SEUPD@CLEF Task 1: Information retrieval for English and French documents: Team JIHUMING

Jesús Moncada Ramírez[1], Isil Atabek[1], Huimin Chen[1], Michele Canale[1], Nicolò Santini[1] and Giovanni Zago[1]

Abstract

Our group will propose an original and efficient information retrieval system for Longitudinal Evaluation of Model Performance (LongEval) by CLEF2023[1]. Focus is on short term and long term temporal persistence of the systems' performance, for both English and French documents. The aim is to find a model giving good results for longitudinal evolving benchmarks, for the subject Search Engines, University of Padova.

Keywords

CLEF 2023, Information retrieval, LongEval, English, French, Search Engines

1. Introduction

This report aims at providing a brief explanation of the Information Retrieval system built as a team project during the Search Engine course 22/23 of the master's degree in Computer Engineering and Data Science at University of Padua, Italy. Task chosen by the group is CLEF LongEval: Longitudinal Evaluation of Model Performance.

Longeval Websearch collection[2] relies on a large set of data provided by a commercial Qwant (a commercial search engine). The idea is to reflect changes of the Web across time, providing evolving document and query sets.

Our approach uses on N-grams, trying to avoid the problem of the sparsity of the data. Our focus was not to produce a high scoring specialized system, but to try developing general ideas in order to approach sentences selection and query expansion.

The paper is organized as follows: Section 2 describes our approach; Section 3 explains our experimental setup; Section 4 discusses our main findings; finally, Section 5 draws some conclusions and outlooks for future work.

"Search Engines", course at the master degree in "Computer Engineering", Department of Information Engineering, and at the master degree in "Data Science", Department of Mathematics "Tullio Levi-Civita", University of Padua, Italy. Academic Year 2022/2023

jesus.moncadaramirez@studenti.unipd.it (J. M. Ramírez); isil.atabek@studenti.unipd.it (I. Atabek); huimin.chen@studenti.unipd.it (H. Chen); michele.canale.1@studenti.unipd.it (M. Canale); nicolo.santini.1@studenti.unipd.it (N. Santini); giovanni.zago.3@studenti.unipd.it (G. Zago)

CEUR Workshop Proceedings (CEUR-WS.org)

2. Methodology

In this section we address how the system was developed starting from the source code of HelloTipster[3] repository developed by Professor Nicola Ferro and presented to us during the Search Engine course. Furthermore, we will present main methodology and approaches using the same structure of the repository[4].

2.1. Analyzer

2.2. Index

2.3. Parse

2.4. Search

3. Experimental Setup

Describe the experimental setup, i.e.

- used collections
- evaluation measures
- url to git repository and its organization
- hardware used for experiments
- ...

4. Results and Discussion

Provide a summary of the performance on the previous year dataset.

Discuss the results and any relevant issues.

5. Conclusions and Future Work

Provide a summary of what are the main achievements and findings.

Discuss future work, e.g. what you may try next and/or how your approach could be further developed.

References

[1] CLEF2023, LongEval CLEF 2023 Lab, https://clef-longeval.github.io/, 2023.

[2] CLEF2023, LongEval Train Collection, https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-5010, 2023.

[3] Nicola Ferro, hello-tipser, https://bitbucket.org/frrncl/se-unipd/src/master/hello-tipster/, 2023.

[4] Canale, Chen, Ramirez, Santini and Zago, jihuming, https://bitbucket.org/upd-dei-stud-prj/seupd2223-jihuming/src/master/, 2023.