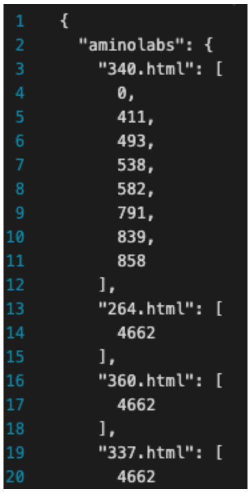
Assignment 2 Retrieval Models & PageRank Report

CS 4250 Web Search and Recommender Systems

Devin Khun, Daniel Ho, Caden Minniefield, Tony Swank, and Thet Wai, Bryan Castaneda

1 Crawling And Indexing

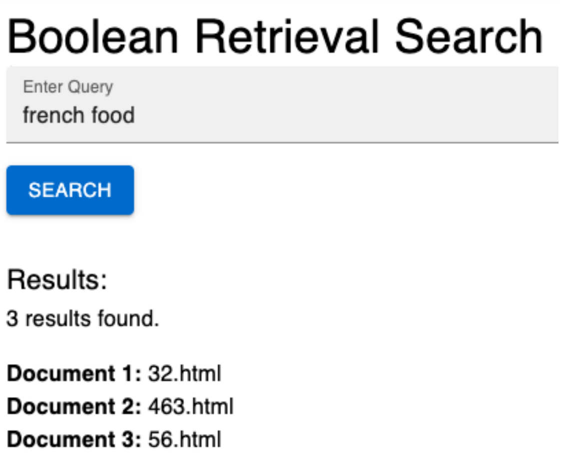
For crawling and indexing 500 web pages, we implemented a web crawler similar to Assignment 1. However, we also implemented an inverted index, which allows for fast and efficient search through documents. In our implementation, we pass in the HTML documents from a repository folder and perform word extraction on each document. For each document, the program extracts the text and breaks it into individual words (converted to lowercase). The program then records each word's position within the document. Then the program creates the index and is saved as a JSON file. As an example, the inverted index of our program for the word "aminolabs". The JSON is formatted in the following hierarchy term: document name: position. This would mean that the term appears in the HTML document at the given position. The following image is a snippet of our inverted index that shows a list of documents and the position the term appears in.



```
{
  "aminolabs": {
    "340.html": [
      0,
      411,
      493,
      538,
      582,
      791,
      839,
      858
    ],
    "264.html": [
      4662
    ],
    "360.html": [
      4662
    ],
    "337.html": [
      4662
```

2 Simple Boolean Retrieval

We implemented a simple Boolean Retrieval system by loading the inverted index from a JSON, which allows us to map words to the documents where they appear. We also allow users to implement a search query for all relevant documents. Our implementation implements a simple Boolean AND search on the inverted index provided. For each unique term in the collection, the search program maintains a list of documents where the term appears. We start with the document list of the first query word the user enters and checks each query term iteratively. For example, in the search, the query “french food” provides a list of three resulting documents that contain that term.



Boolean Retrieval Search

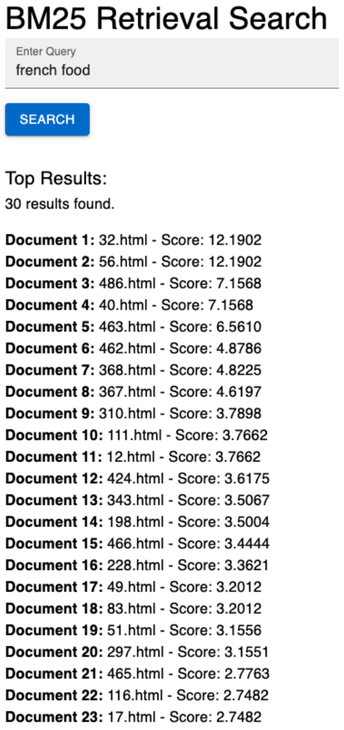
Enter Query
french food

Results:
3 results found.

