Extraction de contexte pour génération de résumés simplifiés : vers une application au domaine juridique

Eric SanJuan, Stéphane Huet (LIA)
{eric.sanjuan,stephane.huet}@univ-avignon.fr
Adrian Chifu (LIS)
adrian.chifu@lis-lab.fr

LABORATOIRE INFORMATIQUE D'AVIGNON

JPEG
Laboratoire des Sciences Juridiques, Politique, Économiques et de Gestion

LIS
LABORATOIRE D'INFORMATIQUE & SYSTÈMES

Objectif de l’extraction = RAG interne

Fournir un contexte aussi précis que possible au LLM qui va produire un résumé simplifié

- contexte exhaustif de taille aussi réduite que possible destinés à être soumis à des LLMs locaux non connectés à un moteur de recherche.
- contrôle des hallucinations : le résumé produit ne doit pas introduire d’élément nouveau mais éventuellement explicité les termes du contexte.

Sommaire

Deux parties

- Leçons de la tâche 1 CLEF Simple Text 2024:
  YeSQL & Mini LM & Ollama
- Application à JudiLibre:
  Des sacs de mots aux sacs de vecteurs

Leçons apprises des épisodes précédents

CLEF SimpleText task 1 2022 - 2023

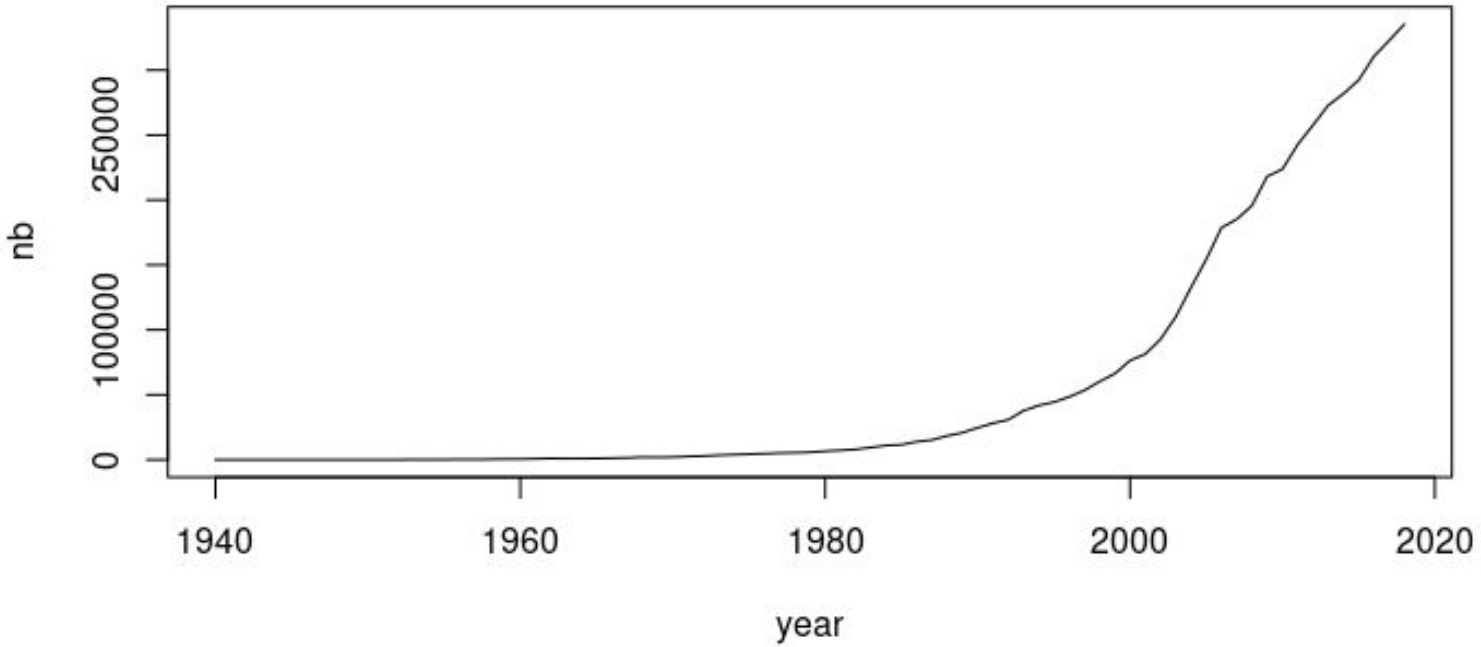
- Documents : DBLP < 2019 titres et résumés + Citations ACM
- Requêtes : Articles de presse grand public sur des questions sociétales ou des technologies liées à l'informatique

