Evaluating Temporal Persistence of Information Retrieval Models at CLEF 2024 LongEval-Retrieval Track

Petra Galuščáková
petra.galuscakova@uis.no

University of Stavanger

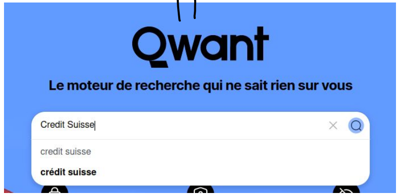
5/4/2024

Outline

1) **Evaluating Temporal Persistence**
2) LongEval-Retrieval CLEF Campaign
    a) Task
    b) Collection
    c) Participation
3) What (did not) work in 2023?
4) How to evaluate relevance?

Web Retrieval

motore francese => focus su privacy utenti



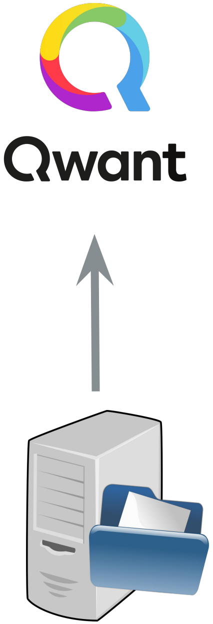
Qwant

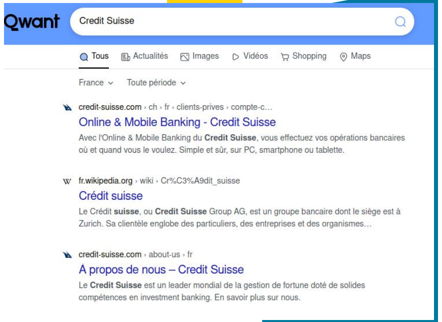
Le moteur de recherche qui ne sait rien sur vous

Credit Suisse

credit suisse

crédit suisse





Qwant Credit Suisse

Tous · Actualités · Images · Vidéos · Shopping · Maps

France · Toute période

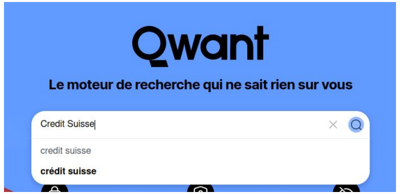
credit-suisse.com › ch › fr › clients-prives › compte-c...
Online & Mobile Banking - Credit Suisse
Avec l'Online & Mobile Banking du **Credit Suisse**, vous effectuez vos opérations bancaires où et quand vous le voulez. Simple et sûr, sur PC, smartphone ou tablette.

fr.wikipedia.org › wiki › Cr%C3%A9dit_suisse
Crédit suisse
Le Crédit **suisse**, ou **Credit Suisse** Group AG, est un groupe bancaire dont le siège est à Zurich. Sa clientèle englobe des particuliers, des entreprises et des organismes…

credit-suisse.com › about-us › fr
A propos de nous – Credit Suisse
Le **Credit Suisse** est un leader mondial de la gestion de fortune doté de solides compétences en investment banking. En savoir plus sur nous.

Evolving Collection (Queries)

ricerche e collezioni cambiano con tempo



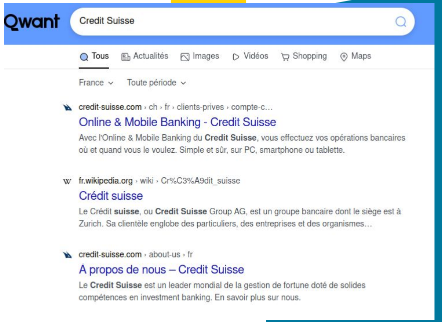
Qwant

Le moteur de recherche qui ne sait rien sur vous

Credit Suisse

credit suisse

crédit suisse



Qwant | Credit Suisse

Tous | Actualités | Images | Vidéos | Shopping | Maps

France ⌄ | Toute période ⌄

credit-suisse.com › ch › fr › clients-prives › compte-c…

Online & Mobile Banking - Credit Suisse

Avec l'Online & Mobile Banking du **Credit Suisse**, vous effectuez vos opérations bancaires où et quand vous le voulez. Simple et sûr, sur PC, smartphone ou tablette.

fr.wikipedia.org › wiki › Cr%C3%A9dit_suisse

Crédit suisse

Le Crédit **suisse**, ou **Credit Suisse** Group AG, est un groupe bancaire dont le siège est à Zurich. Sa clientèle englobe des particuliers, des entreprises et des organismes…

credit-suisse.com › about-us › fr

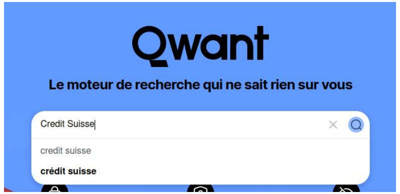
A propos de nous – Credit Suisse

Le **Credit Suisse** est un leader mondial de la gestion de fortune doté de solides compétences en investment banking. En savoir plus sur nous.

March 20, 2023:

1) Motion de censure
2) Assemblée nationale
3) Printemps
4) Grand Prix f1
5) Rennes

Evolving Collection (Queries)



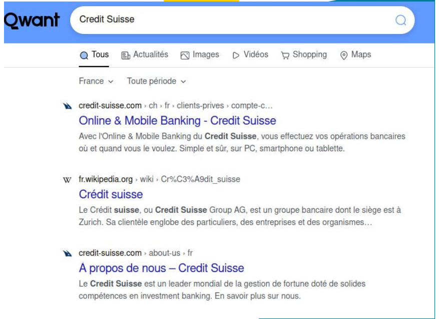
Qwant

Le moteur de recherche qui ne sait rien sur vous

Credit Suisse

credit suisse

crédit suisse



Credit Suisse

credit-suisse.com › ch › fr › clients-prives › compte-c…

Online & Mobile Banking - Credit Suisse

Avec l'Online & Mobile Banking du **Credit Suisse**, vous effectuez vos opérations bancaires où et quand vous le voulez. Simple et sûr, sur PC, smartphone ou tablette.

fr.wikipedia.org › wiki › Cr%C3%A9dit_suisse

Crédit suisse

Le Crédit **suisse**, ou **Credit Suisse** Group AG, est un groupe bancaire dont le siège est à Zurich. Sa clientèle englobe des particuliers, des entreprises et des organismes…

credit-suisse.com › about-us › fr

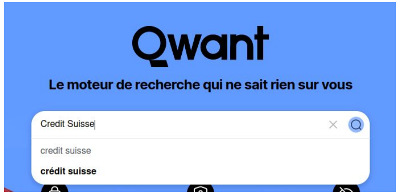
A propos de nous – Credit Suisse

Le **Credit Suisse** est un leader mondial de la gestion de fortune doté de solides compétences en investment banking. En savoir plus sur nous.

March 21, 2023:

1) Paul Grant
2) Muriel Robin Pierre Palmade
3) Bruno Martini
4) Actualite
5) Robbie Williams

Evolving Collection (Documents)



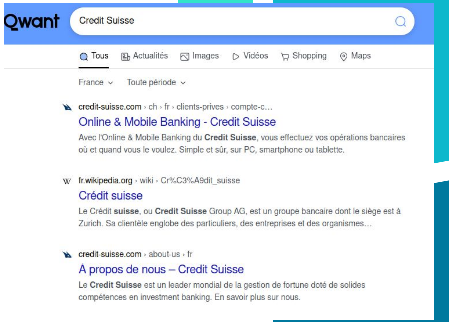
Qwant

Le moteur de recherche qui ne sait rien sur vous

Credit Suisse

credit suisse

crédit suisse



Qwant Credit Suisse

credit-suisse.com › ch › fr › clients-prives › compte-c…

Online & Mobile Banking - Credit Suisse

Avec l'Online & Mobile Banking du **Credit Suisse**, vous effectuez vos opérations bancaires où et quand vous le voulez. Simple et sûr, sur PC, smartphone ou tablette.

fr.wikipedia.org › wiki › Cr%C3%A9dit_suisse

Crédit suisse

Le Crédit **suisse**, ou **Credit Suisse** Group AG, est un groupe bancaire dont le siège est à Zurich. Sa clientèle englobe des particuliers, des entreprises et des organismes…

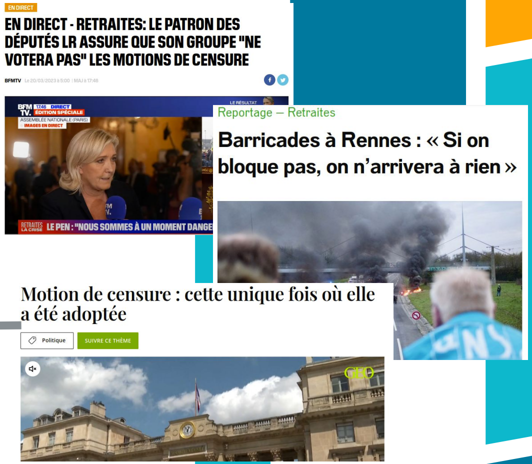
credit-suisse.com › about-us › fr

À propos de nous – Credit Suisse

Le **Credit Suisse** est un leader mondial de la gestion de fortune doté de solides compétences en investment banking. En savoir plus sur nous.

