Studying Retrievability in IR

MS Project

by

Aman Sinha
18MS065

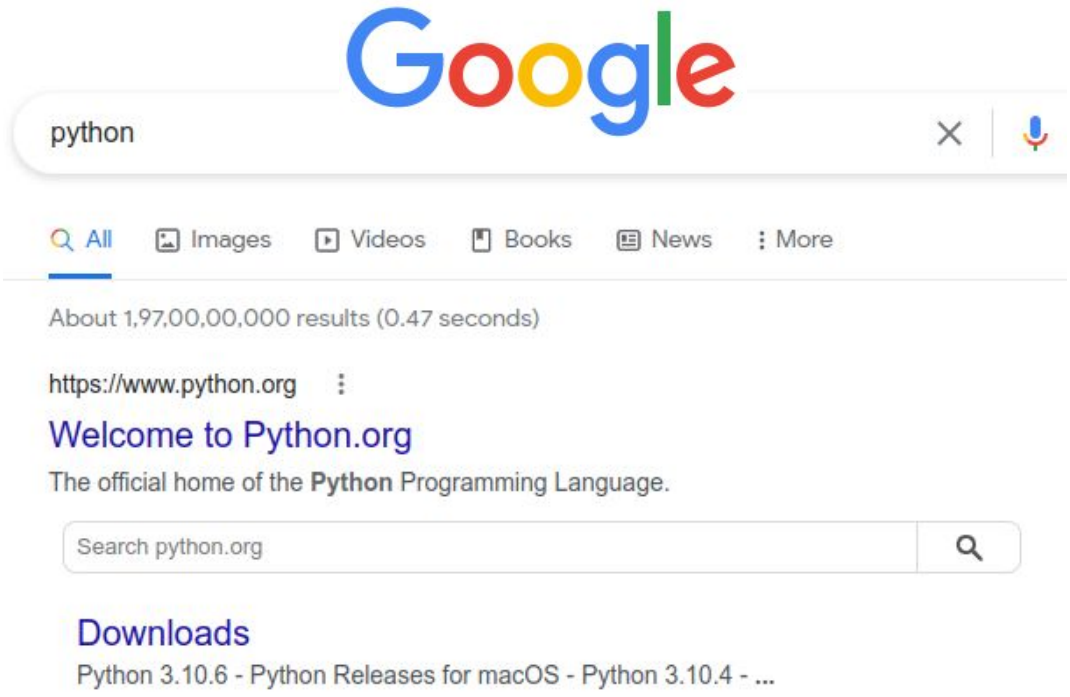
Supervisor:
Dr. Dwaipayan Roy
Department of Computational and Data Sciences
IISER Kolkata

DPS Coordinator:
Prof. Rangeet Bhattacharyya
Department of Physical Sciences
IISER Kolkata

Overview

1. Motivation
2. What is Retrievability? And Retrievability bias?
3. Lorenz curve and Gini coefficient
4. Retrievability for 3 search techniques
5. Retrievability bias in labelled test data?
6. Retrievability bias in Query Expansion
7. Conclusion & Plan for the next semester

Motivation



Favoritism in Search results ⇨ Bias !!!

Motivation

- Considering these biases, some websites are preferred more by a search engine than others - “**Retrievability**” of websites
- **Retrievability** measure - a way of measuring these biases
- Can we use retrievability to improve the quality of search engine results?

Measure of Retrievability

Measure of retrievability of a document **d** is,

$$r(\mathbf{d}) = \sum_{\mathbf{q} \in \mathbf{Q}} f(k_{dq}, c)$$

r(d) = how many times a document d
is retrieved by the search system
within the top rank (say within top 10)
for a large no. of queries

Azzopardi, Leif, and Vishwa Vinay. "Retrievability: An evaluation measure for higher order information access tasks." In Proceedings of the 17th ACM conference on Information and knowledge management, pp. 561-570. 2008.

Retrievability Analysis Framework

5 key steps :

1. Document collection & searching technique selection
2. Query set construction
3. Searching with all the queries
4. Computing document retrievability r(d)
5. Studying the inequalities between r(d) of documents

Retrievability Experiment on TREC 678 collection

TREC - **T**ext **Re**trieval **C**onference

Conference contributes to research in search engine effectiveness

Text REtrieval Conference logo

Publishes large collections of documents for research

Sample search queries and relevant, non-relevant labels for docs for these queries

TREC 678 is one such collection of documents

- ➔ 528,155 documents
- ➔ ~ 2 GB size
- ➔ ~ 1.5 M unique words

Query Set for Retrievability Experiment

2 approaches to construct query set :

