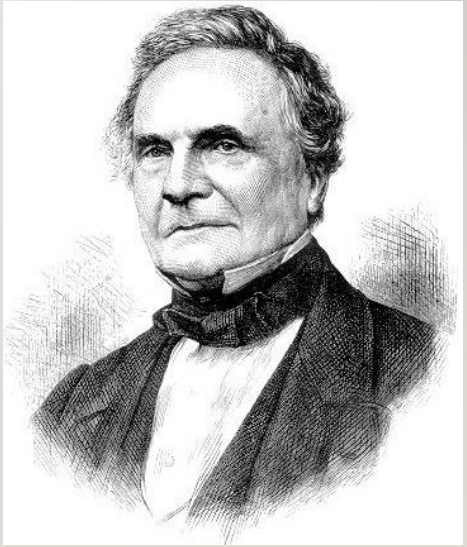
PREPROCESSING LATVIAN PUBLICATIONS FOR CONTENT SIMILARITY DETECTION

Valdis Saulespurens, research and development at NLL

ORIGINS OF THE CHALLENGE

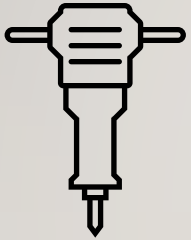
- *On two occasions I have been asked [by members of Parliament], 'Pray, Mr. Babbage, if you put into the machine wrong figures, will the right answers come out?' I am not able rightly to apprehend the kind of confusion of ideas that could provoke such a question.*

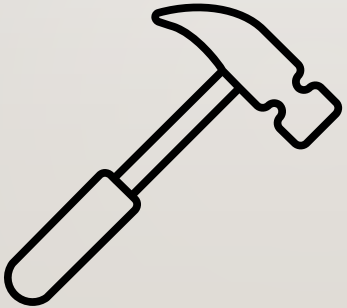


CLEANING UP THE MESS

(CC) http://www.diygenius.com/

TOPICS COVERED



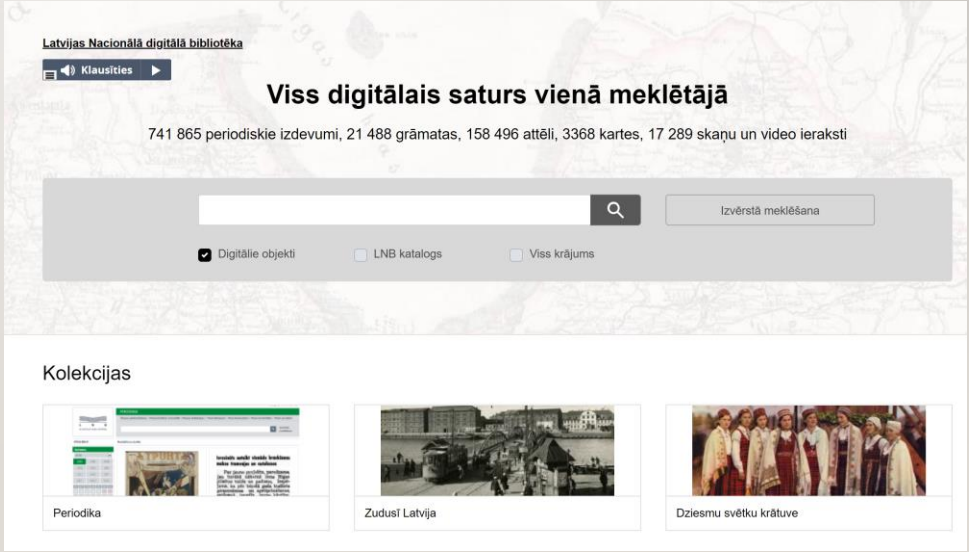


- Goals of the project
- Tools of the trade
- Technologies used for retrieval,cleanup, storage,analysis, deployment