March 20, 2023:

1) Motion de censure
2) Assemblée nationale
3) Printemps
4) Grand Prix f1
5) Rennes



EN DIRECT

EN DIRECT - RETRAITES: LE PATRON DES DÉPUTÉS LR ASSURE QUE SON GROUPE "NE VOTERA PAS" LES MOTIONS DE CENSURE

BFMTV Le 20/03/2023 à 5:00 | MAJ à 17:46

BFM TV. 17:46 DIRECT ÉDITION SPÉCIALE
ASSEMBLÉE NATIONALE (PARIS)
IMAGES EN DIRECT
LE RÉSULTAT
RETRAITES LA CRISE LE PEN : "NOUS SOMMES À UN MOMENT DANGE

Reportage — Retraites

Barricades à Rennes : « Si on bloque pas, on n'arrivera à rien »

Motion de censure : cette unique fois où elle a été adoptée

Politique
SUIVRE CE THÈME
GEO

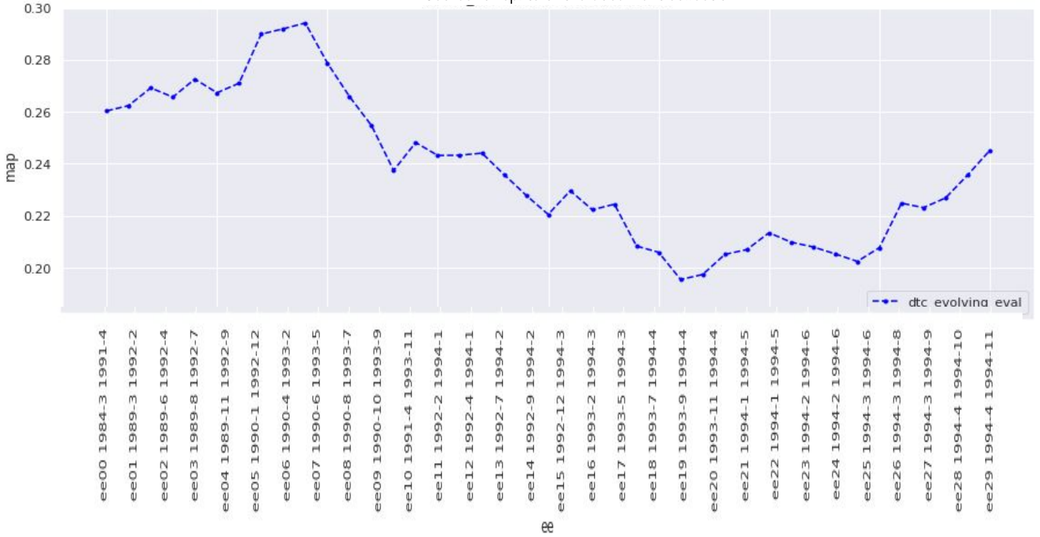
Changes in the IR Framework

- Topics
  - Trends in the topics
- Documents
  - Documents are added to the index
  - Content of the documents changes
  - Documents are removed from the index
- Relevance assessments
  - Trends is general interests (what is considered to be relevant)
  - Subjective relevance for users
- …

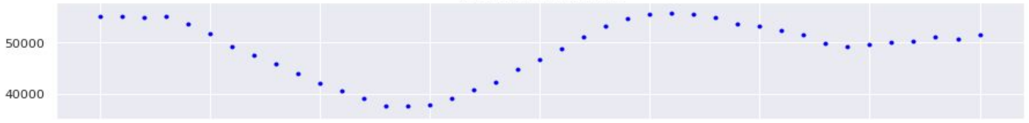
Changes in the IR Framework

- Topics
  - Trends in the topics
- Documents
  - **Documents are added to the index**
  - Content of the documents changes
  - **Documents are removed from the index**
- Relevance assessments
  - Trends is general interests (what is considered to be relevant)
  - Subjective relevance for users
- …

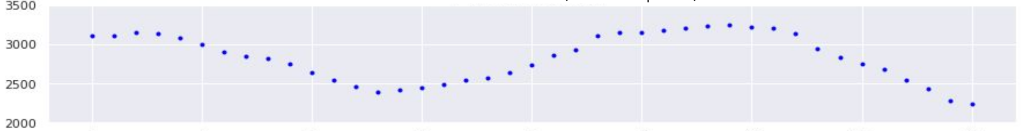
Evolution and IR Performance



MAP Scores for splits of the document collection



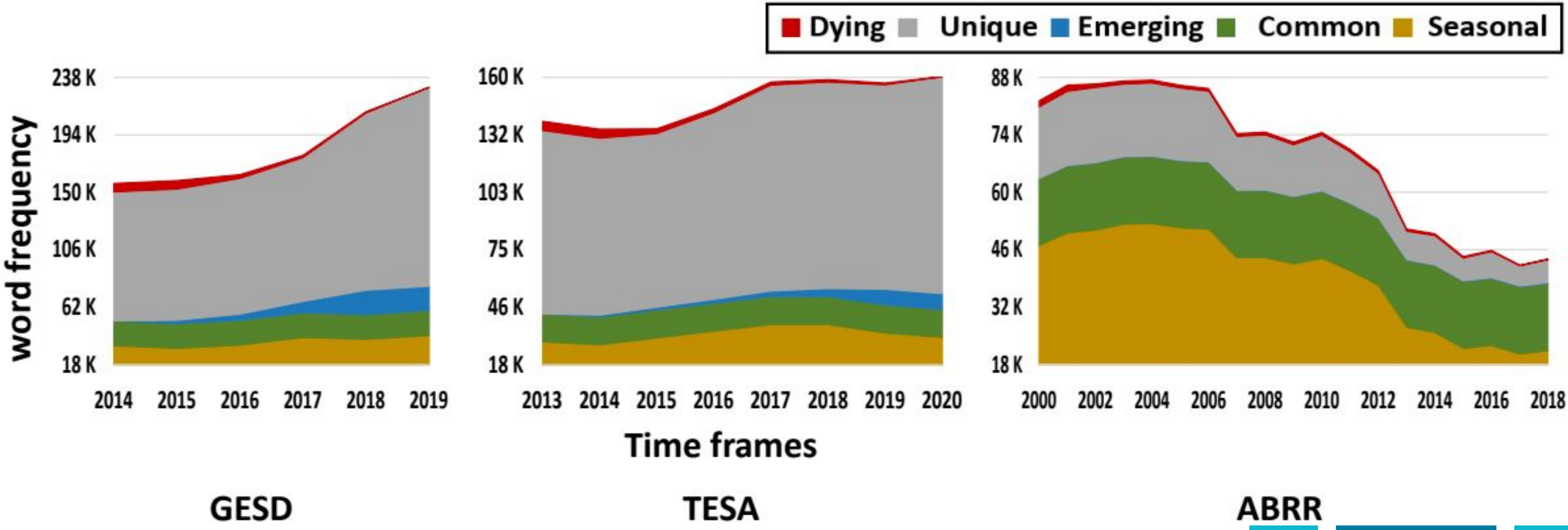
Total Number of Assessments (over 249 queries)



Total Number of Relevant Documents (over 249 queries)

TREC Robust 04 collection (splits of ⅕ of the collection, with 1/10 of the split shift), BM25 run

Evolving Vocabulary

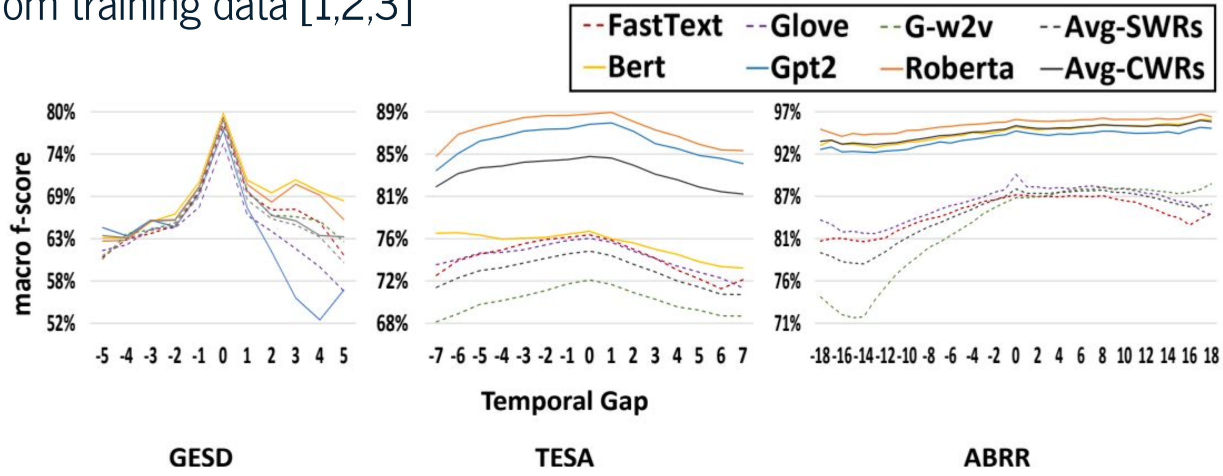


Temporal usage of different word types [1]

[1] Rabab Alkhalifa et al: Building for tomorrow: Assessing the temporal persistence of text classifiers, 2022

Evolution and Neural Models

- The performance of **neural models** in **Text Classification** systems **drops over time** as patterns observed in data change, due to linguistic and societal changes.
- This drop is more pronounced when the testing data is **further away** in time from training data [1,2,3]



Temporal performance of different language representations across the three datasets. [1]

[1] Rabab Alkhalifa et al: Building for tomorrow: Assessing the temporal persistence of text classifiers, 2022
[2] Komal Florio et al.: Time of Your Hate: The Challenge of Time in Hate Speech Detection on Social Media, 2020
[3] Jan Lukes and Anders Søgaard: Sentiment analysis under temporal shift, 2018

Outline

1) Evaluating Temporal Persistence
2) **LongEval-Retrieval CLEF Campaign**
    a) **Task**
    b) Collection
    c) Participation
3) What (did not) work in 2023?
4) How to evaluate relevance?