Quelques résultats inattendus

- Efficacité des plongements vectoriels denses à froid pré-calculés pour tous les titres et résumés
- Performance des bases de données relationnelles vis à vis d'approches NoSQL

Ressources Simple Text 2024 task 1 (Corpus)

Notices DBLP avant 2019 + Résumés + Citations ACM
+ MS Academic Graph 2019 + Sentence Transformers all-MiniLM-L6-v2



Ressources Simple Text 2024 task 1 (Requêtes)

Thèmes extraits d'articles du Guardian publiés après 2019 :

1. How AI systems, especially virtual assistants, can perpetuate gender stereotypes?
2. Concerns related to the handling of sensitive information by voice assistants.
3. How children interact with voice assistants and the design of child-friendly interfaces.
4. Use of AI to improve success rates and speed in the pharmaceutical research field.
5. Application of machine learning algorithms to predict genomic features, functions, and the outcomes of gene-editing interventions like Crispr.
6. Ethical considerations, governance frameworks, and policies for the responsible development and deployment of AI technologies.
7. Use of NLP techniques to detect and analyze misinformation in textual content on social media platforms.
8. Understand the cryptographic underpinnings of blockchain technology, which is the foundation of Bitcoin and other cryptocurrencies
9. Computer science techniques to analyze spatial data and imagery, particularly for reconstructing crime scenes or human rights violation incidents
10. Study of robotic technologies and automated systems that are replacing human labor in various sectors.

Ressources Simple Text 2024 task 1 (Références)

query	nb
G01.C1	127
G02.C1	118
G03.C1	110
G04.C1	157
G05.C1	93
G06.C1	117
G07.C1	110
G08.C1	134
G09.C1	158
G10.C1	158

rel	nb
0	672
2	311
1	299

Evaluation systèmes “officielle” 2024

Run	MRR	Precision		NDCG		Bpref	MAP
		10	20	10	20		
baseline_bool	0.6500	0.4100	0.2550	0.3167	0.2328	0.1216	0.0694
baseline_elasti	0.8192	0.5200	0.4050	0.4434	0.3623	0.3968	0.2000
baseline_meili	0.5124	0.1900	0.1000	0.1254	0.0869	0.0779	0.0209
baseline_vir_abstract	**0.9500**	0.8200	**0.6150**	0.6701	**0.5543**	**0.5533**	**0.2996**
baseline_vir_title	0.9333	**0.8600**	0.5650	**0.7184**	0.5415	0.5196	0.2633

Table 1: Evaluation of SimpleText Task 1 (Test qrels: Guardian).

Evaluation systèmes “étendue” 2024

Run	MRR	Precision		NDCG		Bpref	MAP
		10	20	10	20		
baseline_bool	0.6500	0.4100	0.2550	0.2987	0.2264	0.1026	0.0590
baseline_elastic	0.8192	0.5200	0.4050	0.4065	0.3434	0.3534	0.1743
baseline_meili	0.5124	0.1900	0.1000	0.1161	0.0833	0.0689	0.0187
baseline_vir_abstract	**0.9500**	0.8200	**0.6200**	0.6276	0.5394	0.5164	**0.2730**
baseline_vir_title	0.9333	**0.8700**	0.5950	**0.6806**	**0.5527**	**0.5244**	0.2646

Table 1: Evaluation of SimpleText Task 1 (Test qrels: Guardian).