Document 1: 32.html
Document 2: 463.html
Document 3: 56.html

3 Improved Retrieval: BM25

In our implementation of BM25, we implemented a system to get the inverted index from a JSON file, then calculate the essential stats like document length, computing IDf for each term. Each score is then re-ranked against each other to score against the user query. Results are then returned by these scores. This builds the inverted mapping of words to documents as well as list the positions in order to search through the documents and allow users to enter the search query for their word. In our implementation, we set out k1, the saturation effect to 1.5 in order to prevent common words from occurring in the relevance calculation. The document length normalization is set to 0.75, which means that a longer document will not have an advantage over others, as they would have more keywords in the document. The score output for each query indicates documents that contain more instances of the term, where a higher score indicates a more relevant document in relation to the query. This retrieval model takes account into the term frequency and the inverse frequency in relation to the entire collection



BM25 Retrieval Search

french food

Top Results:

30 results found.

Document 1: 32.html - Score: 12.1902
Document 2: 56.html - Score: 12.1902
Document 3: 486.html - Score: 7.1568
Document 4: 40.html - Score: 7.1568
Document 5: 463.html - Score: 6.5610
Document 6: 462.html - Score: 4.8786
Document 7: 368.html - Score: 4.8225
Document 8: 367.html - Score: 4.6197
Document 9: 310.html - Score: 3.7898
Document 10: 111.html - Score: 3.7662
Document 11: 12.html - Score: 3.7662
Document 12: 424.html - Score: 3.6175
Document 13: 343.html - Score: 3.5067
Document 14: 198.html - Score: 3.5004
Document 15: 466.html - Score: 3.4444
Document 16: 228.html - Score: 3.3621
Document 17: 49.html - Score: 3.2012
Document 18: 83.html - Score: 3.2012
Document 19: 51.html - Score: 3.1556
Document 20: 297.html - Score: 3.1551
Document 21: 465.html - Score: 2.7763
Document 22: 116.html - Score: 2.7482
Document 23: 17.html - Score: 2.7482

4 PageRank

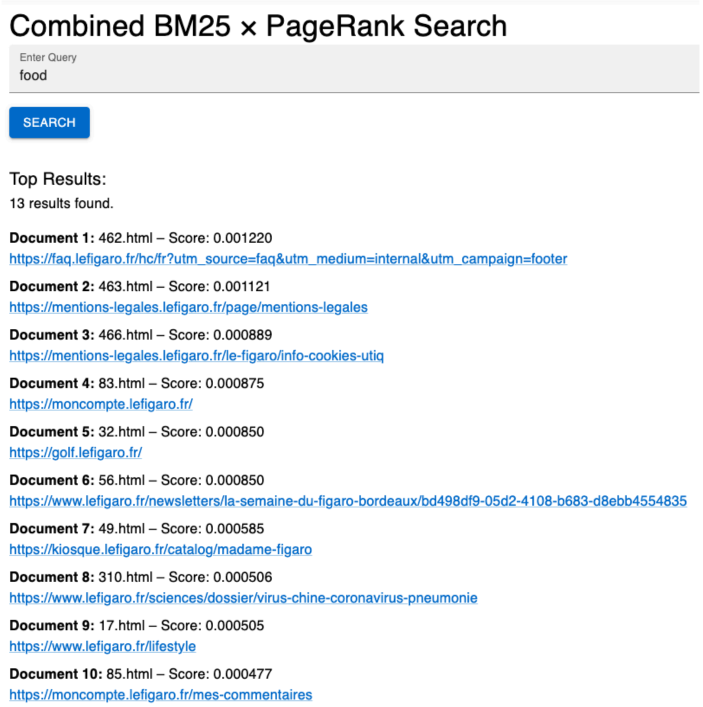
For PageRank, we start by stripping out the links in each of the crawled websites. This is done by taking the links with the href tags, and we found a lot of links. For the ranking, we decided to include the links that were also linking to outside links that are not within the pages we crawled, and it makes sense with the model because the score should be given less and less if the page points to a lot more pages, even if it is not within the pages we crawled. Additionally, another thing we had to do was map between our downloaded web pages and their respective original URL links. The way our web crawler from Assignment 1 worked was that it outputted in multiple HTML pages with a numeric counter at the start as the name such as 0.html, 1.hml, and so on. So, if we didn't map it, we wouldn't be able to give scores to proper documents and also get outside documents that are not the pages we crawled.