DATA STORAGE INFRACTURE AT LNB

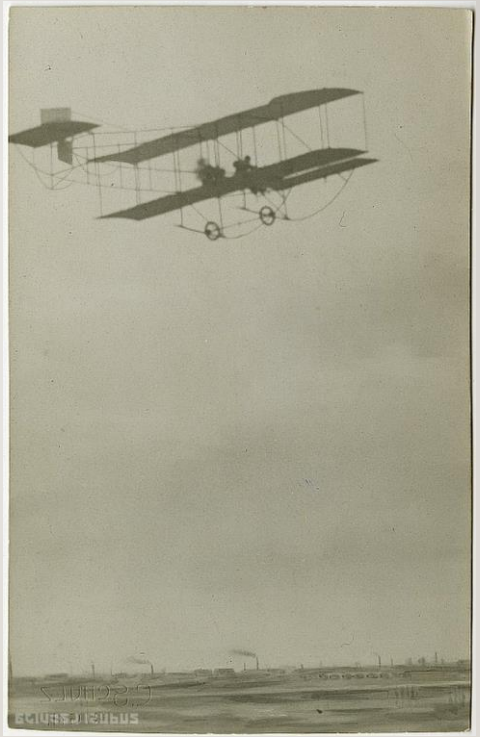
- DOM documents
- (not HTML DOM!)

- Over 700k periodicals
- Size – approaching 1PB
- Storage – a mixture of
- databases(MS SQL and others)
- caching solutions



PILOT PROJECT GOAL

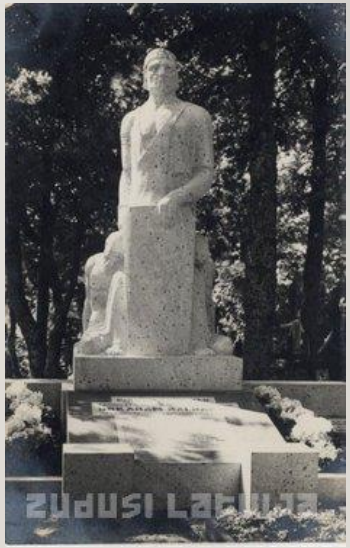
- Pick two corpora from 1920-1940 - (advised by historians)
- Extract
- Clean up
- Tagging - markdown
- Simple similarity metrics among documents
- New document similarity measurement



ANALYSIS OF OSCARS KALPAKS DOCUMENTS

- commander of 1st Latvian Independent Battalion
- 6 January 1882–6 March 1919

- Subgoal
  - Measure similarity in documents mentioning Kalpaks
  - Emphasis on special dates (birth,death, other)
  - Common theme extraction



COLLECTION I

Latvijas Kareivis - Latvian Soldier

Official Ministry of Defense publication

in between WWI – WWII

Many general interest stories

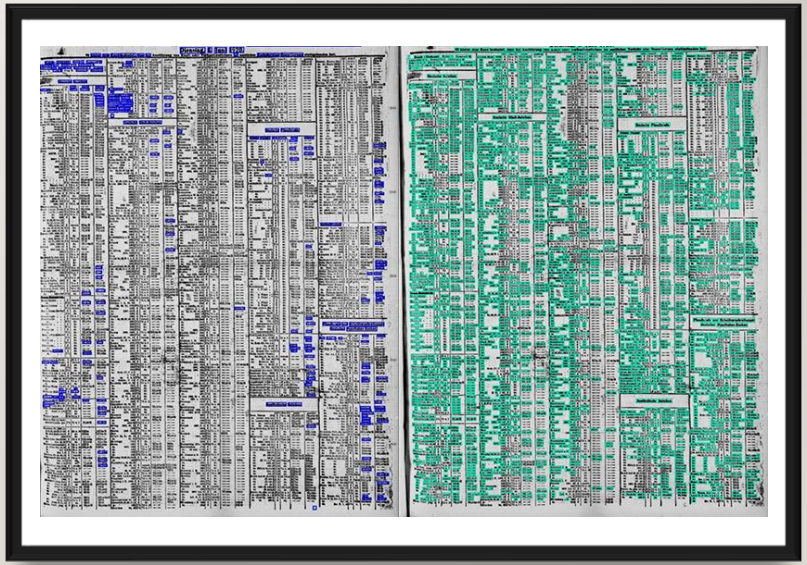

LATVIJAS
KAREIVIS

COLLECTION I - EXTRACTION

198,888 documents -> keep 49k row Excel – cleaned with Pandas – extracted with C# db calls