1. Real query log
    1.1. Google search queries, Bing search queries
2. Artificial auto-generated queries
    2.1. Sampling chunks of words from documents and posing them as queries

Query Generation Method

Query set generated comprise of two subsets:

1. Single word search queries
2. Two word search queries

Use of one word and two words search queries motivated by work of Azzopardi and Vinay (2008)

For one word queries

1. Vocabulary words
2. Cleaning
3. Selecting Nouns only
4. Frequency-based filtering

For two word queries

1. Two consecutive words drawn from sentences
2. Cleaning
3. Selecting Nouns only
4. Frequency-based filtering

Query Set: Queries

No. of Unigram queries = 137,029 ~ 0.137 M

No. of Bigram queries = 447,183 ~ 0.447 M

Total no. of queries = 584,212 ~ 0.584 M

director
sales
area
price
months
agency
shares
law
staff
money
prices
tax
issue
secretary
chairman
document

work
number
commission
program
week
edition
today
interest
county
order
security
department
investment
management
day
committee

financial times
london page
united states
daily report
last year
los angeles
united kingdom
kingdom ec
home edition
prime minister
new york
article type
document type
orange county
type bfn
company news

monetary policy
high school
sports desk
soviet union
human rights
real estate
federal register
russian federation
final rule
business part
sunday home
central bank
stock exchange
united nations
security council
stock market

Searching our queries using 3 Search techniques

Popular searching techniques:

1. TF-IDF
2. BM25
3. LM-Dir

Retrievals on
~0.6M queries
took
~ **21 hrs** of
computational time

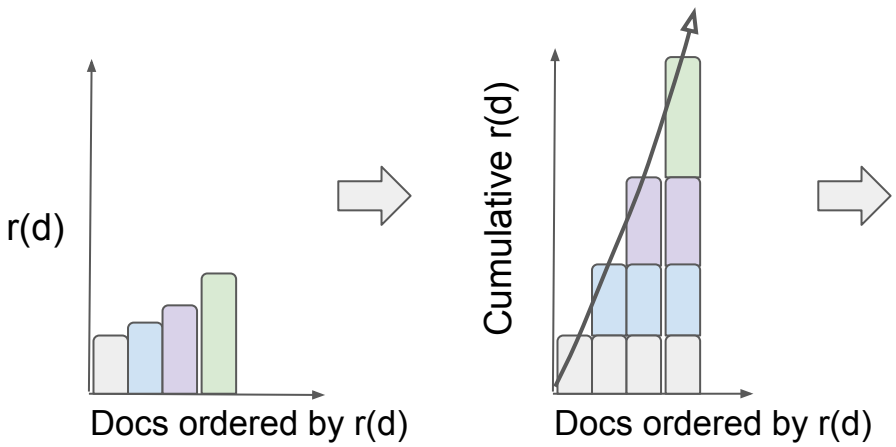
Retrievability score **r(d)**
estimated for all 528,155 documents

Acknowledgment: Experiment codes were run on Dr. Dwaipayan sir's office desktop with 16GB RAM and 12th Gen i7 processor with 20 cores