Questions

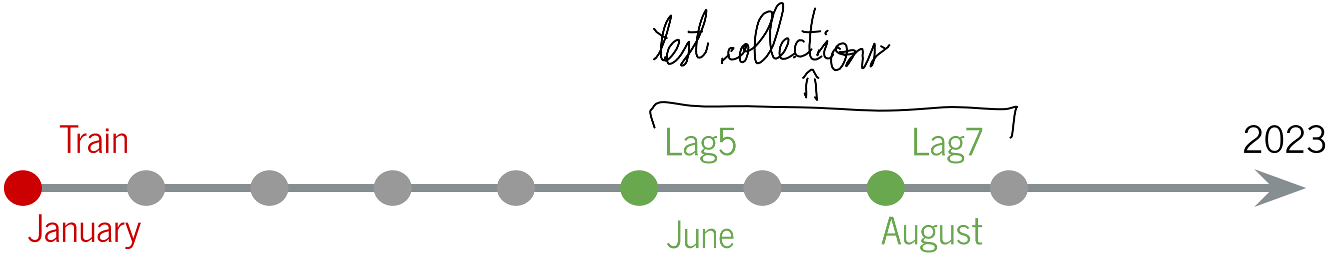
1) How does **search engine** behave as the **collection evolves**?

2) What IR systems are the **most robust** against the **changes** in the collection (and still perform well)?

3) When do we need to **update an IR** system as the **collection evolves?**

LongEval-Retrieval CLEF Campaign

- Build a **succession of train/test collections**
- **Each** of them composed of a set of **documents** from Qwant's index **queries** from Qwant's users
- Designed to reflect the **changes** of the Web across time



test collections

Train

January

Lag5

June

Lag7

August

2023

collezioni create nello stesso modo

LongEval-Classification CLEF Campaign

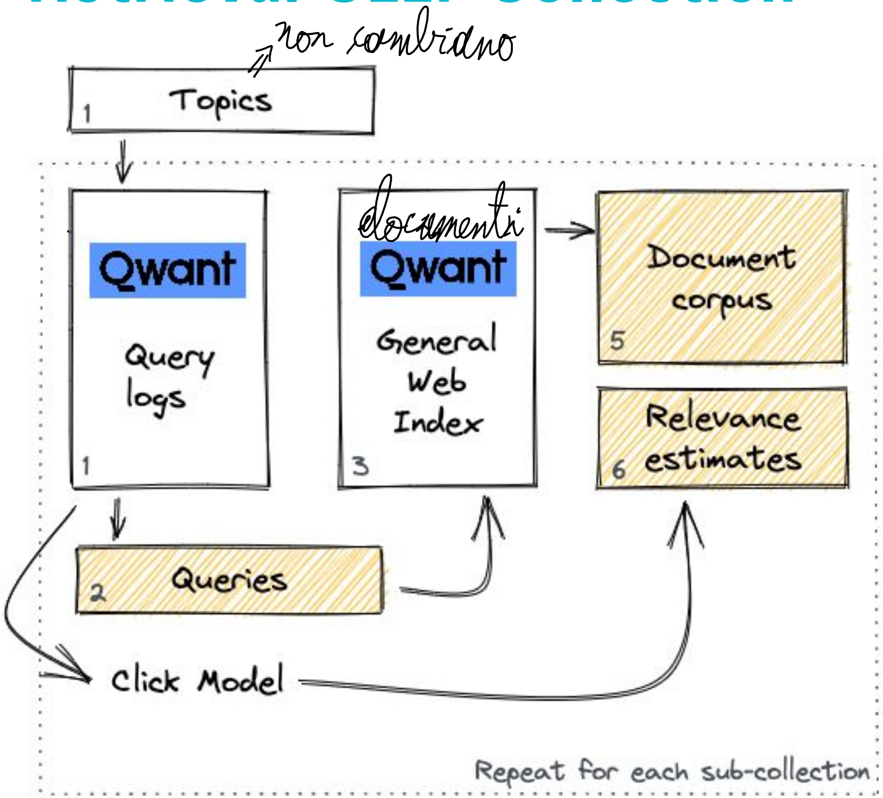
non il nostro

- Explore temporal persistence on **classification of sentiment** of **Tweets**
- TM-Senti Dataset:
  - Large-scale Tweets sentiment dataset
  - Tweets are binary labelled for sentiment as either "positive" or "negative"
- Long-term and short-term subtasks
- Evaluation: Macro-averaged F1-score and Relative Performance Drop

Outline

1) Evaluating Temporal Persistence
2) **LongEval-Retrieval CLEF Campaign**
    a) Task
    b) **Collection**
    c) Participation
3) What (did not) work in 2023?
4) How to evaluate relevance?

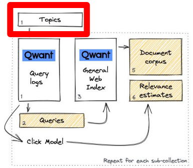
LongEval-Retrieval CLEF Collection



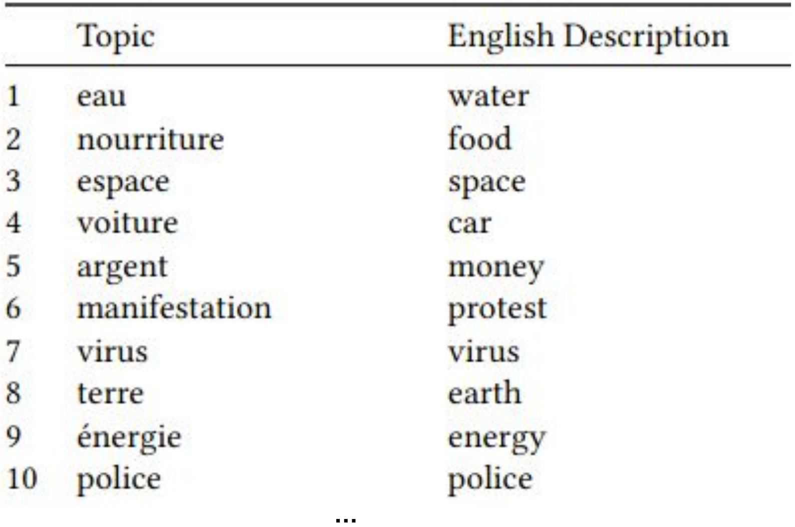
arXiv > cs > arXiv:2303.03229

SIGIR TAIPEI TAIWAN 2023

Topics



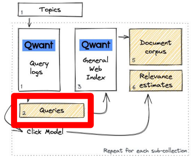
- Selected to ensure a balance between **popularity**, **stability**, **generality**, and **diversity**.
- Selected from the Web and social media based on **trends**.
- Performed only **once** for the entire LongEval-Retrieval collection.



	Topic	English Description
1	eau	water
2	nourriture	food
3	espace	space
4	voiture	car
5	argent	money
6	manifestation	protest
7	virus	virus
8	terre	earth
9	énergie	energy
10	police	police

...

Queries



- In the context of the LongEval-Retrieval test collection, a query is a **multi-word chain** of characters that is **related** to one or more **topics**.

- Topics are used for filtering and selecting the Qwant actual user's queries.
  - Select the queries which **contain topic words**

- Extracted for **each** sub-collection

Translations

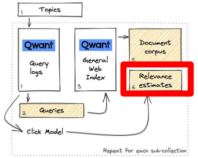
- Initially, all the queries and documents were created in French.
- French-English translation system CUBBITT:
  - https://lindat.mff.cuni.cz/services/translation
- The quality of the translations of the queries is much lower than the quality of the document translation.
- **Multiple** possible **translations** provided for the queries

Query Translations

q01235	**a chapeau**	a hat	a Chief	a chapeau	a chapeau
q01236	**a quoi sert la prise de terre**	What is the purpose of the land grab?	What is the use of the land grab?	a what is the purpose of the land grab	a what is the purpose of the land grab?
q01238	**abonnement ter bordeaux**	subscribing bordeaux	subscribing to bordeaux	subscribing to the bordeaux	bordeaux subscription
q012319	**affutage couteaux**	sharpening knives	knife sharpening	knife-sharpening	Knife sharpening
q012331	**aide a la reprise d'emploi pole emploi**	senior employment assistance	Senior Employment Assistance	Senior employment assistance	assistance senior employment

possibile usare sia una che tutte
traduzioni

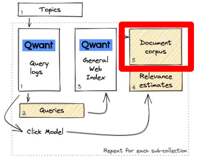
Relevance Assessments



- Based on user **clicks** ⇒ creati automaticamente
- Raw clicks cannot be used:
  - Noisy and biased
- Use click **models**
  - Goal: better model the behaviour of users
    e conservare privacy