Back to SQL

```sql
create table st1_qrels_2024_08 as
  (select topic, query,
  dblp_knn_title(qdblp_v_title(doc),10,-0.8) as doc, rel
  from st1_qrels_2024 where rel=2)
union (select * from  st1_qrels_2024 where rel<2);

CREATE OR REPLACE FUNCTION public.dblp_knn_title(v_query vector(384), nb numeric)
       RETURNS TABLE(id text, ip float, doc json)
       LANGUAGE plpgsql
       AS $$
       BEGIN
        SET LOCAL ivfflat.probes = 65;
        SET LOCAL enable_seqscan = off;
        SET LOCAL min_parallel_table_scan_size = 1;
        SET LOCAL parallel_setup_cost = 1;
        RETURN QUERY
        SELECT J.id, (J.title_v <=> v_query) AS ip, D.doc AS doc
        FROM dblp_v AS J, dblp AS D
        WHERE D.id = J.id
        ORDER BY ip LIMIT nb;
       END;
       $$
       ;
```

Peut-on automatiser les annotations ?

prefix = "Here is a societal question and a scientific paper in computer science. Please, do not recomend papers that are off topic. I do not have time to read them all.",
text = paste("To address the question: ",dfr$query_text,", should I read the paper wich title is:",dfr$title, "ans which abstract is :",dfr$abstract),
prompt = "Answer only returning a relevance score 0, 1 or 2. 0: Not really relevant, 1: relevant, 2: very relevant",
template = "{prefix}{text}\n{prompt}",
system = "You are a journalist writing about a tech topic that raises societal questions. Your are looking for scientific publications that could feed your paper for a large audiance."

Evaluation des LLM locaux (Rollama)

Model	tau	P-value	Accuracy (3)	Accuracy (2)
Qwen	-1,25 %	63,29 %	23,24 %	48,44 %
Qwq	30,00 %	***	32,17 %	52,73 %
Gemma3 :small	33,26 %	***	37,00 %	58,45 %
gemma3:12b	31,95 %	***	38,29 %	59,59 %
Phi4	41,54 %	***	45,16 %	62,79 %
Llama 4	40,04 %	***	52,11 %	69,58 %

RI neuronale juridique

Open data décisions de justice

JSON extrait des PDFs Appel et Cassation

Résumés par extraction

Représentation dense unique
Larges Modèles de Langue

Segmentation en paragraphes

Mini Modèles de Langue

Recherche de passages

Recherche de décisions

Réordonnancement

Modèle relationnel

Jointure et agrégation

Suivi d’affaires (case retrieval)

PoC

Open data décisions de justice

PostGreSQL + pg_vector + curl

Importation JSON

Opérateurs NoSQL

Segmentation en paragraphes

Vecteur dense MiniLM-L12-v2-mmarcoFR

Réduction k-means (ivvflat)