Lorenz Curve

In Economics used for representing inequality of wealth distribution

Developed by Max O. Lorenz in 1905



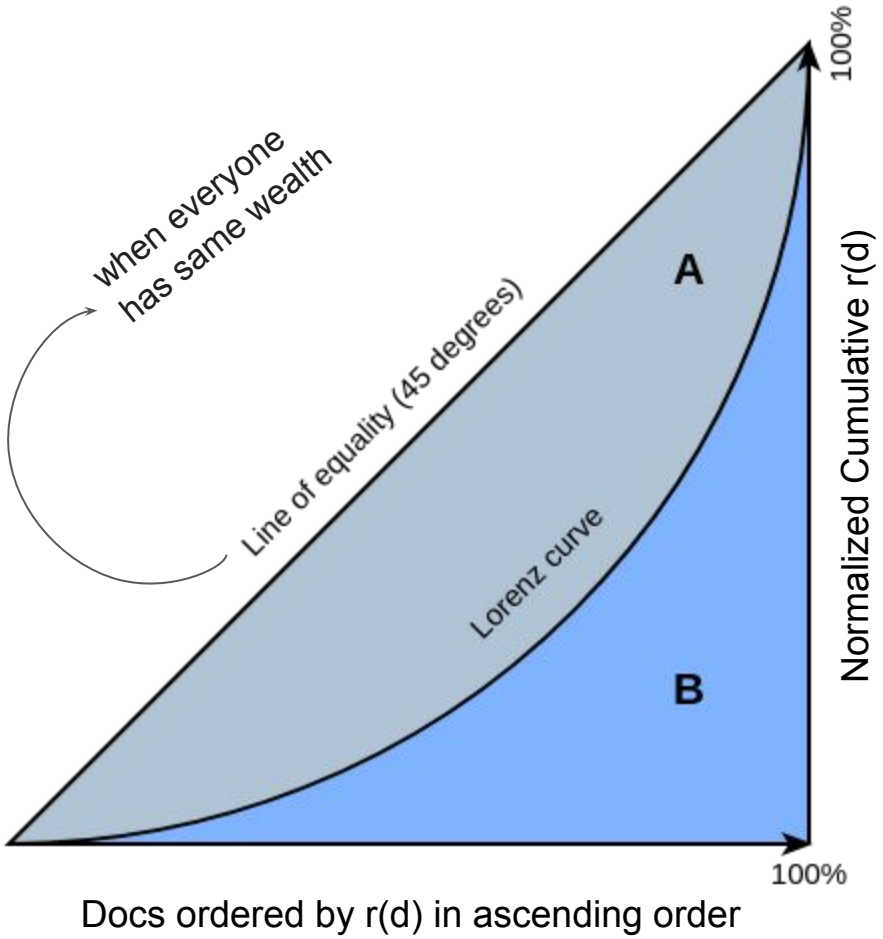


Image Source: Wikipedia. 2022. "Lorenz curve." Wikimedia Foundation. Last modified October 23, 2022. https://en.wikipedia.org/wiki/Lorenz_curve

Gini Coefficient G

$$G = \frac{\sum_{i=1}^{N} (2 * i - N - 1) * r(\mathbf{d_i})}{N \sum_{j=1}^{N} r(\mathbf{d_j})}$$

