Information retrieval project: LiLAS ranking

Presented by:

- Abigail Hayes
- Seoyoung Yoo
- Simon Kral
- Sunggu Kang
- Tom Albrecht

Supervised by:

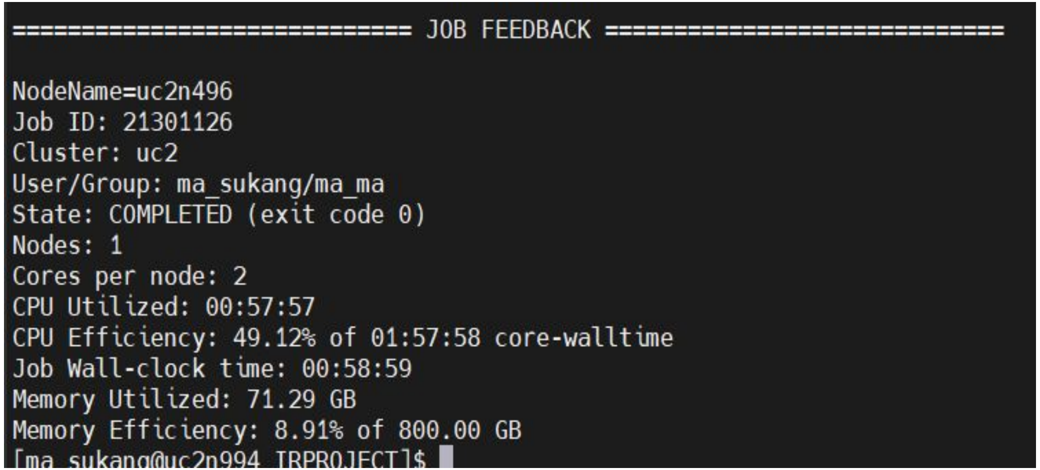
- Prof. Simone Ponzetto
- Dr. Pedro Ortiz Suarez

Our task

- Task had its specificities
- Not a classical retrieval problem as all the documents are relevant
- Rank 100 documents already returned for each of 100 queries
- Query most often just 1 word at most 4
- Queries, title and abstracts in a range of languages

Data extraction

- Problem: The files containing the data were too big - 10 files of ca. 5GB
- The files contained many documents not in the rankings
- Used the BWUniCluster 2.0 to extract the relevant documents for each query



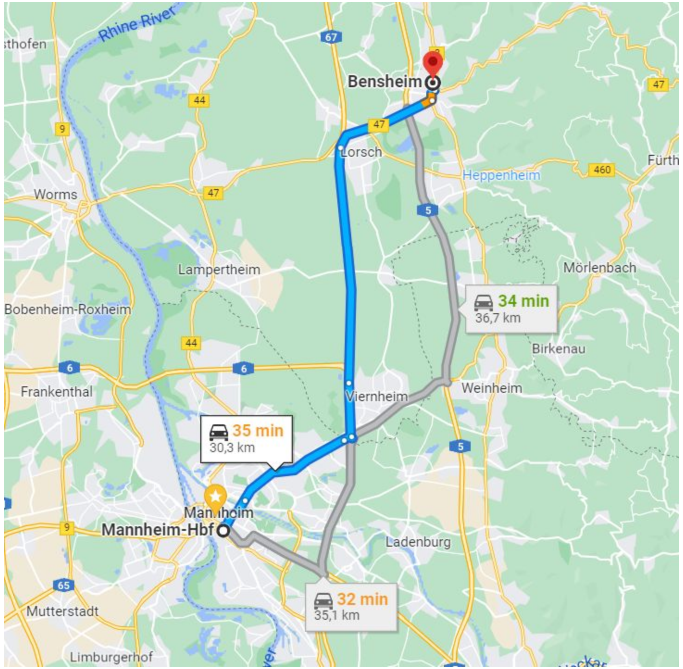
```
============================== JOB FEEDBACK ==============================

NodeName=uc2n496
Job ID: 21301126
Cluster: uc2
User/Group: ma_sukang/ma_ma
State: COMPLETED (exit code 0)
Nodes: 1
Cores per node: 2
CPU Utilized: 00:57:57
CPU Efficiency: 49.12% of 01:57:58 core-walltime
Job Wall-clock time: 00:58:59
Memory Utilized: 71.29 GB
Memory Efficiency: 8.91% of 800.00 GB
[ma_sukang@uc2n994 IRPROJECT]$
```

