Enhancing Search Engine Performance with Character N-Grams, Query Expansion and Named Entity Recognition

TA$K: LongEval CLEF 2023 Lab

Team JIHUMING@UNIPD

Isil Atabek, Huimin Chen, Jesús Moncada-Ramírez

Nicolò Santini, Giovanni Zago

Agenda |

- Introduction
- Methodology
- System Architecture
- Experimental Setup
- Results and Discussion
- Conclusion | Future work

Our Team |

Jesús Moncada-Ramírez

Giovanni Zago

Huimin Chen

Nicolò Santini

Isil Atabek

Introduction |

We introduce a search engine for LongEval at CLEF 2023. Our system focuses on temporal performance in English and French documents.

By analyzing text and using NLP techniques, we refine our system. Implemented in Java with Lucene, we developed five top-performing systems based on MAP and NDCG scores.

Introduction |

Our approach involves analyzing English and French versions of the documents using whitespace tokenization, stopword removal and stemming.

We generate character N-grams to identify recurring word structures repeated over documents.

We use query expansion with synonyms and NLP techniques as NER to further refine our system.

Methodology | Parsing

Documents

- JSON version.
- *Iterator/Iterable* architecture.

Topics

- *TRECTopicReader* not working.
- Our **own** parser treating topic documents as **XML** documents.

Methodology | Index

Always (BM25)

FIELD 1	FIELD 2	FIELD 3	FIELD 4
(processed) English version	(processed) French version	Character **N-grams** of both versions	**NER** information
		*3-grams, 4-grams, 5-grams*	*Apache OpenNLP*

Methodology | English

ENGLISH PROCESSING (ANALYZER)

Whitespace tokenization

Breaking based on special characters

WordDelimiterGraphFilter

Lowercasing

TERRIER stopword list

Query expansion with synonyms

WORDNET

Stemming

English Minimal Stemmer

Methodology | French

FRENCH PROCESSING (ANALYZER)

Whitespace tokenization

Breaking based on special characters

WordDelimiterGraphFilter

Lowercasing

French stopword list

Stemming

French Minimal Stemmer

Methodology | N-grams and NER

N-GRAM GENERATION

- Delete all characters **except letters.**
- Generate char N-GRAMs.

Top ranking terms: (Double-click for more options.)

Rank	Freq	Text
1	1440738	ation
2	1267022	ement
3	1234752	tions
4	1172599	ction
5	957077	ition

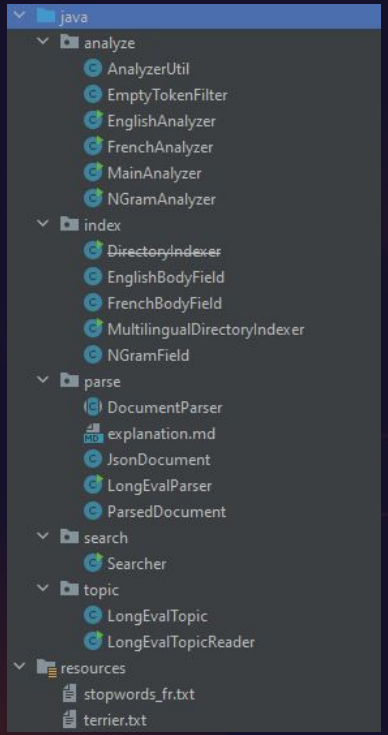
NER GENERATION

- Use the **FRENCH** documents.
- NER about locations, person names and organizations.

System Architecture |

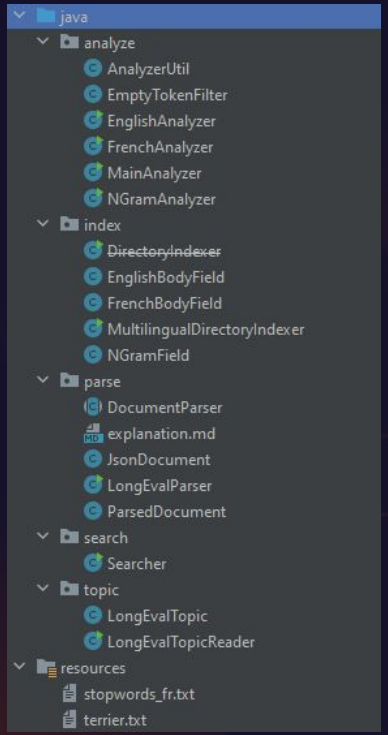
- LongEval data structure (queries and documents), **Document Parser** and **Topic Reader**
- **Analysis** techniques (tokenization, NER, N-Gram)
- **Index**
- **Search**

System Architecture |



- Package division
- Resource files (English and French stopwords)

System Architecture |



REMARKABLE CLASSES

Parse:

- Reading documents in JSON format, in order to convert them in Java classes;
- LongEvalParser is the class that implements DocumentParser (Iterator) and parses all documents.