where,

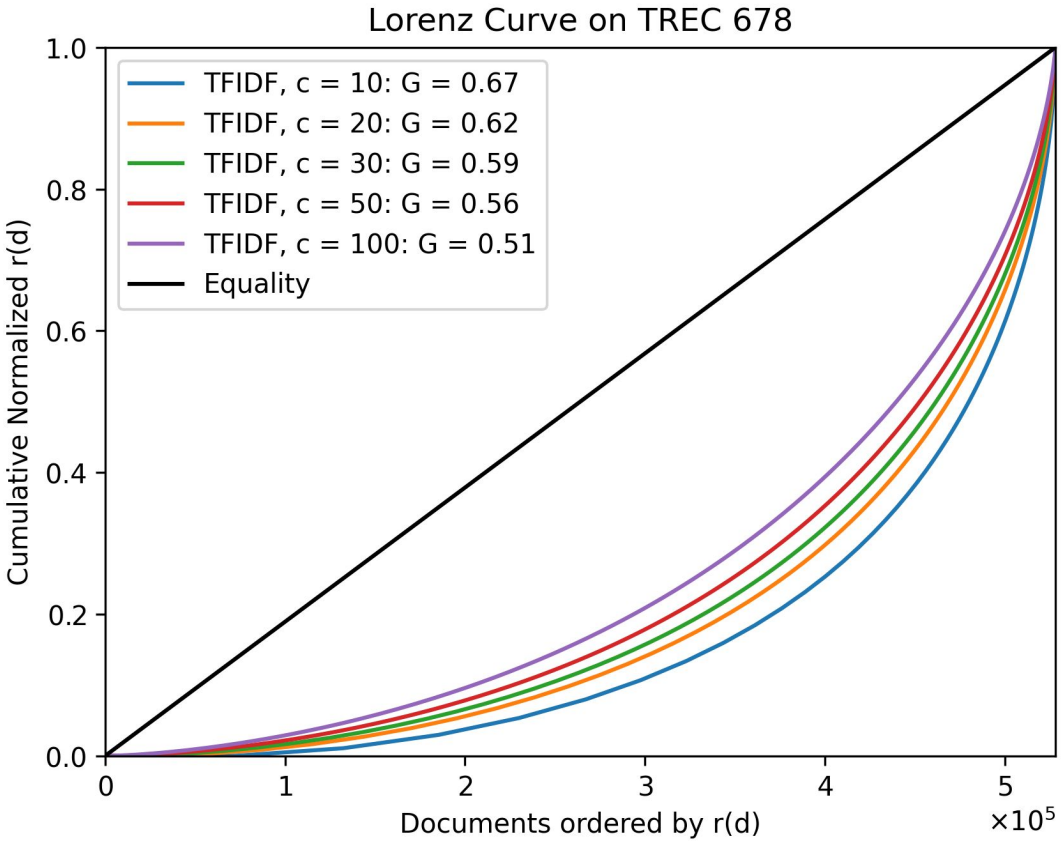
$r(d_i)$ is in ascending order

N is the total number of documents in the collection

G = 0 : absolute equality, no bias

G = 1 : absolute inequality, maximum bias

Lorenz Curve for TF-IDF model