Recherche de décisions

Produit scalaire requête x passages

Agrégation

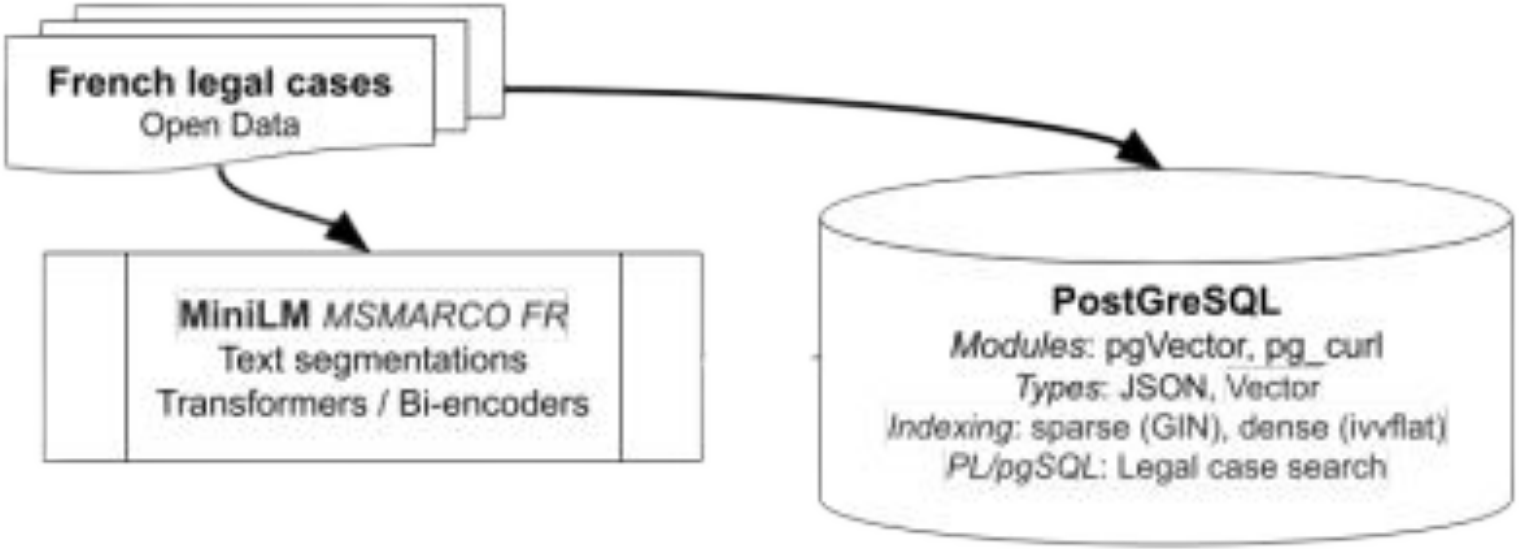
Recherche de passages

Similarité avec la requête

Recherche de Jurisprudence

Contexte instance

Expérimentation



Modèle relationnel
Pas que relationnel

Documents
- Id
- JSON

Extraits
- Id
- Id_doc
- position
- texte

Vecteurs
- id_extrait
- Modèle
- vecteur

Catégories
- Id_extrait
- catégorie

```
SELECT
J.id AS id_dec,
        R.line AS line,
        R.score AS score,
        J.json->'metadata'->>'date' AS date,
        J.json->'metadata'->>'short_title' AS title,
        J.json->'raw_text'->>(line-1) AS passage
    FROM
        (SELECT id_dec, sum(ip) as score, line
                FROM lib_knn(getemb(%s),%s)
                GROUP BY id_dec, line
                ORDER BY score) AS R,
        judilibre_json AS J
    WHERE R.id_dec=J.id;
```

Exemple de fonction

```
-- ### knn judilibre vectors from vector
CREATE OR REPLACE FUNCTION lib_knn(v_query vector(384), nb numeric)
      RETURNS
TABLE(id_dec character varying(100), chunk int, ip float, passage text, line int)
      LANGUAGE plpgsql
      AS $$
      BEGIN
       SET LOCAL ivfflat.probes = 65;
       SET LOCAL enable_seqscan = off;
       SET LOCAL min_parallel_table_scan_size = 1;
       SET LOCAL parallel_setup_cost = 1;
       RETURN QUERY
       SELECT J.id_dec, J.chunk, (J.embedding <#> v_query) AS ip, J.passage, J.line
       FROM judilibre_v AS J
       ORDER BY ip LIMIT nb;
      END;
      $$
      ;
```

Évaluation

Systèmes :

- **0 [CLS]** biencoder-all-MiniLM-L12-v2-mmarcoFR (https://huggingface.co/antoinelouis)
- **1 Booléen** avec index généralisé (PostGreSQL)
- **2 mean pooling** biencoder-mMiniLMv2-L12-mmarcoFR (https://huggingface.co/antoinelouis)

Site: https://guacamole.univ-avignon.fr/jplab/

API:
https://guacamole.univ-avignon.fr/jpvir_test?corpus=0&phrase=contestation%20amende%20pour%20stationnement&length=100

50 requêtes moteur de recherche Juri'Predis

- CONTRAVENTION ET COMPETENCE POLICE MUNICIPALE
- peut-on couper l'électricité chez un occupant sans droit ni titre
- coupure d'électricité chez un occupant sans droit ni titre
- doutes raisonnables volonté quitter état membre visa mariage
- doutes raisonnables volonté quitter territoire état membre visa
- demande de libération du logement de fonction pour loger le successeur

