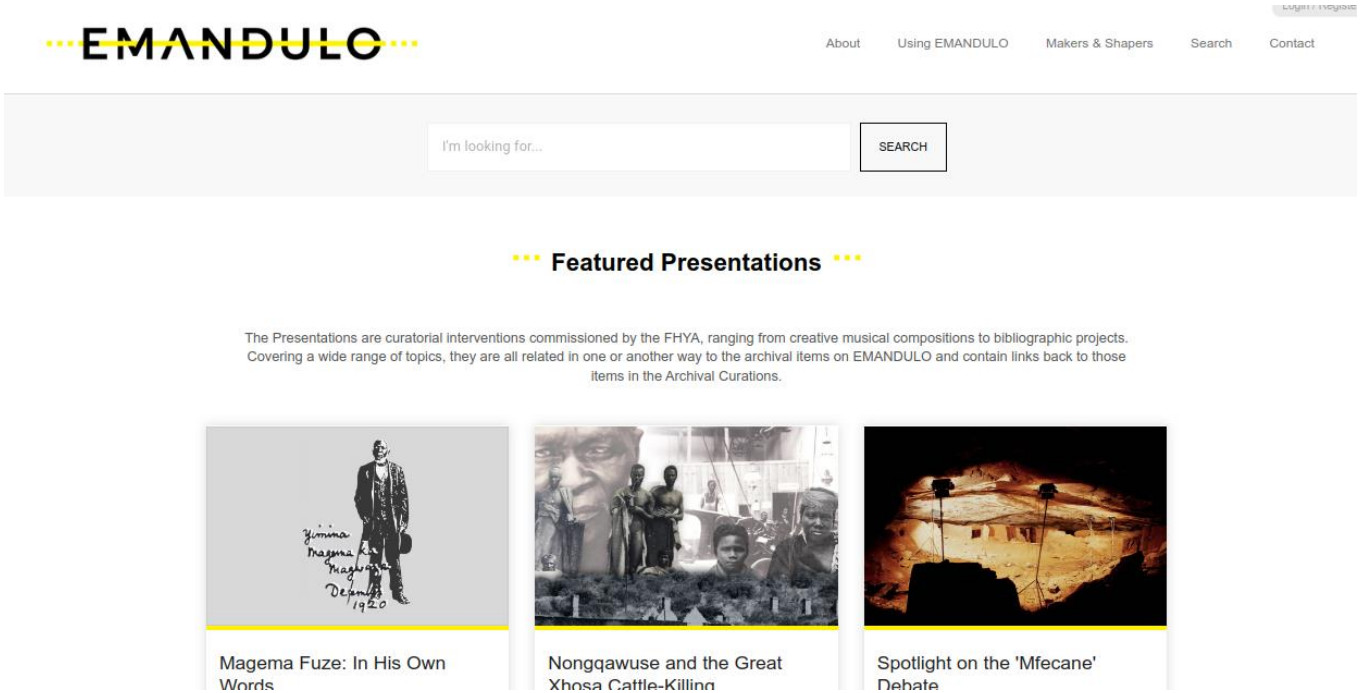
- Will traditional search engine / information retrieval (IR) algorithms work?
- Can stemming work?
  - There are no good stemmers for local agglutinative languages (isiZulu, isiXhosa, etc.).
  - Words have many variations.
- Can image search work?
  - There is little accompanying text.
  - Most people do not have the vocabulary for what they are looking for.

Experimental Goal

- Test common IR text processing algorithms:
  - Stemming, Stopping
  - Thesaurus, ...and all permutations
- Test common IR image processing algorithms:
  - Pyramid histogram of oriented gradients
  - Auto colour correlogram
  - Edge histogram
  - Colour and edge directivity descriptor
  - Colour layout
  - Joint composite descriptor

Dataset

- Five Hundred Year Archive (pre-colonial era heritage)
- 1345 text documents, 5708 images