As for the page rank itself, we used the NetworkX library, which provided a function to run PageRank designed to rank web pages. We set the maximum iteration to 100, and we could even set the random surfer model, but in our case, we decided not to include it for simplicity's sake.

One of the most common urls that appeared was "cloudflare.com/5xx-error-landing". We believe that CloudFlare identified our web crawler as a bot or potential attacker due to the web crawling each page in rapid succession.

5 Combine PageRank And BM25

We found that combining PageRank and BM25 leads to much better results. Since PageRank's main goal is to prioritize credibility and BM25's main goal is to provide the most precise searches for specific terms, we can combine the two models to make the scores even more accurate. Combining the two is relatively simple; with the two scores of a document, we can multiply the scores together, which creates a new combined score. This new combined score can be used to compare against the others in this new combined category, with the new, most credible and accurate results having the greatest value. There are some caveats with this method, the main being that if there is a great difference in the two scores, there can be less precise results, ie. False positives/negatives, where a document that would be the best choice now is no longer that, because of the way PageRank works. This only matters with very niche topics, since with a larger sample size, the results become more accurate.



Combined BM25 × PageRank Search

Enter Query: food

Top Results:

13 results found.

Document 1: 462.html – Score: 0.001220
https://faq.lefigaro.fr/hc/fr?utm_source=faq&utm_medium=internal&utm_campaign=footer

Document 2: 463.html – Score: 0.001121
https://mentions-legales.lefigaro.fr/page/mentions-legales

Document 3: 466.html – Score: 0.000889
https://mentions-legales.lefigaro.fr/le-figaro/info-cookies-utiq

Document 4: 83.html – Score: 0.000875
https://moncompte.lefigaro.fr/

Document 5: 32.html – Score: 0.000850
https://golf.lefigaro.fr/

Document 6: 56.html – Score: 0.000850
https://www.lefigaro.fr/newsletters/la-semaine-du-figaro-bordeaux/bd498df9-05d2-4108-b683-d8ebb4554835

Document 7: 49.html – Score: 0.000585
https://kiosque.lefigaro.fr/catalog/madame-figaro

Document 8: 310.html – Score: 0.000506
https://www.lefigaro.fr/sciences/dossier/virus-chine-coronavirus-pneumonie

Document 9: 17.html – Score: 0.000505
https://www.lefigaro.fr/lifestyle

Document 10: 85.html – Score: 0.000477
https://moncompte.lefigaro.fr/mes-commentaires

6 Conclusions

In this project, we implemented multiple information retrieval techniques including Boolean Retrieval, BM25 ranking, and a hybrid model combining BM25 with PageRank. We built a custom web crawler, generated an inverted index, calculated PageRank scores from hyperlink structures, and developed a full-stack simple search engine capable of handling ranked queries. The hybrid model enhances retrieval quality by combining textual relevance with link-based authority, resulting in more meaningful rankings. Feel free to interact with the retrieval system through the web application. The live web application is hosted at: https://cs4250-assignment2.netlify.app. This provides a convenient way to experiment with the individual retrieval models and the combined BM25 x PageRank ranking directly by inputting a query.

7 Work Contributions

Devin Khun: Served as the project manager and worked on the retrieval modeling system to retrieve the relevant web pages for each query. I also included a simple web application for the user to input a query.

Thet Wai: Developed the page ranking.

Tony Swank: Helped write a report.

Daniel Ho: Wrote the report and BM25/Pagerank ranking combination.

Caden Minniefield: Helped with web crawler.

Bryan Castaneda Mayorga: Helped write the report.