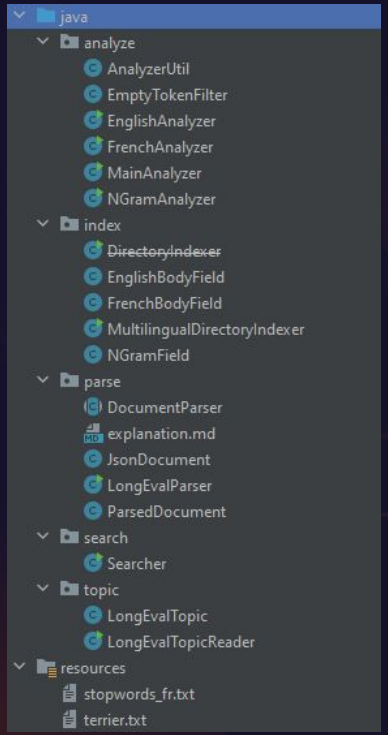
Analyzer:

- Applies the tokenization rules for both English and French documents (Lucene methods and stopwords files);
- Generate character N-grams (Lucene).

Index:

- Initially made by **DirectoryIndexer**, which allowed only to index document with a certain language;
- **MultilingualDirectoryIndexer** reads both English and French documents, indexing all documents;
- Vocabulary statistics.

System Architecture |



REMARKABLE CLASSES

Topic:

- Impossible to use TrecTopicsReader (Lucene) because the format of trec files was too poor;
- Defines a Java class representing a topic;
- Considers the .trec files as XML, parsing them using Java XML library.

Search:

- Do the effective search (specifying analyzers, path of topics, path of index, …)

Experimental Setup | Overview

Goal

Generate **multilingual** indexes

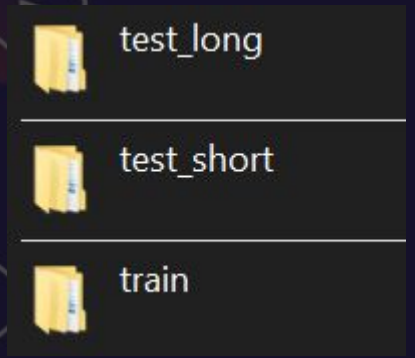
Evaluation:

1. MAP, NDCG and Rprec scores.
2. Two-Ways ANOVA analysis.
3. Tukey Honestly Significant Difference (HSD) test.

Experimental Setup | Indexes

We created indexes for **TRAINING** and **TEST** collections.

- **<date>_multilingual_3gram**
- **<date>_multilingual_3gram_synonym**
- **<date>_multilingual_4gram_synonym**
- **<date>_multilingual_5gram_synonym**
- **<date>_multilingual_4gram_synonym_ner**



All of them are public in Google Drive.

Experimental Setup | Runs

We made **12 experiments** over different configurations.

- *seupd2223-JIHUMING-01_en_en*
- *seupd2223-JIHUMING-02_en_en_3gram*
- *seupd2223-JIHUMING-03_en_en_4gram*
- *seupd2223-JIHUMING-04_en_en_5gram*
- *seupd2223-JIHUMING-05_en_en_fr_5gram*
- *seupd2223-JIHUMING-06_en_en_4gram_ner*
- *seupd2223-JIHUMING-07_fr_fr*
- *seupd2223-JIHUMING-08_fr_fr_3gram*
- *seupd2223-JIHUMING-09_fr_fr_4gram*
- *seupd2223-JIHUMING-10_fr_fr_5gram*
- *seupd2223-JIHUMING-11_fr_en_fr_5gram*
- *seupd2223-JIHUMING-12_fr_fr_4gram_ner*