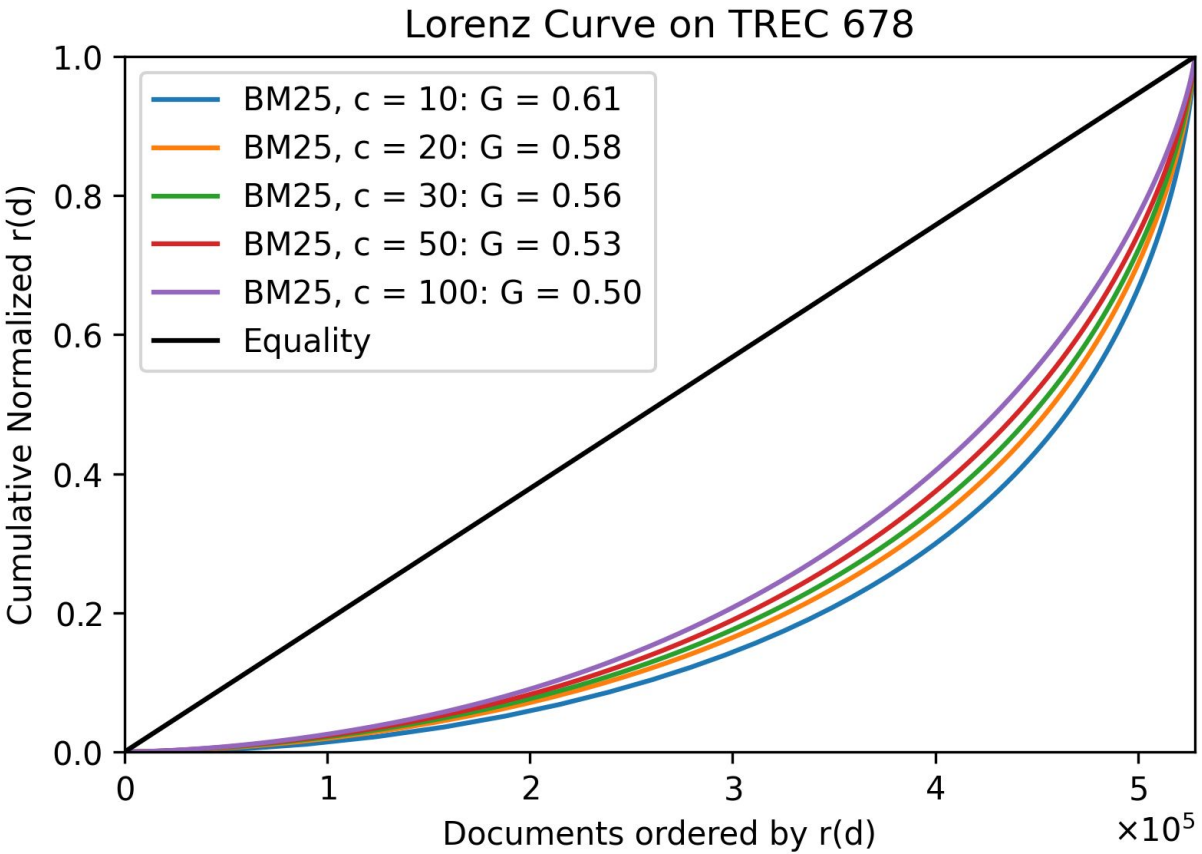


Observation

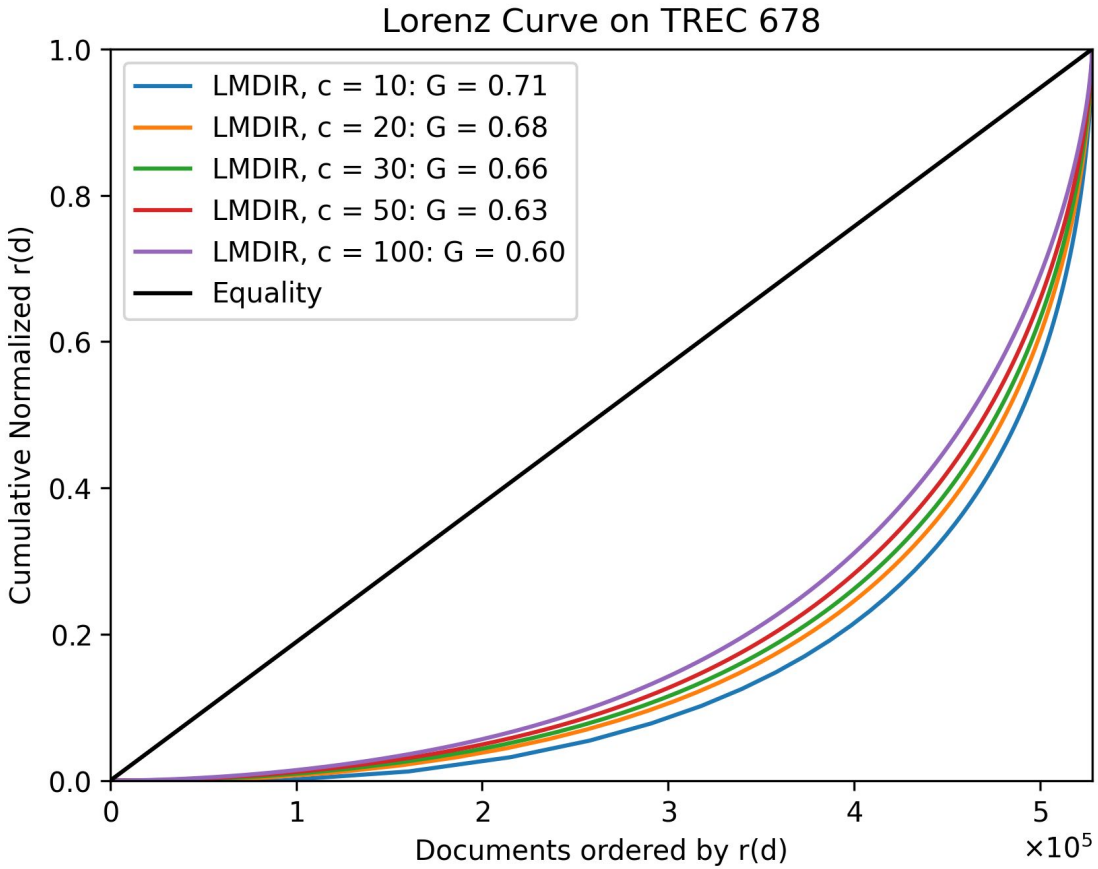
Gini coefficient G is decreasing as the rank cutoff c is increasing.

Suggesting that if explore further down the search results, the lesser we are exposed to document favoritism