COLLECTION I - CHALLENGES

- OCR quality
- Pictures
- Segmentation



COLLECTION II

Valdības Vēstnesis – Offiical Latvian State News

1921-1940

Extra challenge – old typography at the start


Valdības Vēstnesis

Latvijas valdības

oficiāls laikraksts

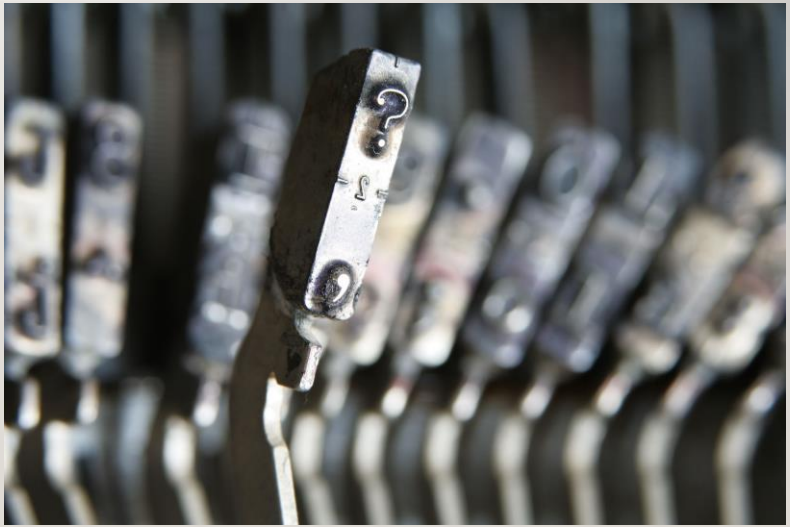
COLLECTION II - EXTRACTION

6206 publications x 50 files each

Filtered to 4125 publications x aprox 22 files each

DATA TYPES EXTRACTED

- Ocr xml
- Metadata.xml
- Jpg – for sanity checking results
- PDF – of whole page is possible



METADATA

- Types of metadata collected
- Date
- Segment
- Issue
- Publication - given
- URI - into DOM

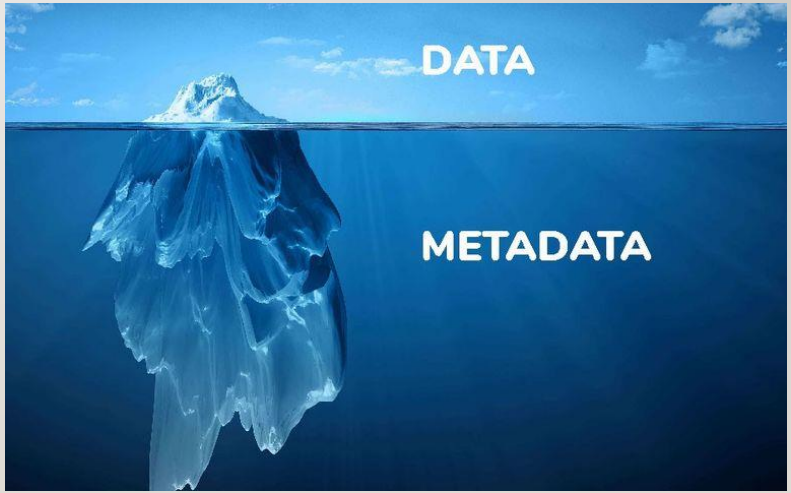


Image courtesy of Zeenea

TOOLS OF THE TRADE

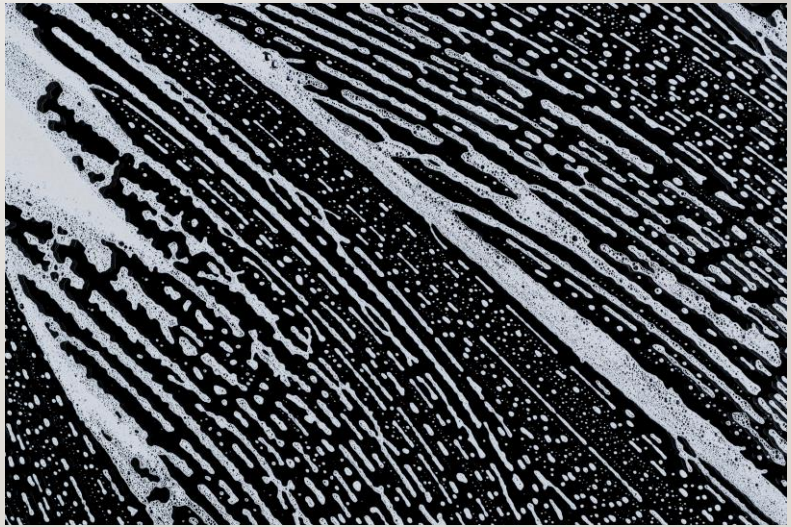
- Jupyter Based Notebooks (mostly Python)
- Local deployment on library premises
- Git -> local Gitlab installation/private Github