REFERENCES

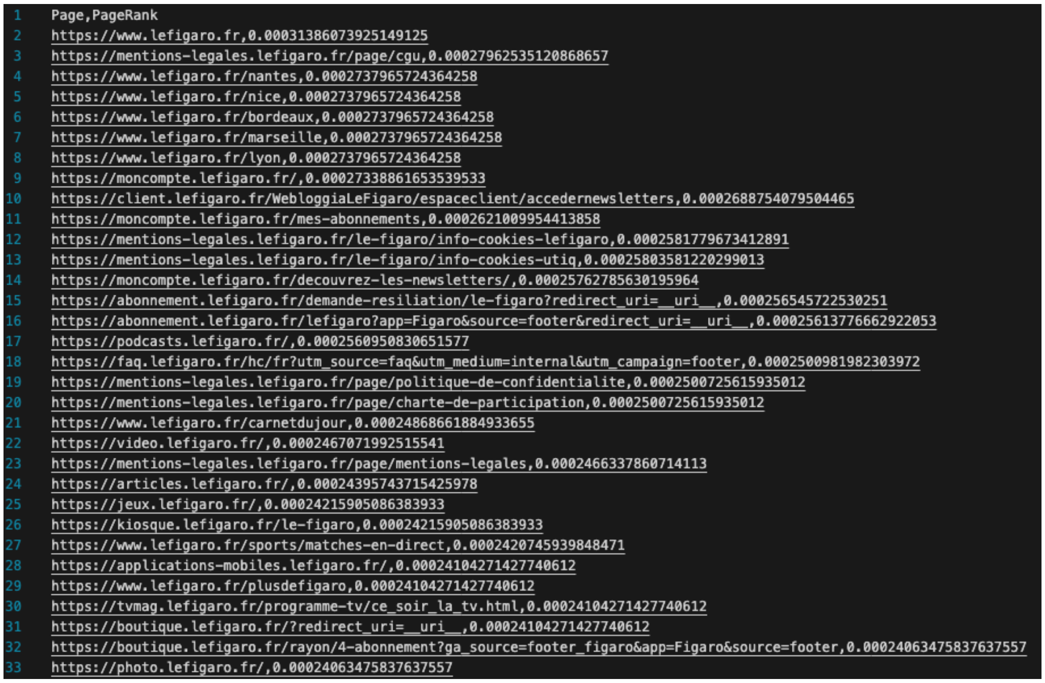
[1] https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.link_analysis.pagerank_alg.pagerank.html

[2] Code: https://github.com/devingineer/CS4250-Assignments

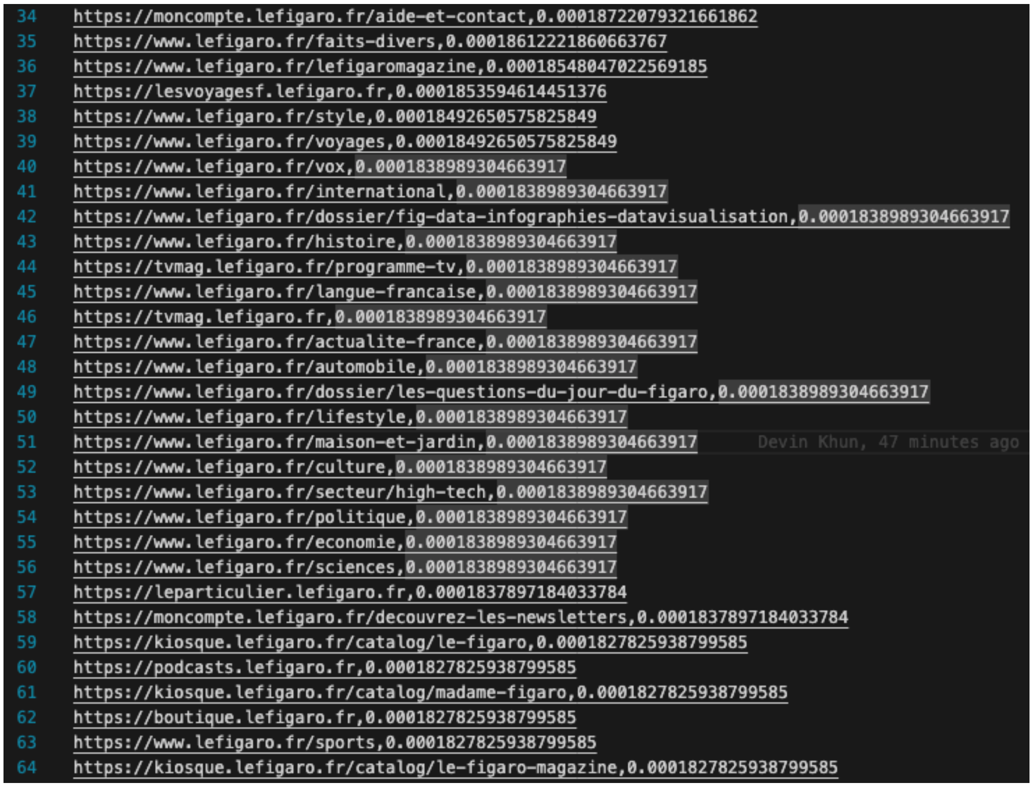
[3] Web Application: https://cs4250-assignment2.netlify.app