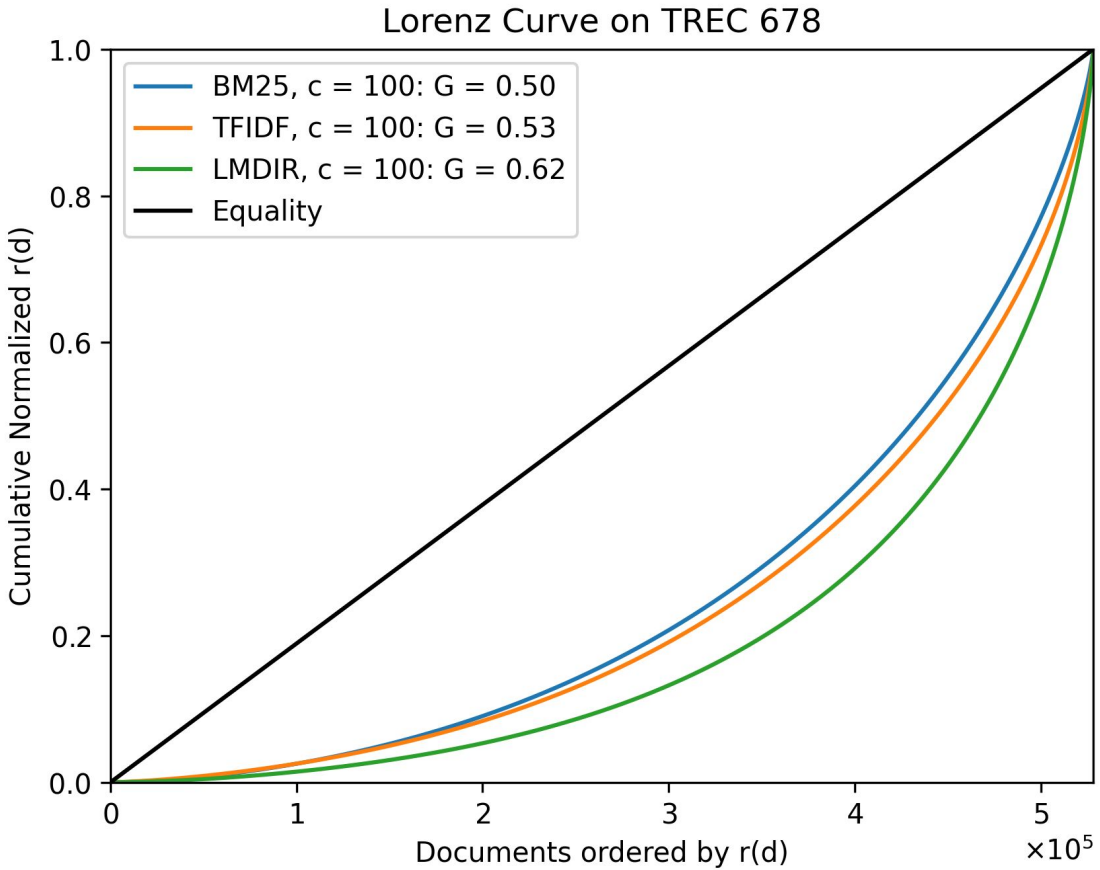
Lorenz Curve for BM25 model



Lorenz Curve for LMDir model



Lorenz Curve for all 3 searching techniques for c = 100

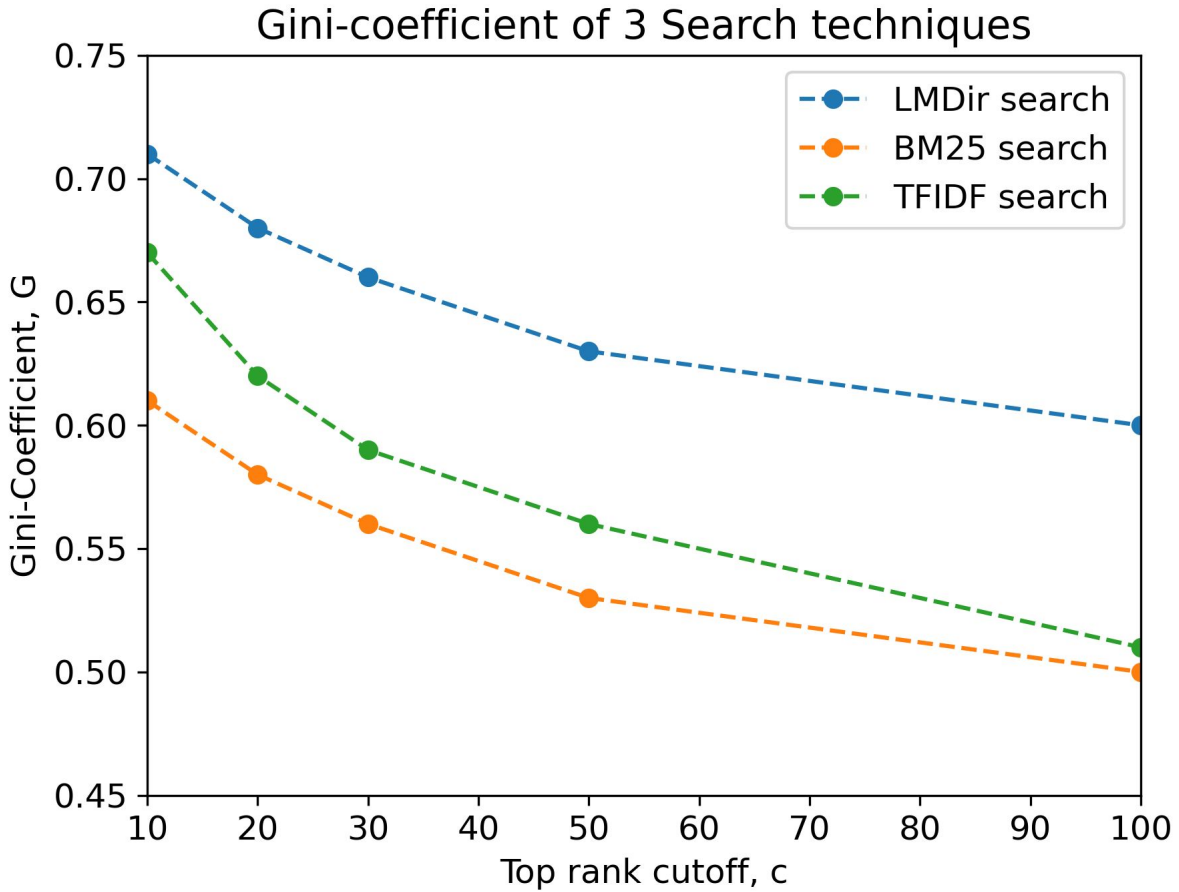


Observation

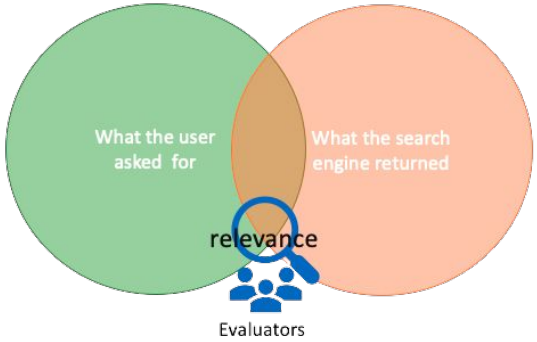
Bias is least for BM25 and maximum for LMDir among the 3 retrieval models explored