CONVERTING XML TO PLAIN TEXT

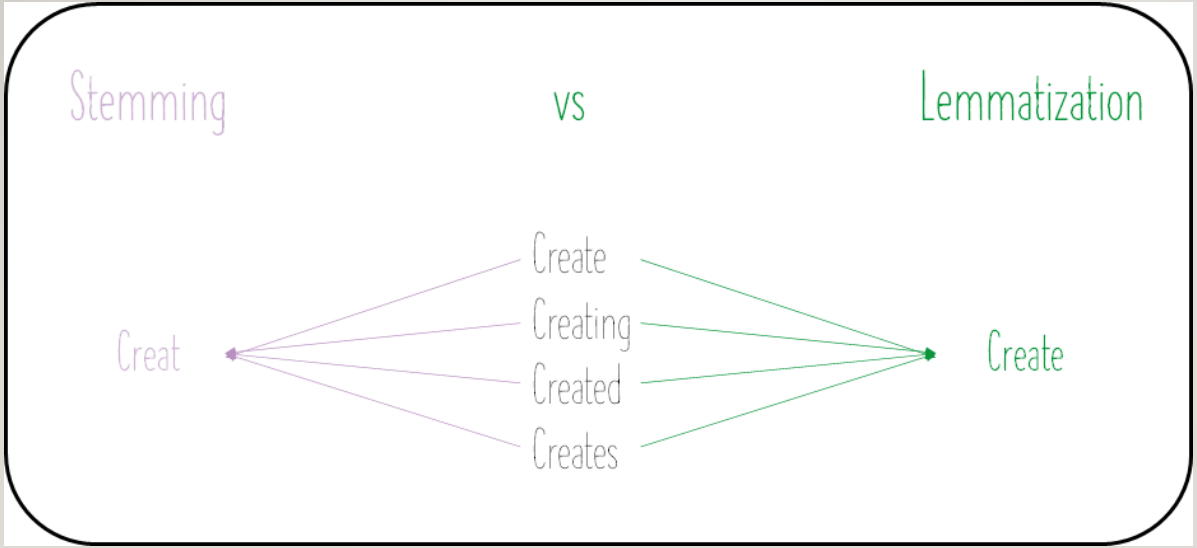
- Python Scripts
- xml.etree.ElementTree
- lxml

CLEANUP

- Special characters
- Bad characters
- Regular Expressions
- Dictionaries of good values



NORMALIZATION

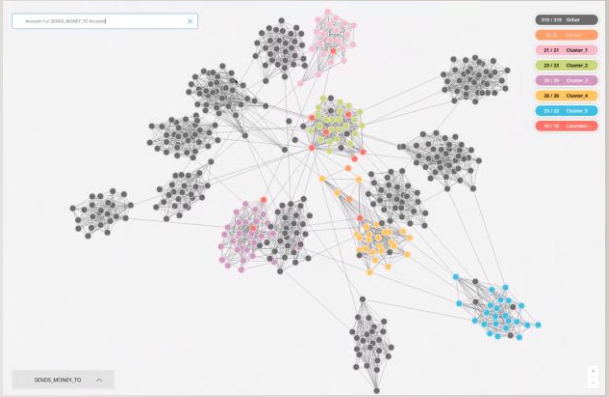


TAGGING, SENTENCE STRUCTURE,

- English Language has many good tagging options
- Latvian has one decent option:
- Latvian University http://nlp.ailab.lv/
- Stemming, lemmas
- Parts of Speech
- Named Entity (not too good)