- 0 = not relevant, 1 = relevant, 2 = highly relevant

Documents



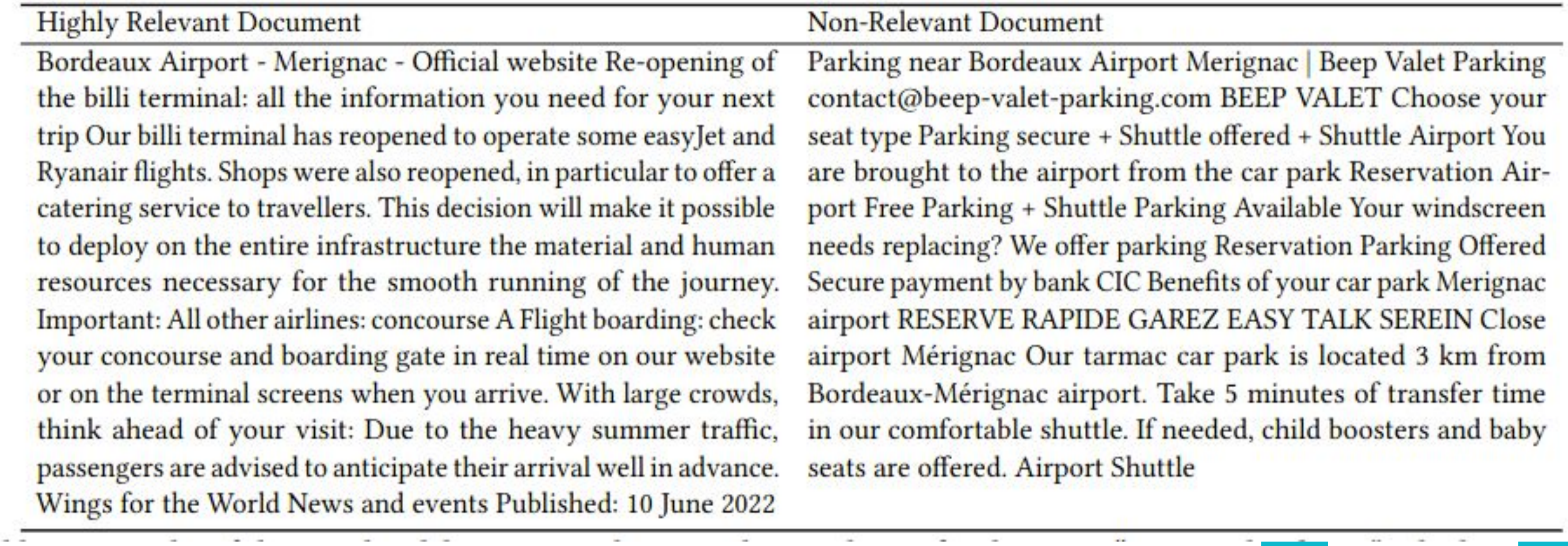
- 2 approaches to selection:

  risultati ricerche

  - Documents that were **displayed in SERPs** for the queries that we selected
  - Potentially non-relevant documents are **randomly sampled** from Qwant index in order to better represent the nature of a Web test collection (~100,000 documents per topic)

- Web documents, **downloaded and cleaned**

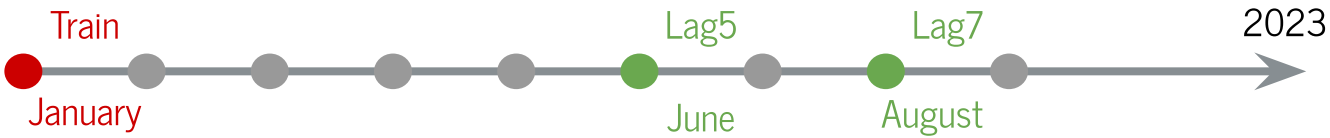
Documents Example



Highly Relevant Document	Non-Relevant Document
Bordeaux Airport - Merignac - Official website Re-opening of the billi terminal: all the information you need for your next trip Our billi terminal has reopened to operate some easyJet and Ryanair flights. Shops were also reopened, in particular to offer a catering service to travellers. This decision will make it possible to deploy on the entire infrastructure the material and human resources necessary for the smooth running of the journey. Important: All other airlines: concourse A Flight boarding: check your concourse and boarding gate in real time on our website or on the terminal screens when you arrive. With large crowds, think ahead of your visit: Due to the heavy summer traffic, passengers are advised to anticipate their arrival well in advance. Wings for the World News and events Published: 10 June 2022	Parking near Bordeaux Airport Merignac | Beep Valet Parking contact@beep-valet-parking.com BEEP VALET Choose your seat type Parking secure + Shuttle offered + Shuttle Airport You are brought to the airport from the car park Reservation Airport Free Parking + Shuttle Parking Available Your windscreen needs replacing? We offer parking Reservation Parking Offered Secure payment by bank CIC Benefits of your car park Merignac airport RESERVE RAPIDE GAREZ EASY TALK SEREIN Close airport Mérignac Our tarmac car park is located 3 km from Bordeaux-Mérignac airport. Take 5 minutes of transfer time in our comfortable shuttle. If needed, child boosters and baby seats are offered. Airport Shuttle

Query "**aeroport bordeaux"**, which was translated as "**airport**"

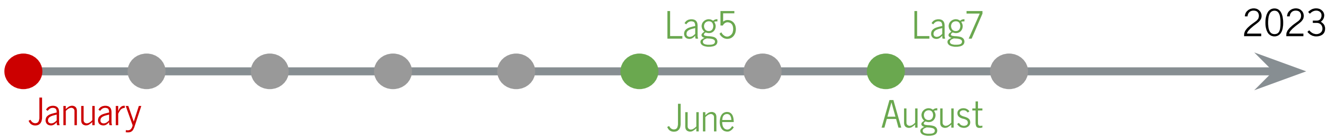
Train and Test Sets



- **Train collection**: ~2M Web pages, 30k queries, 133k assessments (60% are non-relevant, 19% are relevant, and 21% are highly relevant)
- **Lag5 Test Collection**: ~3M Web pages, ~20k queries
- **Lag7 Test Collection**: ~3M Web pages, ~25k queries

6% were highly relevant in 2023

Evaluation

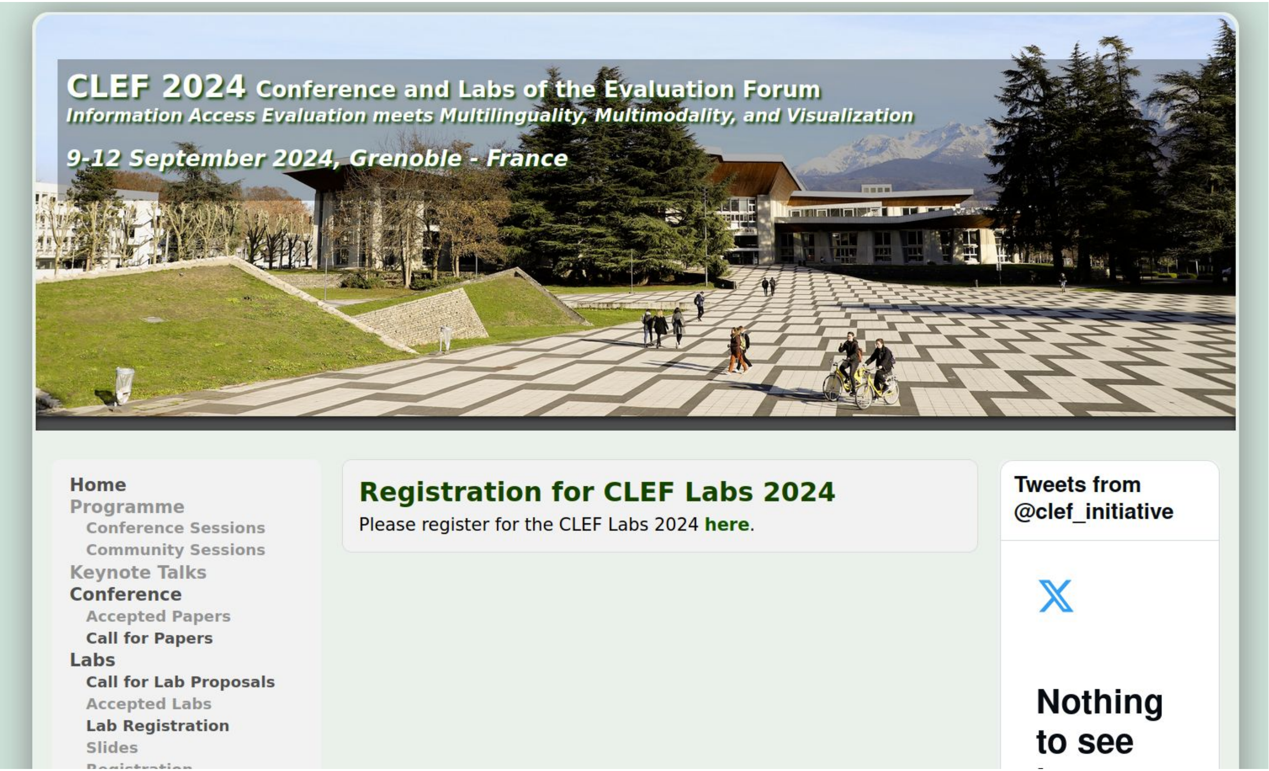


- **nDCG** (Lag5 and Lag7)
- **Relative nDCG Drop** between Lag5 and Lag7

Outline

1) Evaluating Temporal Persistence
2) **LongEval-Retrieval CLEF Campaign**
    a) Task
    b) Collection
    c) **Participation**
3) What (did not) work in 2023?
4) How to evaluate relevance?