Even for least biased model here BM25, Gini coefficient is considerably high

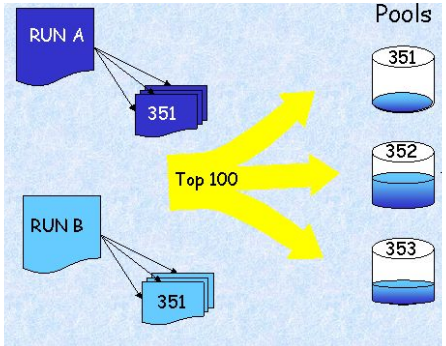
Results Plot



Relevance Judgements by TREC



- To access search result quality and effectiveness, TREC has an evaluation scheme
- Some sample queries,
  and their relevant and non-relevant documents are labelled by humans → “Relevance Judgement” of those queries
- Not all documents are labelled
- Which ones are selected to be label?
  - Top Results from all good search engines



Retrievability of TREC Judged vs Non-judged documents

	Judged documents	Non-judged documents
Count	174787	353368
Mean r(d)	**131.03**	**80.15**
Min r(d)	0.0	0.0
Max r(d)	3220	1534

Judged documents more retrievable !!

Query Expansion (QE)

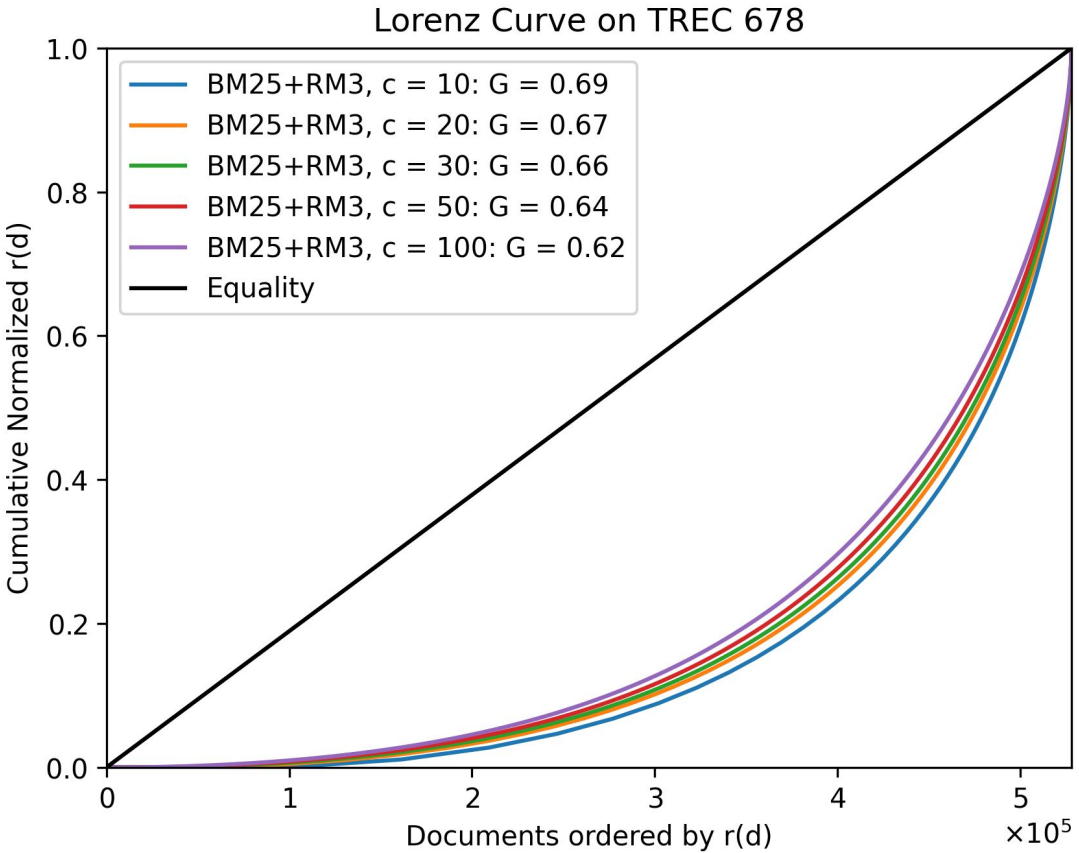
Process of adding extra terms to the original query from deemed relevant docs to improve search results