INVESTIGATING STORAGE IN A GRAPH DATABASE

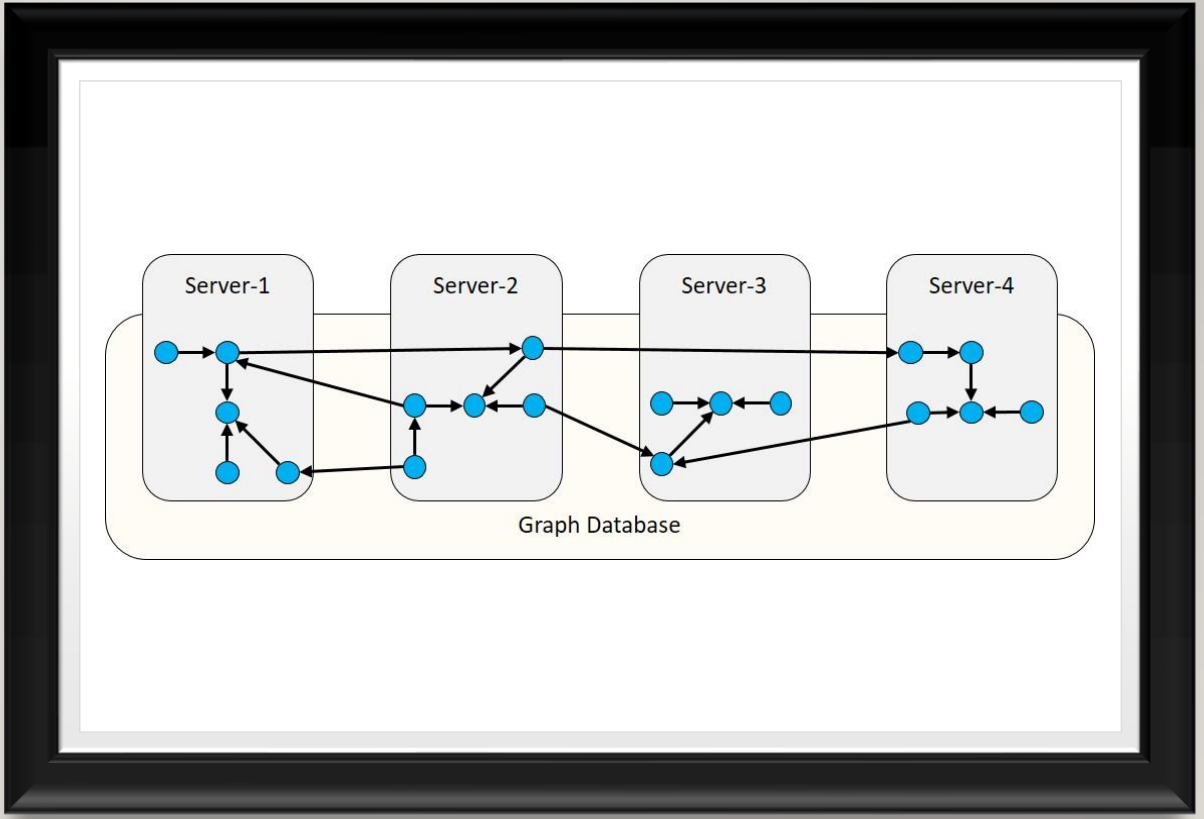
- Nodes would be documents
- Relationships – edges
- Multiples edges of similarity supported
- Nice visual and query tools