Results and Discussion | Train

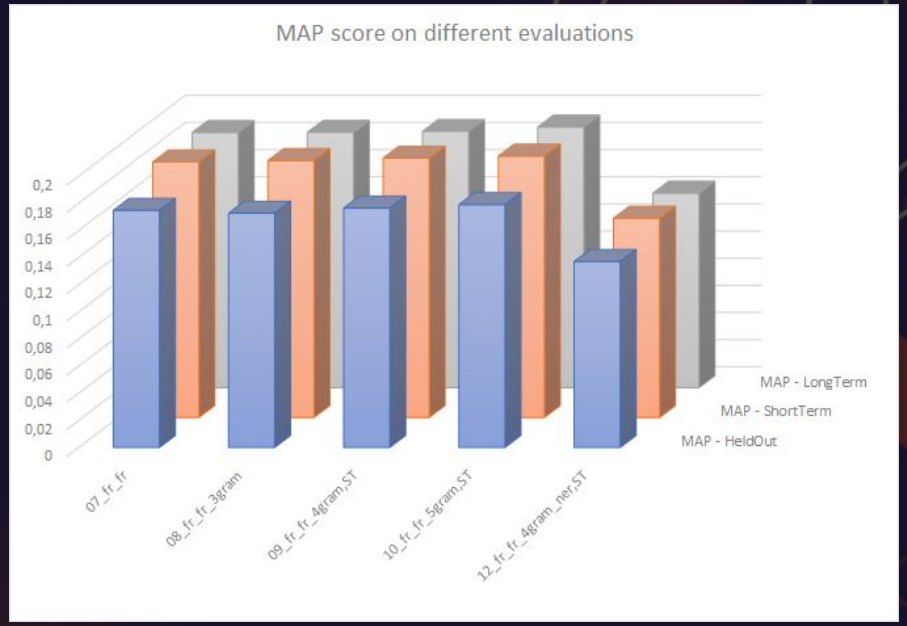
Index	Run	MAP Score	NCDG Score
01	en_en	0.0700	0.1614
02	en_en_3gram	0.0704	0.1661
03	en_en_4gram	0.0874	0.2025
04	en_en_5gram	0.1028	0.2288
05	en_en_fr_5gram	0.0669	0.1525
06	en_en_4gram_ner	0.0360	0.1098
07	fr_fr	0.1656	0.3135
08	fr_fr_3gram	0.1698	0.3208
09	fr_fr_4gram	0.1737	0.3269
10	fr_fr_5gram	0.1748	0.3285
11	fr_en_fr_5gram	0.1288	0.2797
12	fr_fr_4gram_ner	0.1362	0.2881

- Five best systems with five best scores
  - Fr_fr_5gram
  - Fr_fr_4gram
  - Fr_fr_3gram
  - Fr_fr
  - Fr_fr_4gram_ner

Results and Discussion |

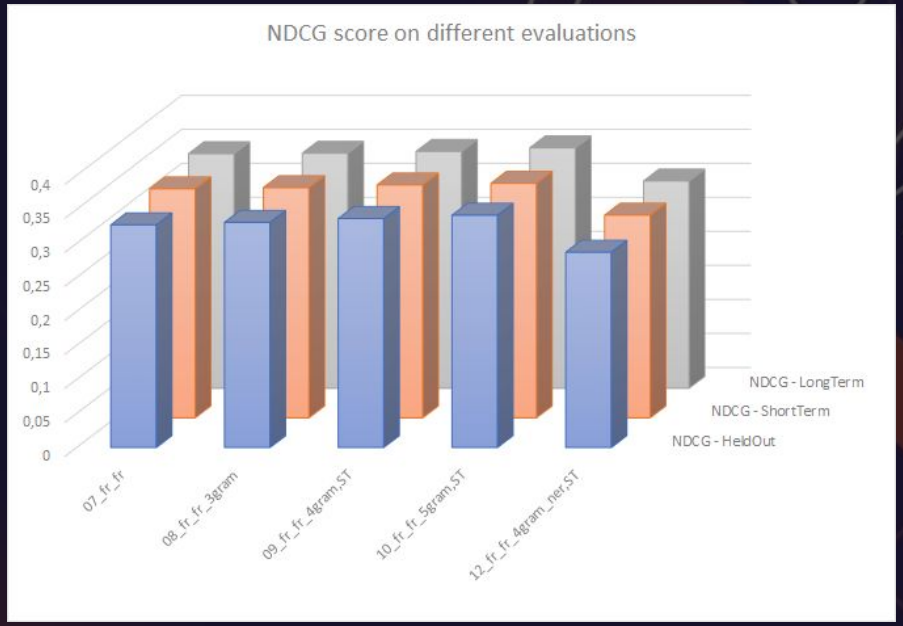
- **French queries** perform better than their English counterparts.
- IR system’s effectiveness generally increases with a **larger N-gram size.**
- The inclusion of NER in the indexing process has a **negative impact** on the scores.