A APPENDICES

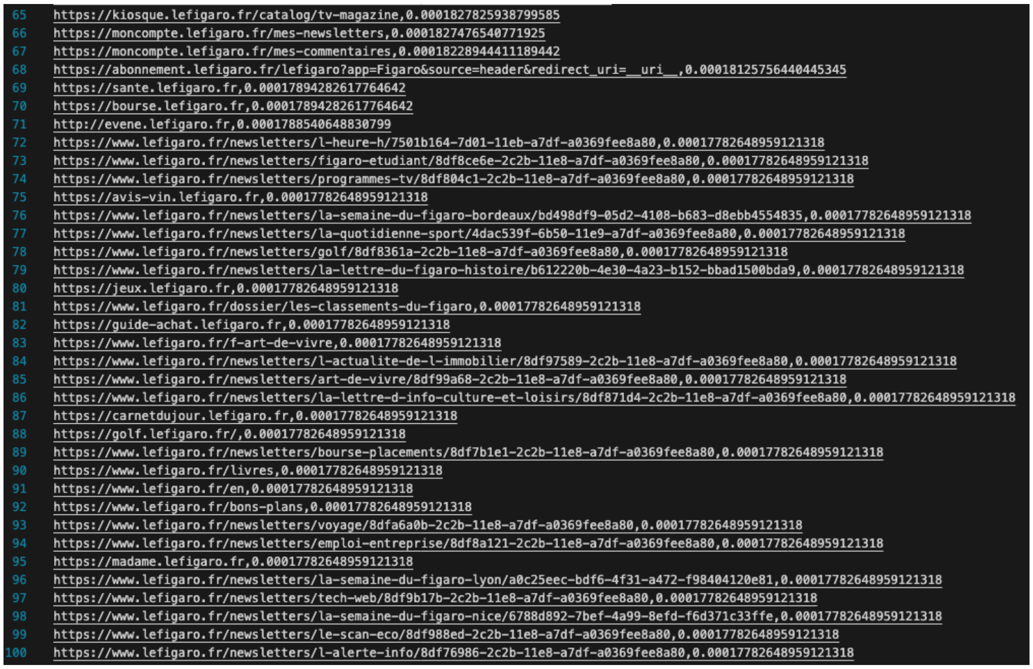
A.1 Top 100 Crawled Pages from PageRank



```
Page,PageRank
https://www.lefigaro.fr,0.00031386073925149125
https://mentions-legales.lefigaro.fr/page/cgu,0.00027962535120868657
https://www.lefigaro.fr/nantes,0.0002737965724364258
https://www.lefigaro.fr/nice,0.0002737965724364258
https://www.lefigaro.fr/bordeaux,0.0002737965724364258
https://www.lefigaro.fr/marseille,0.0002737965724364258
https://www.lefigaro.fr/lyon,0.0002737965724364258
https://moncompte.lefigaro.fr/,0.00027338861653539533
https://client.lefigaro.fr/WebloggiaLeFigaro/espaceclient/accedernewsletters,0.0002688754079504465
https://moncompte.lefigaro.fr/mes-abonnements,0.0002621009954413858
https://mentions-legales.lefigaro.fr/le-figaro/info-cookies-lefigaro,0.0002581779673412891
https://mentions-legales.lefigaro.fr/le-figaro/info-cookies-utiq,0.00025803581220299013
https://moncompte.lefigaro.fr/decouvrez-les-newsletters/,0.00025762785630195964
https://abonnement.lefigaro.fr/demande-resiliation/le-figaro?redirect_uri=__uri__,0.000256545722530251
https://abonnement.lefigaro.fr/lefigaro?app=Figaro&source=footer&redirect_uri=__uri__,0.00025613776662922053
https://podcasts.lefigaro.fr/,0.0002560950830651577
https://faq.lefigaro.fr/hc/fr?utm_source=faq&utm_medium=internal&utm_campaign=footer,0.0002500981982303972
https://mentions-legales.lefigaro.fr/page/politique-de-confidentialite,0.0002500725615935012
https://mentions-legales.lefigaro.fr/page/charte-de-participation,0.0002500725615935012
https://www.lefigaro.fr/carnetdujour,0.00024868661884933655
https://video.lefigaro.fr/,0.0002467071992515541
https://mentions-legales.lefigaro.fr/page/mentions-legales,0.0002466337860714113
https://articles.lefigaro.fr/,0.00024395743715425978
https://jeux.lefigaro.fr/,0.00024215905086383933
https://kiosque.lefigaro.fr/le-figaro,0.00024215905086383933
https://www.lefigaro.fr/sports/matches-en-direct,0.0002420745939848471
https://applications-mobiles.lefigaro.fr/,0.00024104271427740612
https://www.lefigaro.fr/plusdefigaro,0.00024104271427740612
https://tvmag.lefigaro.fr/programme-tv/ce_soir_la_tv.html,0.00024104271427740612
https://boutique.lefigaro.fr/?redirect_uri=__uri__,0.00024104271427740612
https://boutique.lefigaro.fr/rayon/4-abonnement?ga_source=footer_figaro&app=Figaro&source=footer,0.00024063475837637557
https://photo.lefigaro.fr/,0.00024063475837637557
```



```
https://moncompte.lefigaro.fr/aide-et-contact,0.00018722079321661862