neo4j

EVALUATING GRAPH DATABASE SUITABILITY

DATABASE VENDORS LIKE TO EMBELLISH CAPABILITY

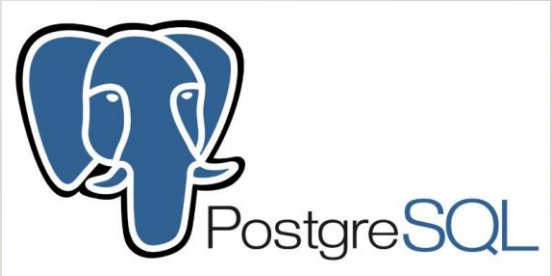


Source: Ojectivity, Inc

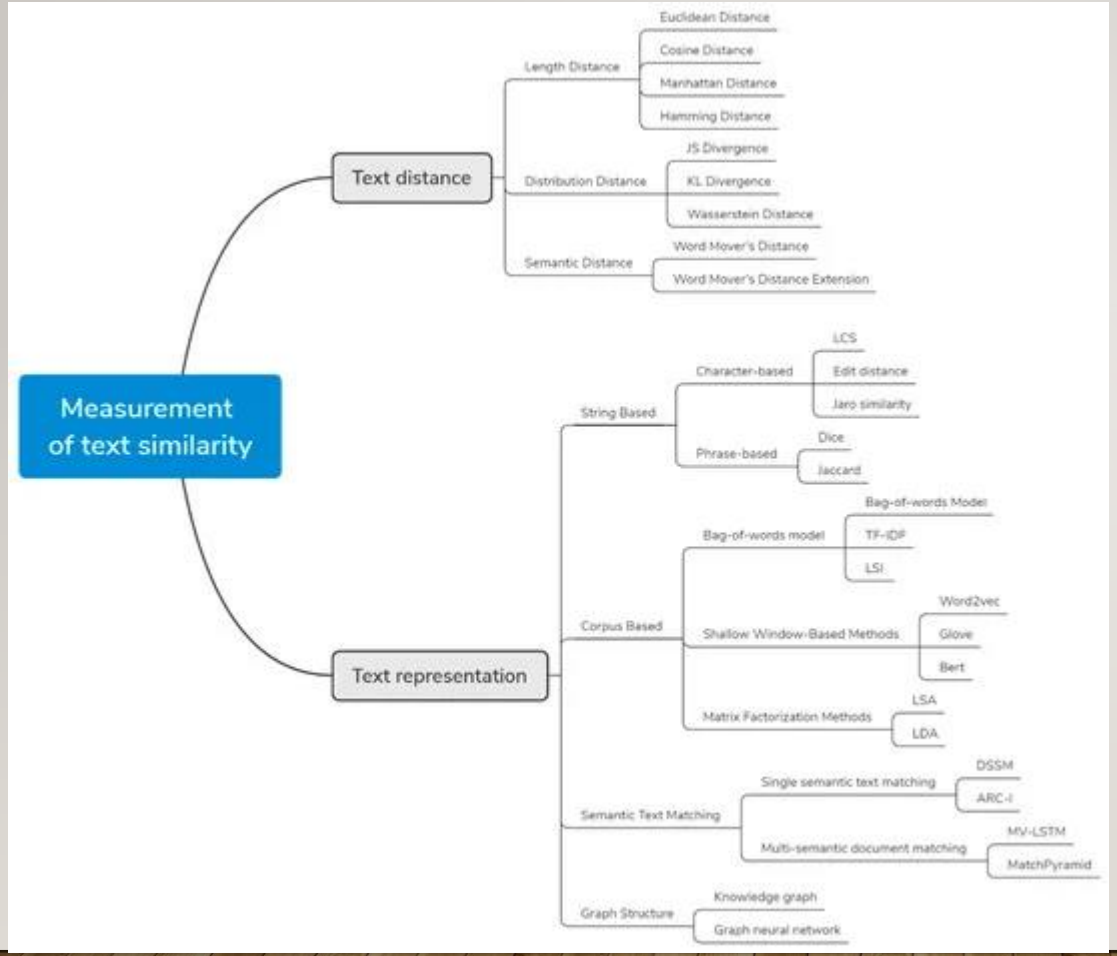
STORAGE/RETRIEVAL ALTERNATIVES

- Normal relational database - PostgreSQL, SQLite
- Custom built solution

- Challenges - proper relationship structure
- Indexing
- Batch(pre-compute) vs Real Time Processing



BIG PICTURE OF CONTENT SIMILARITY DETECTION



From: Measurement of Text Similarity: A Survey

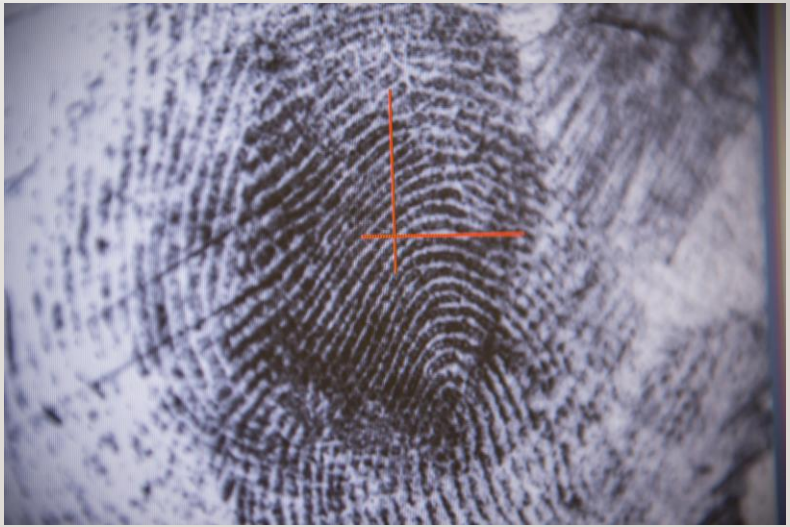
Jiapeng Wang and Yihong Dong

SIMILARITY DETECTION ALGORITHMS

- Basics
- Jaccard Similarity
- Cosine Similarity - (requires vectorization of sentences)
- Euclidian Similarity - also vectorization
- Tried TF/IDF, testing BERT - challenge encoding speed