Registration to CLEF 2024



CLEF 2024 Conference and Labs of the Evaluation Forum

Information Access Evaluation meets Multilinguality, Multimodality, and Visualization

9-12 September 2024, Grenoble - France

- Home
- Programme
  - Conference Sessions
  - Community Sessions
- Keynote Talks
- Conference
  - Accepted Papers
  - Call for Papers
- Labs
  - Call for Lab Proposals
  - Accepted Labs
  - Lab Registration
  - Slides
  - Registration

Registration for CLEF Labs 2024

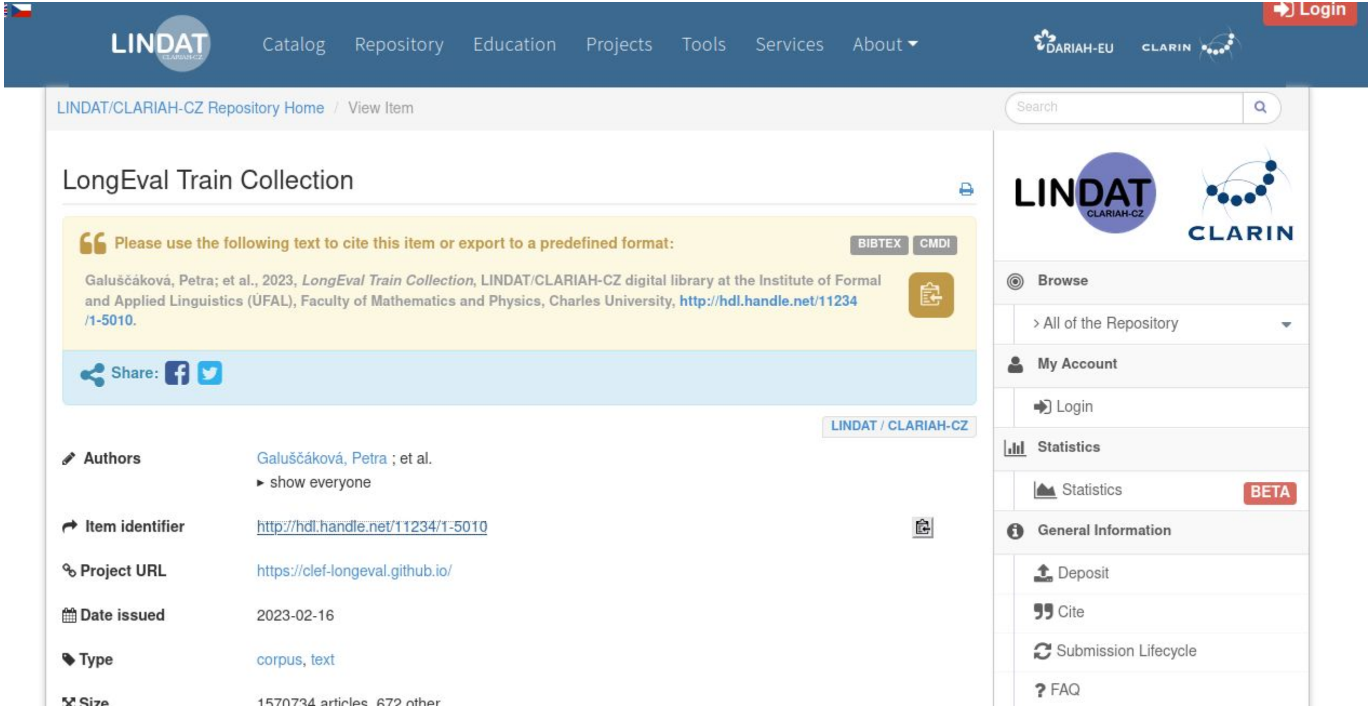
Please register for the CLEF Labs 2024 **here**.

Tweets from @clef_initiative

Nothing to see

CLEF Web site: https://clef2024.imag.fr

Data Download



LINDAT/CLARIAH-CZ Repository Home / View Item

LongEval Train Collection

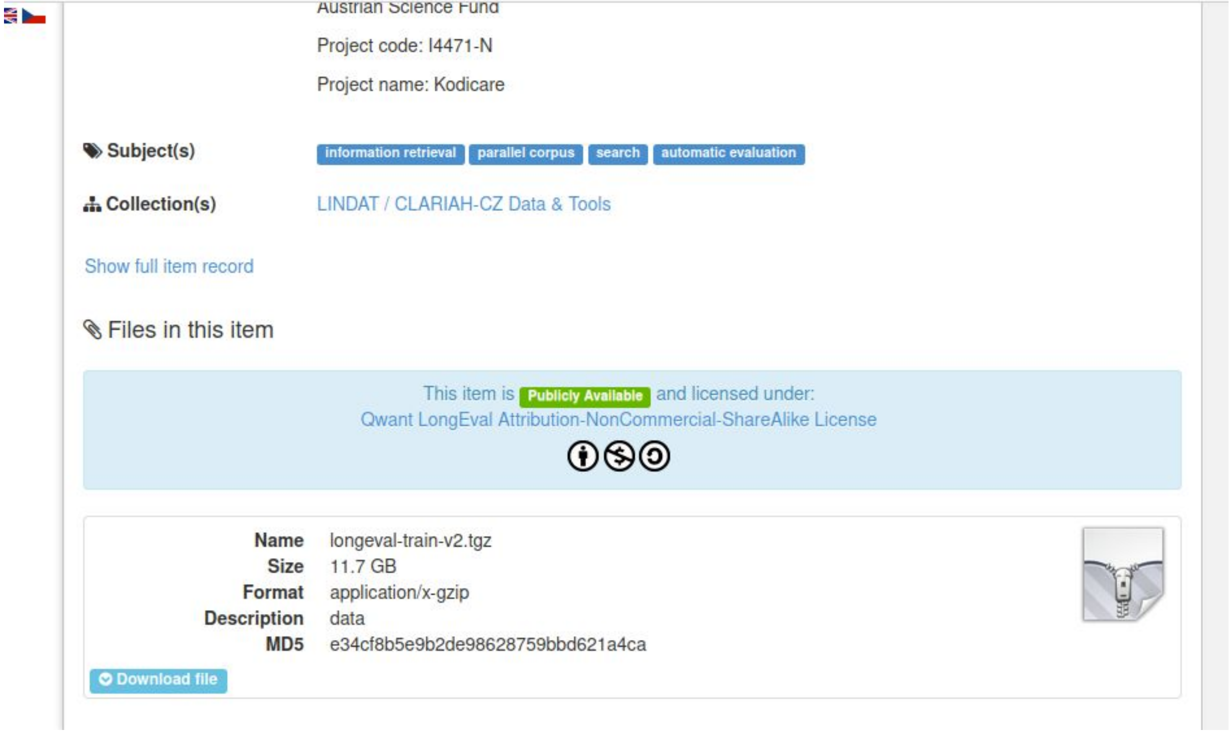
Please use the following text to cite this item or export to a predefined format:

Galuščáková, Petra; et al., 2023, *LongEval Train Collection*, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1-5010.

LINDAT / CLARIAH-CZ

Authors	Galuščáková, Petra ; et al. ▸ show everyone
Item identifier	http://hdl.handle.net/11234/1-5010
Project URL	https://clef-longeval.github.io/
Date issued	2023-02-16
Type	corpus, text
Size	1570734 articles, 672 other

Data Download



Austrian Science Fund

Project code: I4471-N

Project name: Kodicare

Subject(s) information retrieval, parallel corpus, search, automatic evaluation

Collection(s) LINDAT / CLARIAH-CZ Data & Tools

Show full item record

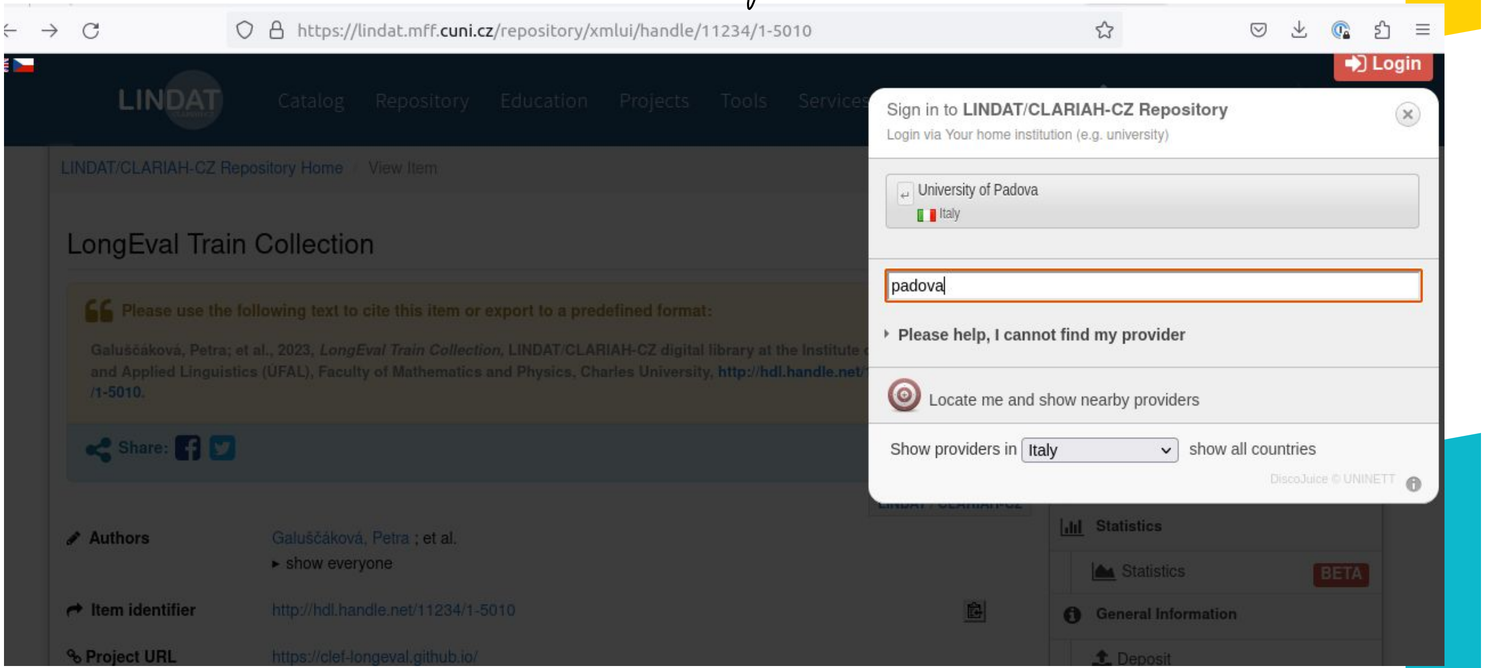
Files in this item

This item is Publicly Available and licensed under:
Qwant LongEval Attribution-NonCommercial-ShareAlike License

Name	longeval-train-v2.tgz
Size	11.7 GB
Format	application/x-gzip
Description	data
MD5	e34cf8b5e9b2de98628759bbd621a4ca

Data Download

loggare per scaricare



https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-5010

LINDAT/CLARIAH-CZ Repository Home / View Item

LongEval Train Collection

Please use the following text to cite this item or export to a predefined format:

Galuščáková, Petra; et al., 2023, *LongEval Train Collection*, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1-5010.