Experimentation

https://guacamole.univ-avignon.fr/jplab/

Id: JURITEXT63c10971bf9fd47c90a13c0d

Titre: CA Paris, 12 janv. 2023, RG n° 21/19806
Date: 2023-01-12
Passage: Les stipulations du contrat de cession d'actions prévoient donc expressément que pour être valide, une réclamation doit préciser le fondement, l'acte ou l'événement sur lequel elle est fondée, indiquer le montant ou une estimation du montant réclamé et comporter les documents qui permettent d'établir l'existence de la perte.
Ligne: 167
Score: -28.410415649414062

☆☆☆☆☆

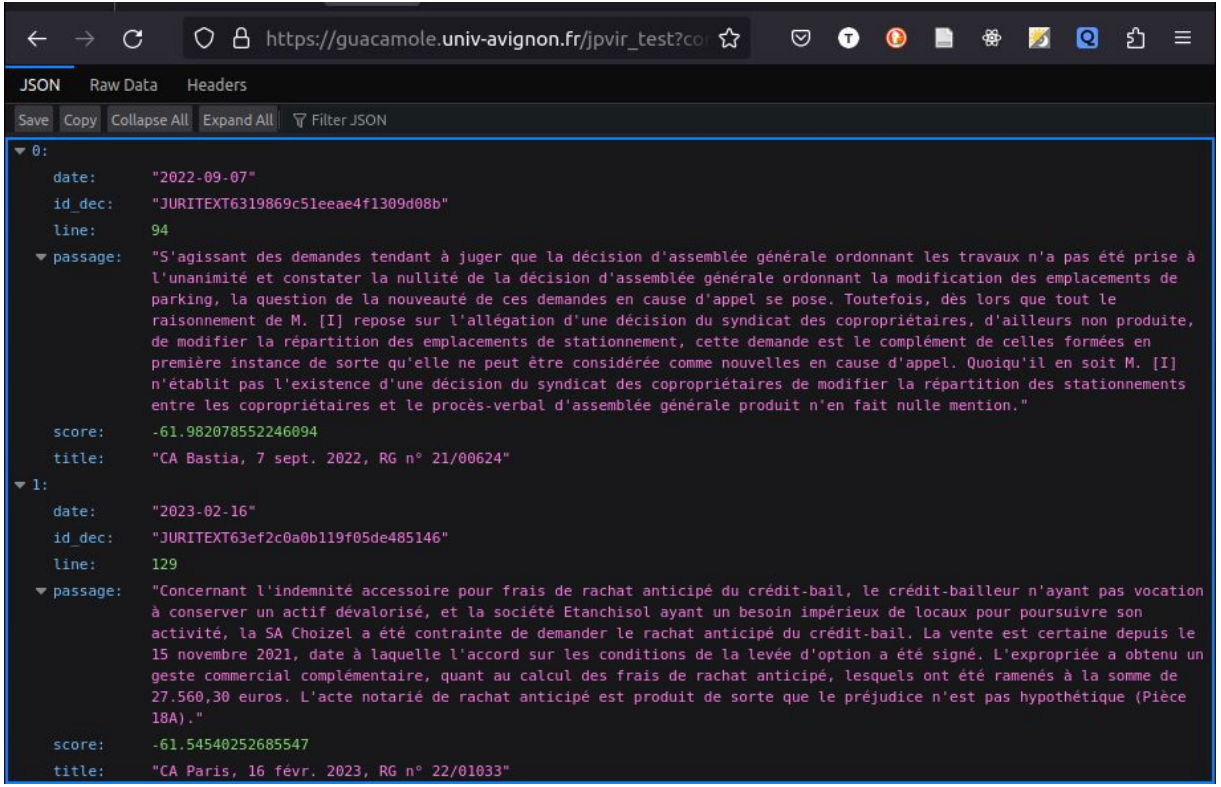
Id: JURITEXT64019dd4546e3305deed5bca

Titre: CA Aix-en-Provence, 2 mars 2023, RG n° 21/18317
Date: 2023-03-02
Passage: Le dommage imminent s'entend du dommage qui n'est pas encore réalisé mais qui se produira sûrement si la situation présente doit perdurer. Il s'ensuit que, pour que la mesure sollicitée soit prononcée, il doit nécessairement être constaté, avec l'évidence qui s'impose à la juridiction des référés, l'imminence d'un dommage, d'un préjudice ou la méconnaissance d'un droit, sur le point de se réaliser et dont la survenance et la réalité sont certaines ; un dommage purement éventuel ne saurait être retenu pour fonder l'intervention du juge des référés.
Ligne: 83
Score: -0.22006282210350037

