Open Data Beta

Where can I find pdf of articles in French, their titles and abstracts/summaries?

Asked 1 year, 2 months ago Modified 3 months ago Viewed 63 times

2

I have a task to find a title/make a summary of market research, in pdfs, in French. So I'm trying to find a dataset of the closest to it to create a model.

I was thinking of press articles in pdf as long as we have their title already extracted, and the pdf. I then thought about scientific articles. If you have better ideas, I'm interested.

For the moment I have found:

- A code on Kaggle that allows me to create a dataframe of scientific articles from Arxivx, with their abstracts and links to the pdfs. However it seems that it is only in English.

data-request nlp pdf french research-papers

Share Edit Follow Flag

edited Mar 24, 2023 at 1:49

Nicolas Raoul ♦

8,426 5 28 61

asked Mar 13, 2023 at 21:46

Revolucion for Monica

183 7

2 Answers

Sorted by: Highest score (default)

3

There's the **HAL scientific open archive**, which has a lot of content in French and various other languages. As far as I know, they don't offer a dump of their data, but they do have an API that you can request for free (with rate limitations per second, so be careful about that if you do not want to be blacklisted). Not all articles have PDF attached to them, but they still have tens of thousands of PDF files in French with titles and abstracts.

Here is **the documentation of the API** (in French, so probably not a problem for you as I guess that you're a French speaker, otherwise automatic translation tools may help). And here is an example of a json request for 30 articles in French with their titles, abstracts, and link to the PDF urls: **https://api.archives-ouvertes.fr/search/?q=language_s:fr&wt=json&fq=submitType_s:file&fq=docType_s:ART&fq=openAccess_bool:true&fl=files_s,abstract_s,title_s**
If you want to request more articles than that, they explain how to do it in the API documentation (look at the "Nombre de résultats" and "Pagination" sections).

Sometimes articles do not have abstract, sometimes they're translated in multiple languages, so it may require some work to get what you want from it, but hopefully it should be useful to you.

Share Edit Follow Flag

answered Mar 16, 2023 at 17:10

J-J-J
300 1 9

2

You can use openalex data dump. https://api.openalex.org/works?group_by=language There are `6843788` publications in French.

```
{
  "key": "fr",
  "key_display_name": "French",
  "count": 6843788
}
```

Share Edit Follow Flag

answered Feb 13 at 7:25

Dmitry Zagorulkin
121 2

Really nice resource (+1), however the metadata relative to language seem to be quite inaccurate. The first few articles I get when using this request to retrieve documents in French are in fact in English: api.openalex.org/works?filter=language:fr . I suspect the actual number of

documents in French might be largely overestimated, but it would require to examine a random sample of these results to get a fair estimation of the problem. Anyway, users with a use case like in the question will have to double-check the results returned by the API. – J-J-J Feb 13 at 20:54

Yep I know, it's really easy to filter all french metadata with `lang detect model` for instance fasttext.cc/blog/2017/10/02/blog-post.html – Dmitry Zagorulkin Feb 21 at 9:52