Share:

Authors: Galuščáková, Petra ; et al.
show everyone

Item identifier: http://hdl.handle.net/11234/1-5010

Project URL: https://clef-longeval.github.io/

Sign in to **LINDAT/CLARIAH-CZ Repository**
Login via Your home institution (e.g. university)

University of Padova
Italy

padova

Please help, I cannot find my provider

Locate me and show nearby providers

Show providers in Italy show all countries

DiscoJuice © UNINETT

Submission Requirements

- **For each system submitted, the participants need to submit the results acquired by this system on both Lag5 and Lag7 queries**
- Participants also need to provide a short **description** of each of the submitted systems.
- Each team can submit up to **5 systems**.

Submission Format for a System

Each system should be submitted in a **single zipped file** consisting of a following tuple:

- team_system.L5 (Lag5 test set)
- team_system.L7 (Lag7 test set)
- team_system.meta (run description)

⇒ codice deve generarli, noi carichiamo anche su TIRA

system.** is a **TREC run** in TREC format (max 1000 documents)

If run is BM25, team UGA, in-time set: **UGA_BM25.L5**

Submissions

- Will be done using **git**
- Each team will have its own private repository created on bitbucket

LongEval at CLEF 2024

- **April:** Data release
- **May:** Participants' submissions (Still time to register and participate !)
- **June:** Participants' papers submissions
- **June:** Evaluation results release
- **July:** Camera ready paper submissions
- **September:** CLEF Conference in Grenoble

LongEval Web site: https://clef-longeval.github.io

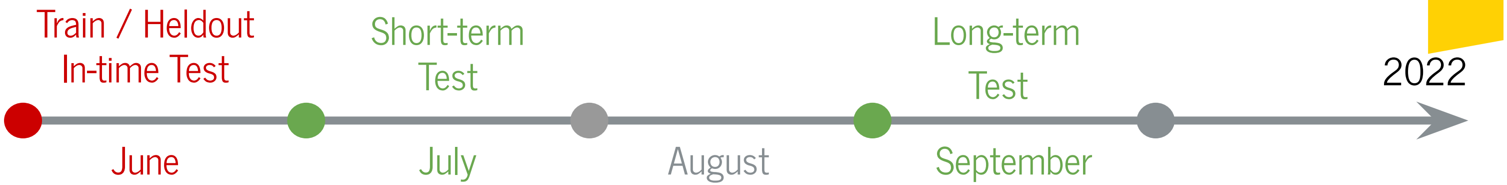
LongEval CLEF 2024 Lab

Longitudinal Evaluation of Model Performance

Outline

1) Evaluating Temporal Persistence
2) LongEval-Retrieval CLEF Campaign
    a) Task
    b) Collection
    c) Participation
3) **What (did not) work in 2023?**
4) How to evaluate relevance?