One such very good QE method is **RM3**

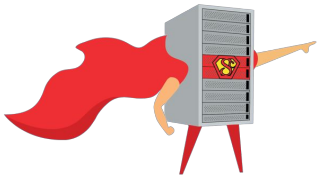
Retrievability Experiment for
BM25 search technique with
RM3 query expansion method

Pseudo-Relevant docs = 10 ; Expansion terms = 10 ; Original query weight = 0.4

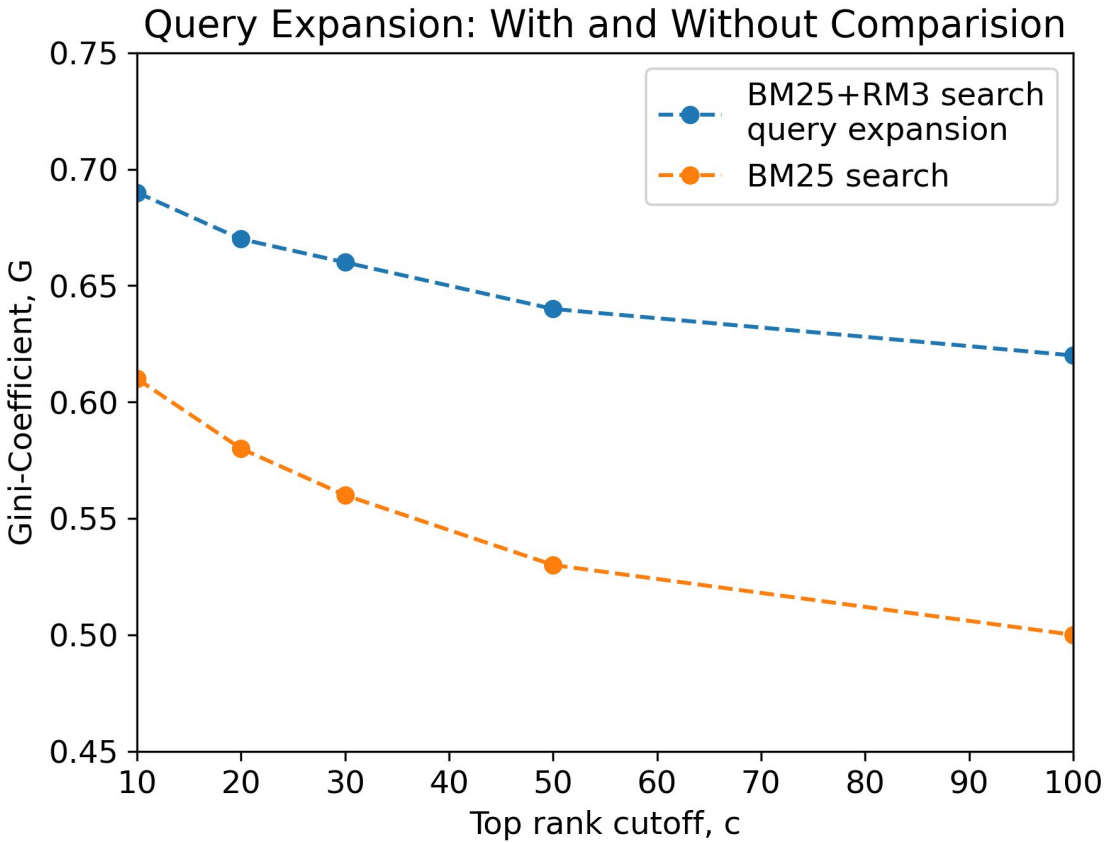
Lorenz Curve for BM25+RM3 Query Expansion



Ran the experiment code on Dirac Supercomputer for ~ 2 days



Comparison Plot



Observation

Query expansion technique increased bias in comparison to the initial search results
(even though RM3 is known to increase search effectiveness)

Conclusion

- ❖ Retrievability analysis estimated bias of different searching techniques
- ❖ BM25 provided best accessibility and is also known for strong performance
- ❖ Judged documents can be biased towards highly retrievable documents
- ❖ RM3 along with boosting performance is increasing inequality in retrievability

Next Semester

- PageRank and Retrievability correlation for Wikipedia dataset
- Exploration of Retrievability in increasing search effectiveness

Thank You!