- BWUniCluster High performance computing

- Detail of extraction in the cluster

Data extraction- CO2 emission

- **5.8kg of CO2 was emitted**



30 km driving = 5.8kg of CO2

If we assume to run.. With Hyundai 2018 Santafe

- 1 day : Mannheim - Bensheim (30km)
- 1 week : Mannheim - Cologne (250km)
- 1 month : Mannheim - Birmingham, UK (985km)

Pre-processing

Starting point:

- Lower case
- Remove any ‘AND’ or ‘OR’ (split the query)

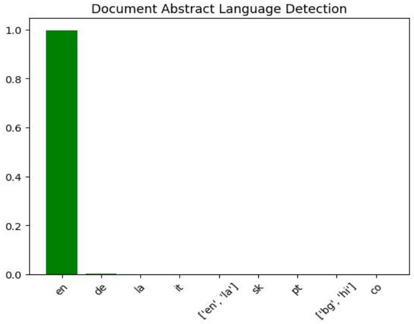
  e.g. mental AND health -> ['mental', 'health']
- Tokenization (no stemming, no stopword removal)

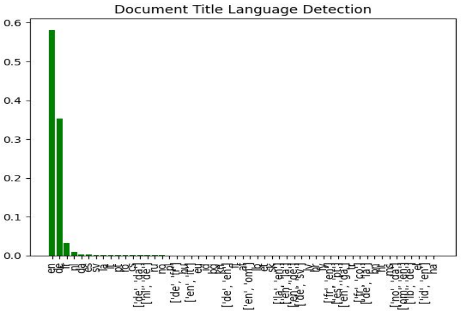
Other variants with an example (Query:mbsr)

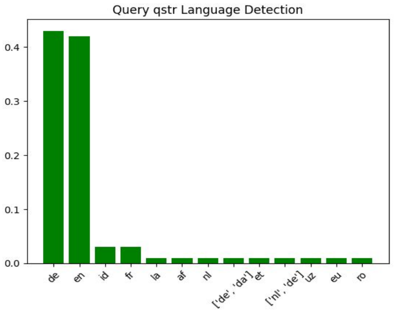
Original Abstract	No Stopwords, No stemming	No Stopwords, Stemming	Stopwords, No stemming	Stopwords, Stemming
["The purpose**#!** of this study was to characterize sympathetic activity…]	['purpose', 'study', '**characterize**', 'sympathetic', '**activity**', 'using' …]	['purpos', 'studi', '**character**', 'sympathet', '**activ**', 'use'...]	['**the**', 'purpose', 'of', '**this**', 'study', 'was', 'to', 'characterize' …]	['the', 'purpos', 'of', 'thi', 'studi', 'wa', 'to', 'character' …]

Translation

- Workflow :
  - Methodology 1. translate all document titles/abstracts into ‘en’ using Googletrans library
  - Methodology 2. translate document titles/abstracts into query’s detected language
- Result of Detection:
  - Document Abstract Language : 9 languages
  - Document Title Language : 34 languages
  - Query qstr Language : 12 languages
- Potential improvement :
  - Single word queries (e.g. Pandemie) don’t detect well







Model Selection and adaptation

- Chose BM25 as main model and VSM as a comparison
- Vector of 1-query terms very sparse → Probabilities are better (modelling uncertainty)
- Ranking based on abstracts
- Missing abstracts → Add small scores based on the title in those cases

BM25

- There were some problems with the BM25 computation:
- We chose formula with no relevance judgment since N = R and nt = rt

$$wt = \log(0.5 * N/Nt)$$

- Log got negative for some terms, because 0.5*(N/Nt) got smaller than 1
  ➔ calculate without the log