2023 Collection



Train / Heldout In-time Test

Short-term Test

Long-term Test

2022

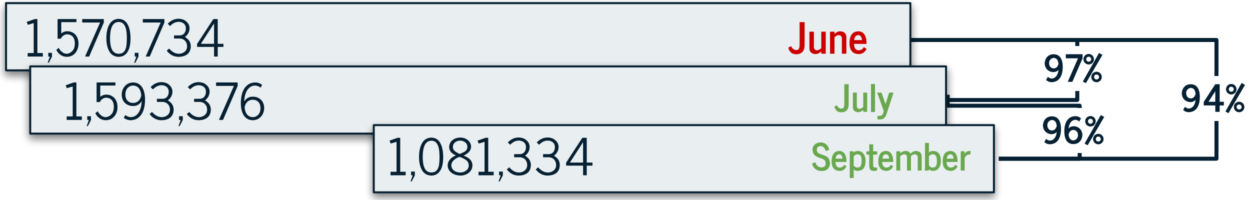
June

July

August

September

Documents



1,570,734 June

1,593,376 July

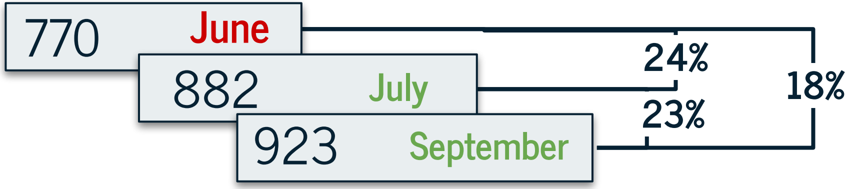
1,081,334 September

97%

96%

94%

Queries



770 June

882 July

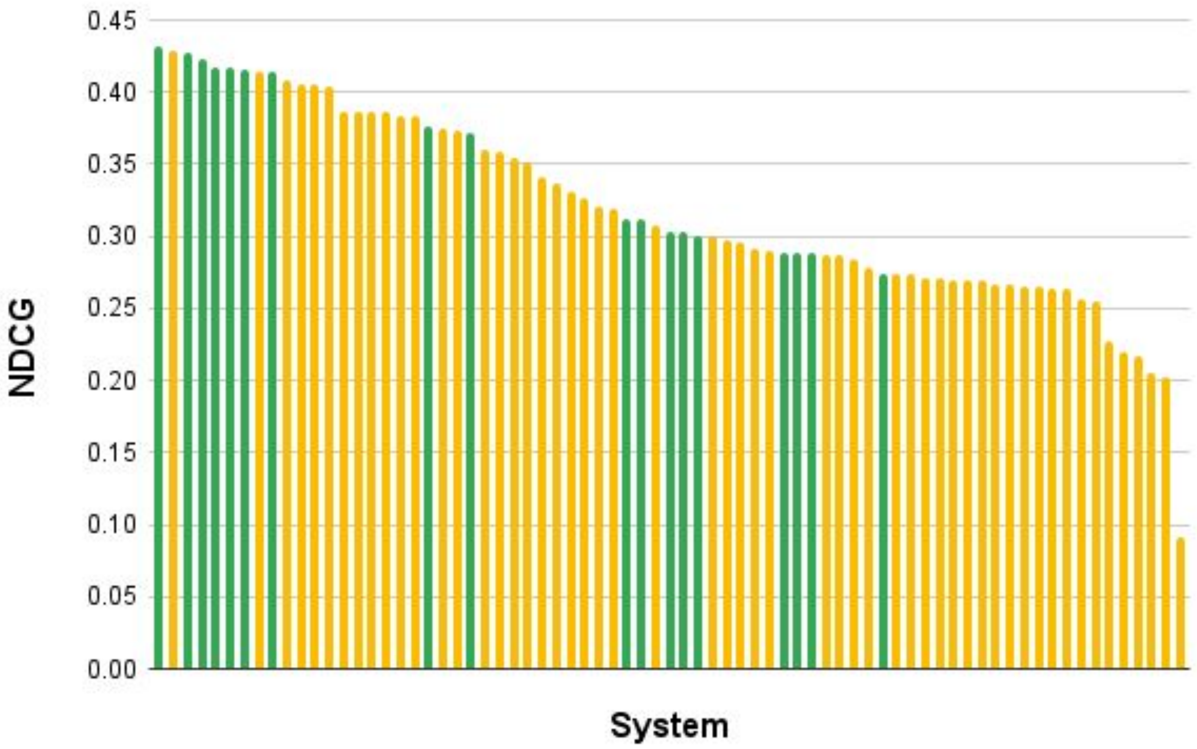
923 September

24%

23%

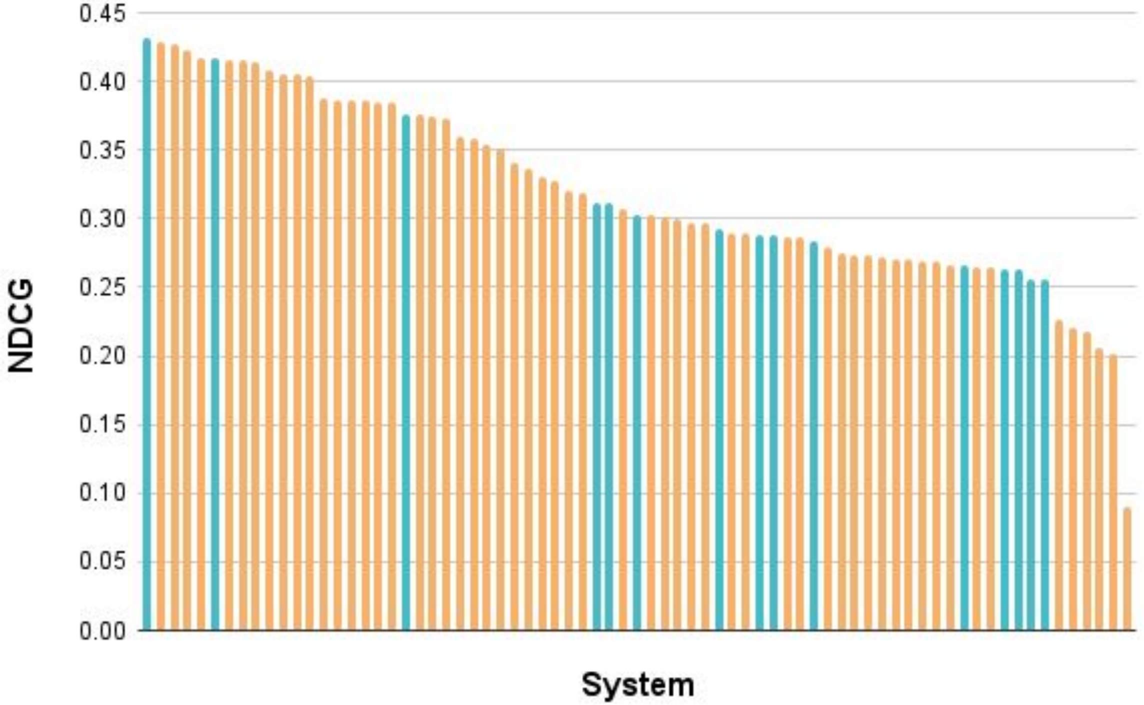
18%

Neural vs Statistical Models



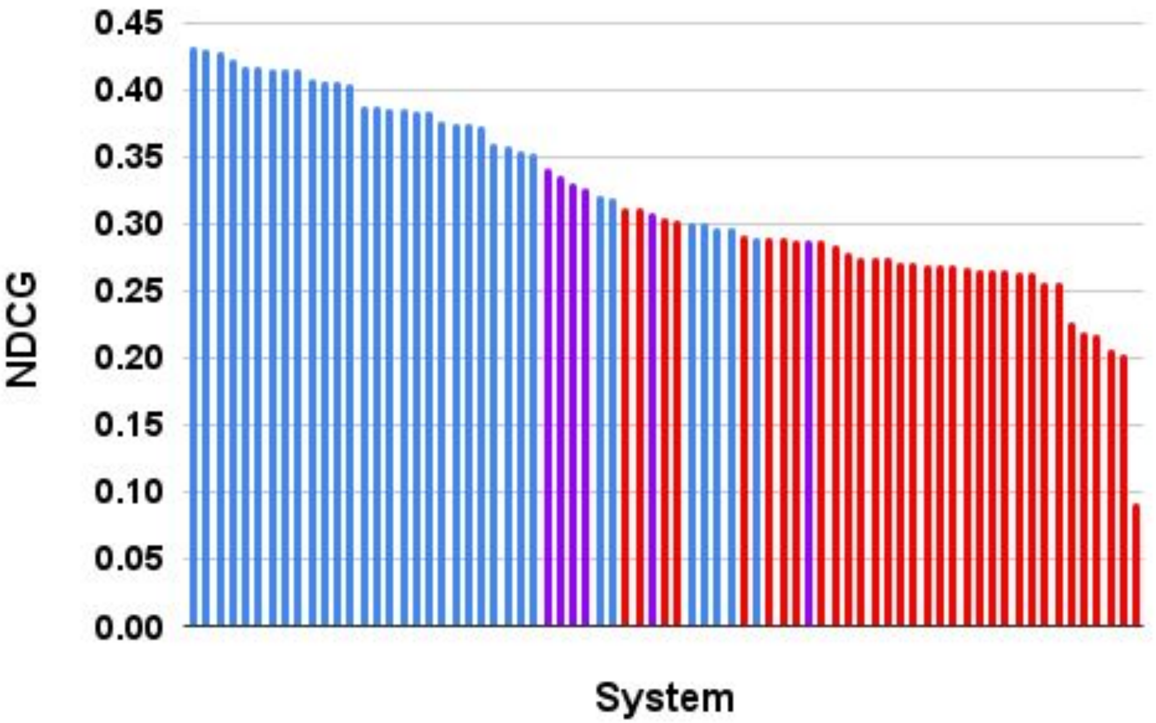
Performance of neural (green) and non-neural (yellow) approaches on In-Time test

Single System vs. Combination



Performance of single approach (orange) and combinations (cyan) on In-Time test

French vs. English



Performance of systems which use French (blue), English (red) or both languages (purple)
on In-Time test

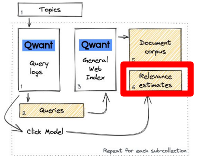
Ideas for 2024

- Use overlaps and changes between data
- Use larger training data
  - Fine-tune models on train data
- Use n-best translations
- Make mixture of models effective
- …

Outline

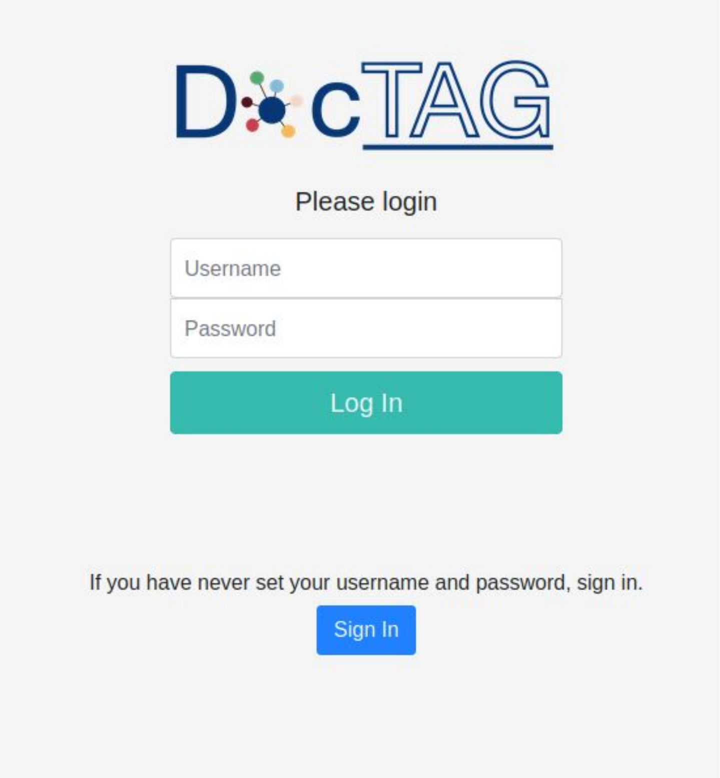
1) Evaluating Temporal Persistence
2) LongEval-Retrieval CLEF Campaign
    a) Task
    b) Collection
    c) Participation
3) What (did not) work in 2023?
4) **How to evaluate relevance?**

Relevance Assessments



- Based on user **clicks**
- Raw clicks cannot be used:
  - Noisy and biased
- Use click **models**
  - Goal: better model the behaviour of users
- 0 = not relevant, 1 = relevant, 2 = highly relevant

We also aim to acquire manual relevance judgements

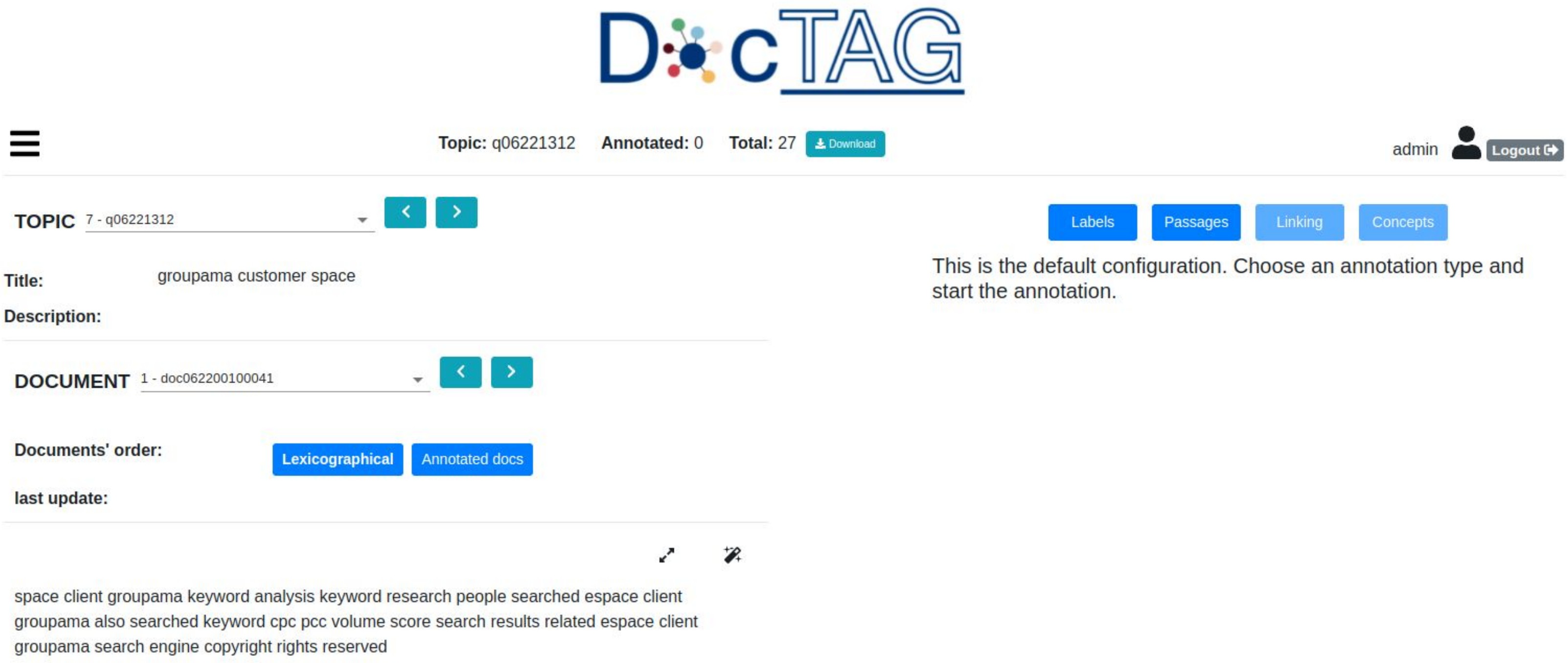


Please login

If you have never set your username and password, sign in.

https://dscapp01.researchstudio.at/

DocTAG



Topic: q06221312 Annotated: 0 Total: 27

TOPIC 7 - q06221312

Title: groupama customer space

Description:

This is the default configuration. Choose an annotation type and start the annotation.