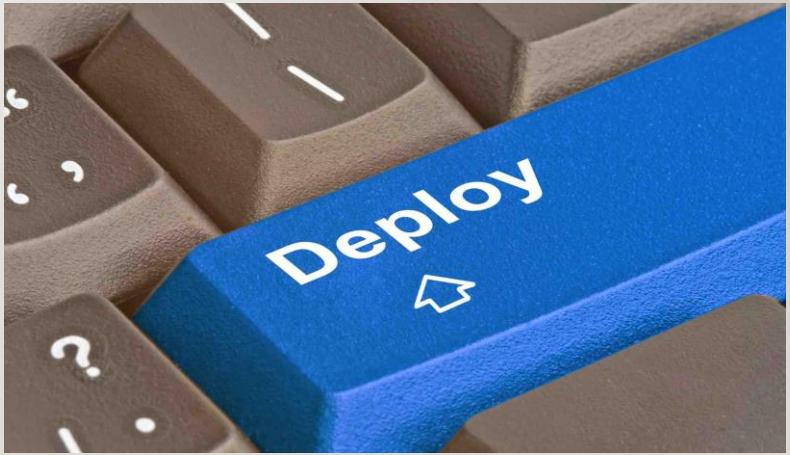
SIMILARITY DETECTION ALGORITHMS

- Fingerprinting - checking multiple ngrams - needs paring down to be effective
- Stylometry - promising but untested on our end



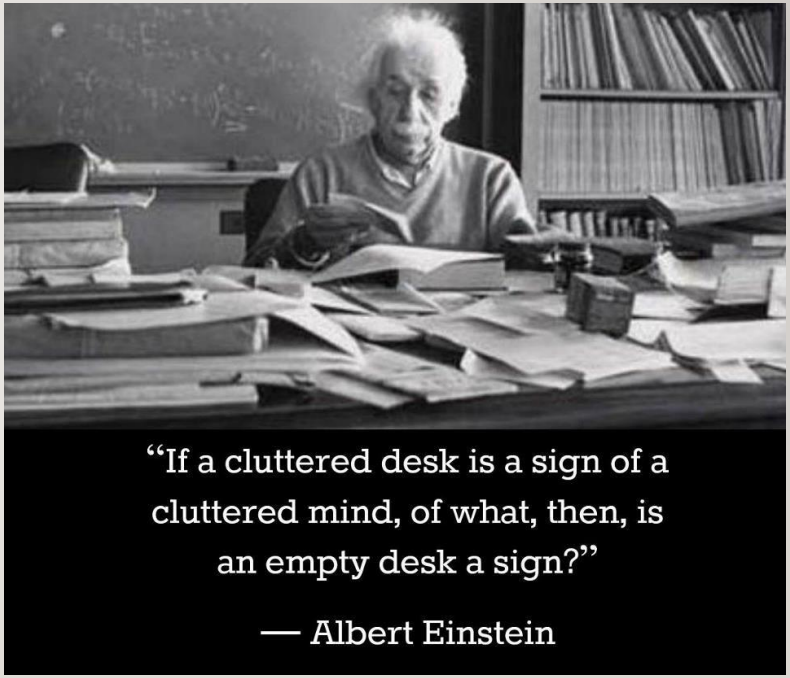
DEPLOYMENT GOALS AND CHALLENGES

- Support exploration not limited to few topics
- Python -> Scala rewrite
- Scalable to larger corpora / data sets
- Accessible to researchers



CONCLUSION

- DATA is MESSY – CLEAN IT
- USE PROVEN TECHNOLOGIES
- KISS - KEEP THINGS AS SIMPLE AS POSSIBLE BUT NO MORE SO



Anecdotical: https://quoteinvestigator.com/2017/09/02/clutter/