☆☆☆☆☆

Experimentation

https://guacamole.univ-avignon.fr/jpvir_test?corpus=0&phrase=contestation%20amende%20pour%20stationnement&length=100



Evaluation

Table 1: Evaluation based on expert qrels . * indicates highly significant differences (p < 0.01, t-test)**

System	Description	MAP	P[5]	P[10]	Recip_Rank	IPRec[0]	IPRec[10]	NDCG
M	MeiliSearch Document	0.1344	0.3407	0.3259	0.5180	0.5654	0.3825	0.2456
B	Boolean Passage Search	0.0580	0.1556	0.1185	0.2509	0.2735	0.1840	0.1360
T	Sentence Transformer	**0.3252*****	**0.4815**	**0.4667**	**0.7727**	**0.7948**	**0.6258**	**0.4675*****
E	Sentence bi-Encoder	0.1214	0.2519	0.2444	0.4050	0.4662	0.3617	0.2476
TB	Transformer + Boolean	0.3099	0.4593	0.4481	0.7171	0.7636	0.6073	0.4647
EB	bi-Encoder + Boolean	0.1332	0.2667	0.2593	0.4174	0.4829	0.3770	0.2808
TE	Transformer + Encoder	0.2283	0.4593	0.4481				

Table 2: Evaluation based on keyword overlap

System	Description	MAP	P[5]	P[10]	Recip_Rank	IPRec[0]	IPRec[10]	NDCG
M	MeiliSearch Document	0.0027	0.0043	0.0036	0.7407	0.0241	0.0213	0.0043
T	Sentence Transformer	**0.0530**	**0.0645**	**0.1064**	**0.7727**	0.1887	**0.1351**	0.1115
E	Sentence bi-Encoder	0.0204	0.0511	0.0291	0.4050	0.1309	0.0523	0.0636
TE	Transformer + Encoder	0.2283	0.1021	0.0464	0.7171	**0.2026**	0.1124	**0.1148**

Requêtes et intensions

clause de non concurrence agent commercial
- clause concurrence agent commercial (42)

caution disproportion manifeste 12000 15000 aucun patrimoine
- caution disproportion aucun patrimoine (8)

référé 521-3 écoulement des eaux et fuite canalisatio
- eaux fuite canalisation (97)

avenant compromis de vente droit de rétractation
- compromis vente rétractation (16)

harcèlement moral critiques
- harcèlement moral critiques (56)

chevauchement deux arrêts de travail
- chevauchement arrêt travail (2)

conduite sous l’empire d’un état alcoolique procédure de contrôle et vices de procédure
- conduite alcool (97)

COMMISSION DISCIPLINE BACCALAUREAT SANCTION FRAUDE BAC
- commission BAC (5)

Ressources sur https://www.madics.fr/actions/simpletext/

Contener image debian autonome (30 Go)

- *Logiciels* : PostGreSQL + pg_vector + services mini LM + R + Rstudio + Shiny + Rollama
- *Corpus*: DBLP (< 2019) + JudiLibre (< 2019)
- *Données*: SimpleTest task1 2024 + evaluations Master Gouvernance des Données avignon
- *A venir*: collaboration NHS Scotland

Conclusions et perspectives

- Les mini LMs ont permis l'intégration dans les SGBDR de systèmes RI efficaces.
- Les LLMs internes peinent à se substituer aux annotateurs humains et … les externes non plus à un coût raisonnable.
- Les SBGDR permettent de gérer les documents comme de larges ensembles de vecteurs à l'échelle de l'open data juridique.
- Perspectives d'applications à la gestion de l'information au sein de la NHS Scotland.
- Comment approcher la double nature discrète/continue des mots (graphie/vecteur) ?