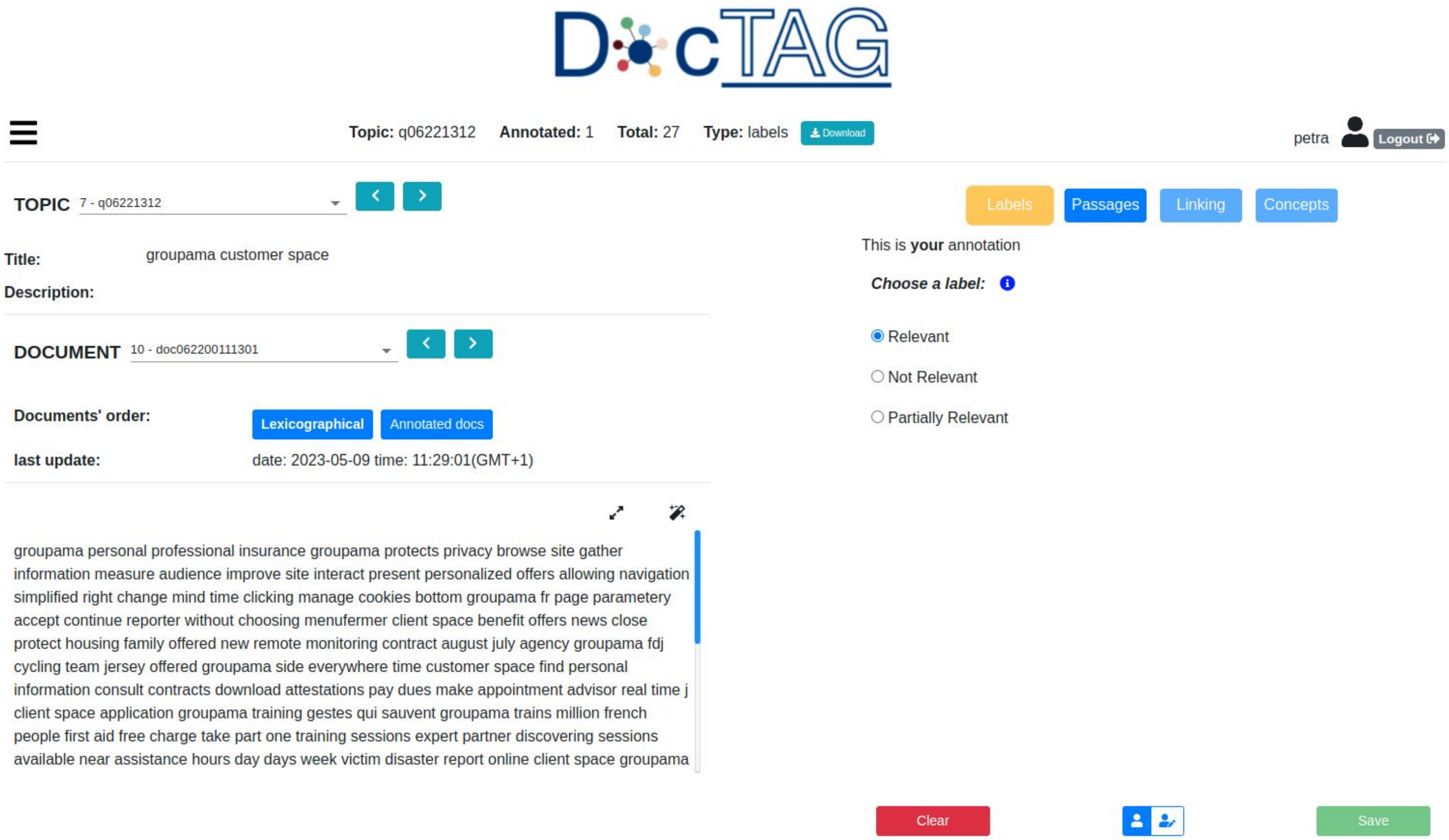
DOCUMENT 1 - doc062200100041

Documents' order: Lexicographical Annotated docs

last update:

space client groupama keyword analysis keyword research people searched espace client groupama also searched keyword cpc pcc volume score search results related espace client groupama search engine copyright rights reserved

DocTAG



Topic: q06221312 **Annotated:** 1 **Total:** 27 **Type:** labels

petra

TOPIC 7 - q06221312

Title: groupama customer space

Description:

DOCUMENT 10 - doc062200111301

Documents' order: Lexicographical Annotated docs

last update: date: 2023-05-09 time: 11:29:01(GMT+1)

groupama personal professional insurance groupama protects privacy browse site gather information measure audience improve site interact present personalized offers allowing navigation simplified right change mind time clicking manage cookies bottom groupama fr page parametery accept continue reporter without choosing menufermer client space benefit offers news close protect housing family offered new remote monitoring contract august july agency groupama fdj cycling team jersey offered groupama side everywhere time customer space find personal information consult contracts download attestations pay dues make appointment advisor real time j client space application groupama training gestes qui sauvent groupama trains million french people first aid free charge take part one training sessions expert partner discovering sessions available near assistance hours day days week victim disaster report online client space groupama

This is **your** annotation

Choose a label:

● Relevant

○ Not Relevant

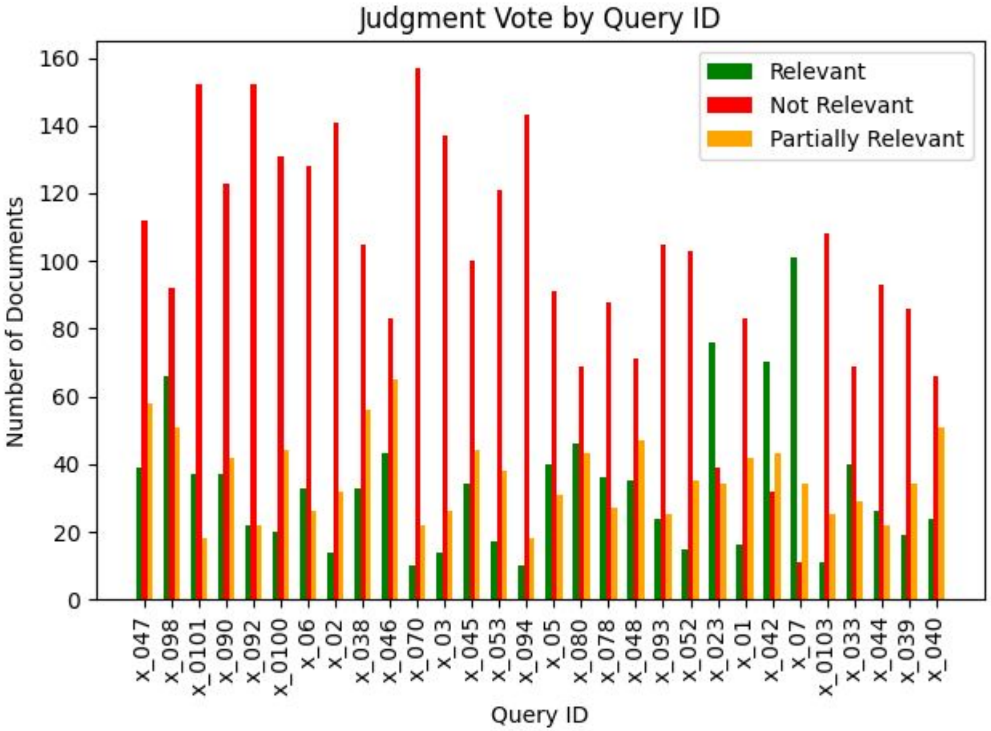
○ Partially Relevant

Document Relevance

- If you are a user, you search using the query terms of the topic and find the Web page, **would you stop your search** and find the search **successful**?
  - Yes -> **Relevant**.
  - Continue searching -> **Not Relevant**
- If you found some information relevant to your query, but you would still continue searching
    - -> **Partially Relevant**
- The labeling option '**I don't know**' should be used only if it is not possible to do any assessment.

Human Relevance Judgments in 2023

- 37 judges
- 150 queries
- ~130 jud/q
- 14,953 jud.



LongEval-Retrieval Organizers

Romain Deveaud; Qwant

Alaa El-Ebshihy; Research Studios Austria

Gabriela Nicole Gonzalez Saez; Université Grenoble-Alpes

Lorraine Goeuriot; Université Grenoble-Alpes

David Iommi; Research Studios Austria

Philippe Mulhem; Université Grenoble-Alpes

Florina Piroi; TU Wien

Martin Popel; Charles University

Question?

https://clef-longeval.github.io/

longeval-ir-task@univ-grenoble-alpes.fr