Results and Discussion | Test



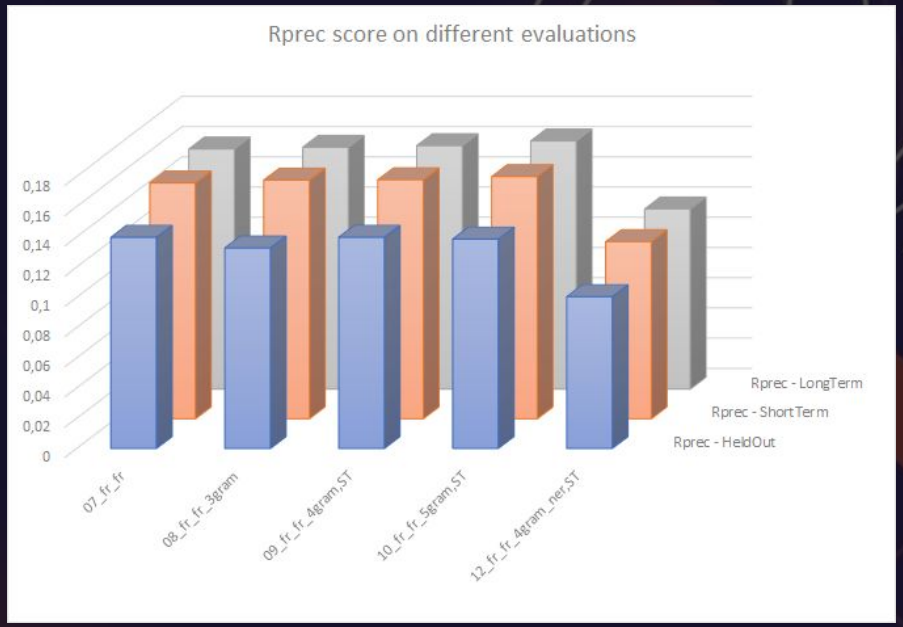
- Mean Average Precision (MAP) evaluating the effectiveness of an IR system in ranking documents/items.
- Indicating same score ranks in three data sets.
- Demonstrate the best performance at the long-term.

Results and Discussion | Test



- nDCG (normalized Discounted Cumulative Gain) assesses the quality of the ranking produced by an IR system.
- Indicating same score ranks in three data sets.
- Demonstrate the best performance at the long-term.

Results and Discussion | Test



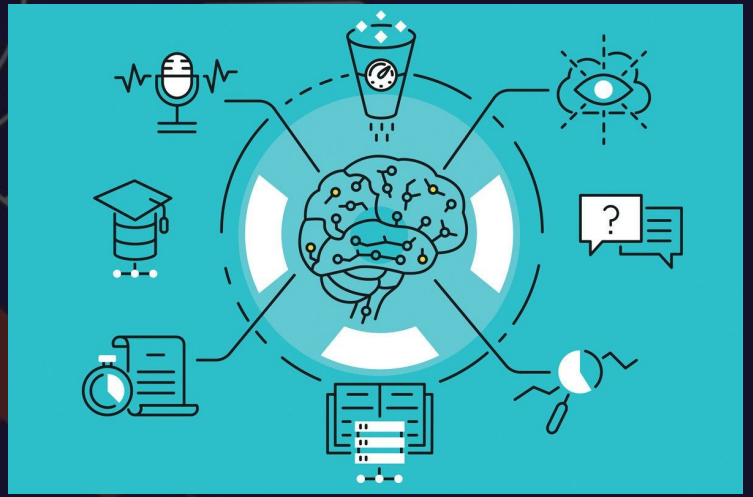
- Rprec (Rank Precision) measures the precision of the retrieved documents/items
- Indicating same score ranks in three data sets.
- Demonstrate the best performance at the long-term.

Results and Discussion |

- From the training data and test data, the three metrics shows the same rankings of five systems' effectiveness.
- The **long-term data** presents the best performance score.

Conclusion | Future work

Our results (MAP, NCDG, ANOVA etc) reflect our approach: many simple approaches, as if we were searching for a heuristic start of a more complex information retrieval. The only untapped approach would be including a machine-learning based IR system, to achieve a more adaptable one.



THANK YOU

Team JIHUMING@UNIPD

Isil Atabek

Huimin Chen

Jesús Moncada-Ramírez

Nicolò Santini

Giovanni Zago