https://www.lefigaro.fr/faits-divers,0.00018612221860663767
https://www.lefigaro.fr/lefigaromagazine,0.00018548047022569185
https://lesvoyagesf.lefigaro.fr,0.0001853594614451376
https://www.lefigaro.fr/style,0.00018492650575825849
https://www.lefigaro.fr/voyages,0.00018492650575825849
https://www.lefigaro.fr/vox,0.00018389893046663917
https://www.lefigaro.fr/international,0.00018389893046663917
https://www.lefigaro.fr/dossier/fig-data-infographies-datavisualisation,0.00018389893046663917
https://www.lefigaro.fr/histoire,0.00018389893046663917
https://tvmag.lefigaro.fr/programme-tv,0.00018389893046663917
https://www.lefigaro.fr/langue-francaise,0.00018389893046663917
https://tvmag.lefigaro.fr,0.00018389893046663917
https://www.lefigaro.fr/actualite-france,0.00018389893046663917
https://www.lefigaro.fr/automobile,0.00018389893046663917
https://www.lefigaro.fr/dossier/les-questions-du-jour-du-figaro,0.00018389893046663917
https://www.lefigaro.fr/lifestyle,0.00018389893046663917
https://www.lefigaro.fr/maison-et-jardin,0.00018389893046663917        Devin Khun, 47 minutes ago
https://www.lefigaro.fr/culture,0.00018389893046663917
https://www.lefigaro.fr/secteur/high-tech,0.00018389893046663917
https://www.lefigaro.fr/politique,0.00018389893046663917
https://www.lefigaro.fr/economie,0.00018389893046663917
https://www.lefigaro.fr/sciences,0.00018389893046663917
https://leparticulier.lefigaro.fr,0.0001837897184033784
https://moncompte.lefigaro.fr/decouvrez-les-newsletters,0.0001837897184033784
https://kiosque.lefigaro.fr/catalog/le-figaro,0.0001827825938799585
https://podcasts.lefigaro.fr,0.0001827825938799585
https://kiosque.lefigaro.fr/catalog/madame-figaro,0.0001827825938799585
https://boutique.lefigaro.fr,0.0001827825938799585
https://www.lefigaro.fr/sports,0.0001827825938799585
https://kiosque.lefigaro.fr/catalog/le-figaro-magazine,0.0001827825938799585
```



```
https://kiosque.lefigaro.fr/catalog/tv-magazine,0.0001827825938799585
https://moncompte.lefigaro.fr/mes-newsletters,0.0001827476540771925
https://moncompte.lefigaro.fr/mes-commentaires,0.00018228944411189442
https://abonnement.lefigaro.fr/lefigaro?app=Figaro&source=header&redirect_uri=__uri__,0.0001812575644045345
https://sante.lefigaro.fr,0.0001789428261776464 2
https://bourse.lefigaro.fr,0.00017894282617764642
http://evene.lefigaro.fr,0.0001788540648830799
https://www.lefigaro.fr/newsletters/l-heure-h/7501b164-7d01-11eb-a7df-a0369fee8a80,0.00017782648959121318
https://www.lefigaro.fr/newsletters/figaro-etudiant/8df8ce6e-2c2b-11e8-a7df-a0369fee8a80,0.00017782648959121318