VSM

- Challenge: many 1 word queries
- Implemented solution: Differentiation of cases
  - multiple word query ➔ cosine similarity
  - one word query ➔ length of tf-term

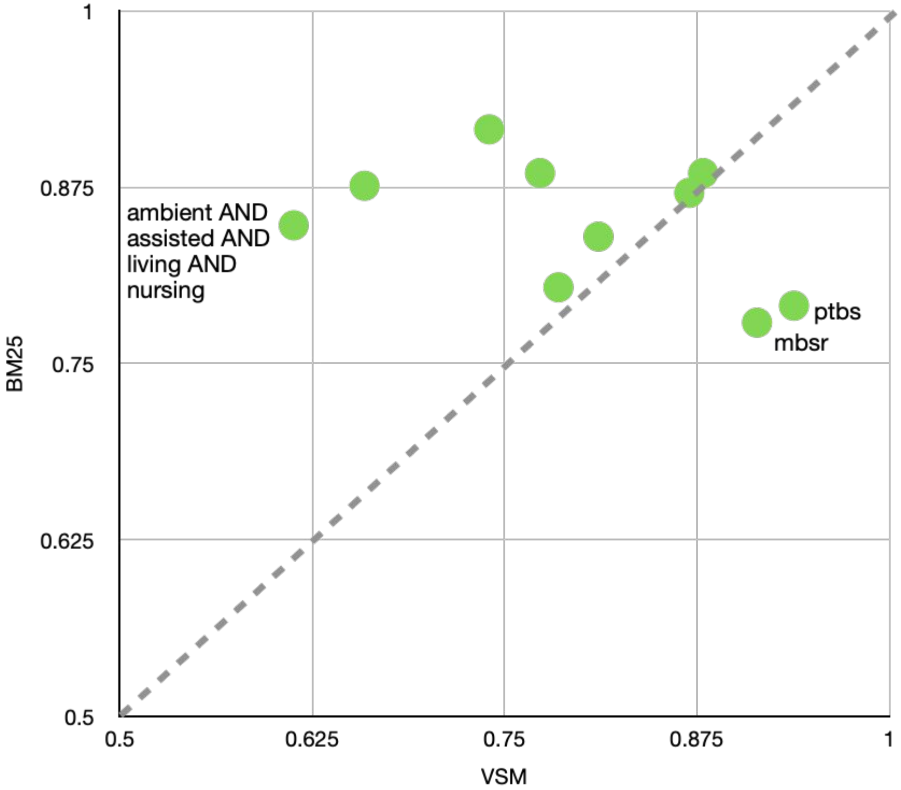
Evaluation

- Planned to use the STELLA evaluation
- Instead used nDCG (normalised Discounted Cumulative Gain)
- Used documents pooled from the results of BM25 and VSM
- Limitations given all docs are already judged as relevant

	No stemming	No stemming	Stemming	Stemming
	No stopwords	Stopwords	No stopwords	Stopwords
Raw frequency	0.728	0.728	0.709	0.709
BM25	0.846	0.850	0.836	0.834
VSM	0.794	0.798	0.758	0.757

Evaluation

- No stemming
- No stopword removal



Further development

- Interpretation of abbreviations
- Support for phrase queries
- Author’s name as a search term
- Quality of translation ➔ Multiple translation variants

THANK
YOU

Sources:

- Lecture slides
- Usage of BwUnicluster:
  - https://wiki.bwhpc.de/e/BwUniCluster2.0
- CO2 emission calculation:
  - https://www.carbonfootprint.com/calculator.aspx
  - https://ark.intel.com/content/www/us/en/ark/products/192437/intel-xeon-gold-6230-processor-27-5m-cache-2-10-ghz.html)
  - https://dl.acm.org/doi/10.1145/2989081.2989088
  - https://www.statista.com/statistics/1229367/data-center-average-annual-pue-worldwide/
  - https://www.enbw.com/unternehmen/nachhaltigkeit/environment/umweltschutz/co-fussabdruck.html#spezifische-co-emissionen
  - https://aclanthology.org/P19-1355/
- Googletrans Library
  - https://pypi.org/project/googletrans/