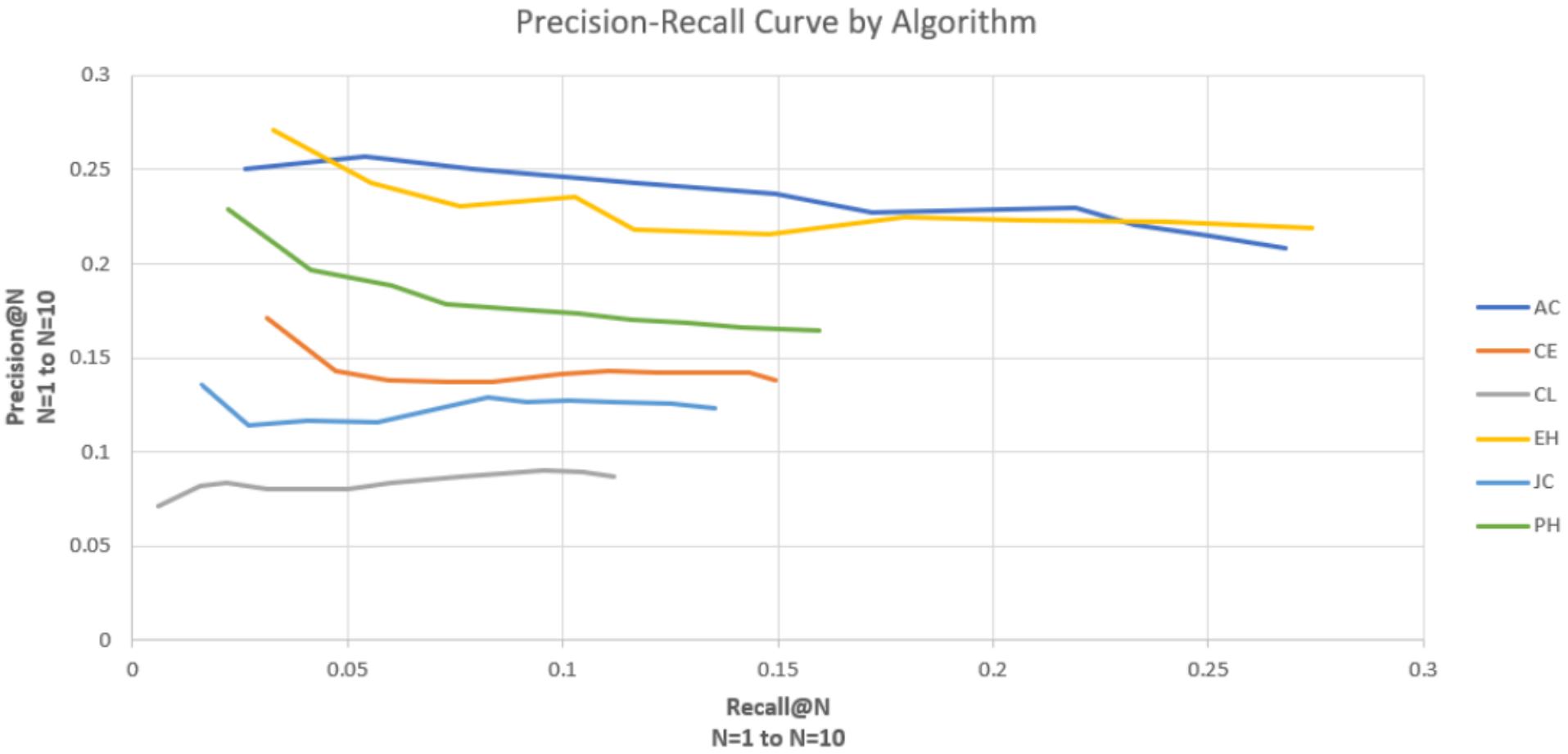


Experimental Procedure

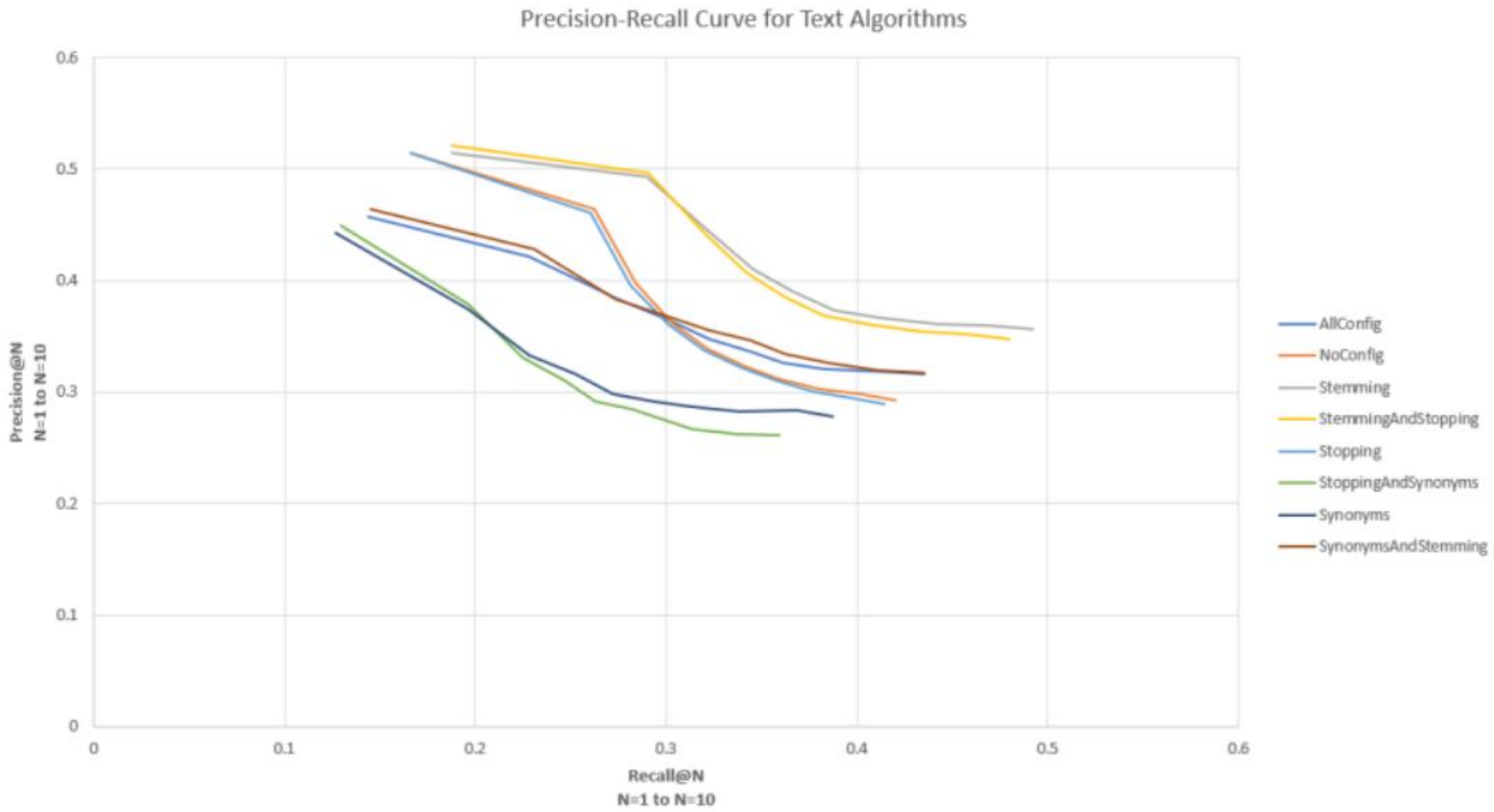
- SOLR+LIRE used for IR engine.
- 70 text queries
  - Based on 54 information needs (from FHYA/Google)
- 70 image queries
  - Based on Smithsonian Institute collections

- De-duplicated supersets of results were judged for relevance by 3 participants for each query.
- Results were then computed for each algorithm.

Results – Image Algorithms



Results – Text Algorithms



Conclusions

- Stemming seems more promising than anticipated.
  - Multilingual African language group stemmer could also help.
- Thesaurus use was not promising.
  - We may need a specialised thesaurus for collection.
- Image algorithms based on shape performed better than those based on colour.
  - Selection of algorithms depends on dataset.
- Search/processing algorithms can improve retrieval of historical low-resource collections if algorithms are chosen carefully.

Questions / Comments / Discussion

hussein@cs.uct.ac.za

IT | SCHOOL OF IT