https://www.lefigaro.fr/newsletters/programmes-tv/8df804c1-2c2b-11e8-a7df-a0369fee8a80,0.00017782648959121318
https://avis-vin.lefigaro.fr,0.00017782648959121318
https://www.lefigaro.fr/newsletters/la-semaine-du-figaro-bordeaux/bd498df9-05d2-4108-b683-d8ebb4554835,0.00017782648959121318
https://www.lefigaro.fr/newsletters/la-quotidienne-sport/4dac539f-6b50-11e9-a7df-a0369fee8a80,0.00017782648959121318
https://www.lefigaro.fr/newsletters/golf/8df8361a-2c2b-11e8-a7df-a0369fee8a80,0.00017782648959121318
https://www.lefigaro.fr/newsletters/la-lettre-du-figaro-histoire/b612220b-4e30-4a23-b152-bbad1500bda9,0.00017782648959121318
https://jeux.lefigaro.fr,0.00017782648959121318
https://www.lefigaro.fr/dossier/les-classements-du-figaro,0.00017782648959121318
https://guide-achat.lefigaro.fr,0.00017782648959121318
https://www.lefigaro.fr/f-art-de-vivre,0.00017782648959121318
https://www.lefigaro.fr/newsletters/l-actualite-de-l-immobilier/8df97589-2c2b-11e8-a7df-a0369fee8a80,0.00017782648959121318
https://www.lefigaro.fr/newsletters/art-de-vivre/8df99a68-2c2b-11e8-a7df-a0369fee8a80,0.00017782648959121318
https://www.lefigaro.fr/newsletters/la-lettre-d-info-culture-et-loisirs/8df871d4-2c2b-11e8-a7df-a0369fee8a80,0.00017782648959121318
https://carnetdujour.lefigaro.fr,0.00017782648959121318
https://golf.lefigaro.fr/,0.00017782648959121318
https://www.lefigaro.fr/newsletters/bourse-placements/8df7b1e1-2c2b-11e8-a7df-a0369fee8a80,0.00017782648959121318
https://www.lefigaro.fr/livres,0.00017782648959121318
https://www.lefigaro.fr/en,0.00017782648959121318
https://www.lefigaro.fr/bons-plans,0.00017782648959121318
https://www.lefigaro.fr/newsletters/voyage/8dfa6a0b-2c2b-11e8-a7df-a0369fee8a80,0.00017782648959121318
https://www.lefigaro.fr/newsletters/emploi-entreprise/8df8a121-2c2b-11e8-a7df-a0369fee8a80,0.00017782648959121318
https://madame.lefigaro.fr,0.00017782648959121318
https://www.lefigaro.fr/newsletters/la-semaine-du-figaro-lyon/a0c25eec-bdf6-4f31-a472-f98404120e81,0.00017782648959121318
https://www.lefigaro.fr/newsletters/tech-web/8df9b17b-2c2b-11e8-a7df-a0369fee8a80,0.00017782648959121318
https://www.lefigaro.fr/newsletters/la-semaine-du-figaro-nice/6788d892-7bef-4a99-8efd-f6d371c33ffe,0.00017782648959121318
https://www.lefigaro.fr/newsletters/le-scan-eco/8df988ed-2c2b-11e8-a7df-a0369fee8a80,0.00017782648959121318
https://www.lefigaro.fr/newsletters/l-alerte-info/8df76986-2c2b-11e8-a7df-a0369fee8a80,0.00017782648959121318
```