Requêtes Utilisées :

1-Récupération de tous les noms des datasets sur HuggingFace :

```
dataset_names = list_datasets()
```

2-Récupération des métadonnées d'un dataset :

```
url = f"https://huggingface.co/api/datasets/{name}"
response = session.get(url, params={"full": "True"})
```

params:
"name" correspond à l'id du dataset (auteur/nom_dataset)

results:


```
{'_id': '621ffdd236468d709f181d58', 'id': 'acronym_identification', 'sha': '15ef643450d589d5883e289ffadeb03563e80a9e', 'lastModified':
'2024-01-09T11:39:57.000Z', 'private': False, 'gated': False, 'disabled': False, 'description': '\n\t\n\t\t\n\t\n\t\n\t\tDataset Card for Acronym
Identification Dataset\n\t\n\n\n\t\n\t\t\n\t\n\t\n\t\tDataset Summary\n\t\n\nThis dataset contains the training, validation, and test data for the
Shared Task 1: Acronym Identification of the AAAI-21 Workshop on Scientific Document Understanding.\n\n\t\n\t\t\n\t\n\t\n\t\tSupported Tasks and
Leaderboards\n\t\n\nThe dataset supports an acronym-identification task, where the aim is to predic which tokens in a pre-tokenized sentence
correspond to acronyms. The dataset was released for a Shared… See the full description on the dataset page: https://huggingface
.co/datasets/acronym_identification.', 'paperswithcode_id': 'acronym-identification', 'downloads': 632, 'likes': 18, 'cardData':
{'annotations_creators': ['expert-generated'], 'language_creators': ['found'], 'language': ['en'], 'license': ['mit'], 'multilinguality':
['monolingual'], 'size_categories': ['10K<n<100K'], 'source_datasets': ['original'], 'task_categories': ['token-classification'], 'task_ids': [],
'paperswithcode_id': 'acronym-identification', 'pretty_name': 'Acronym Identification Dataset', 'tags': ['acronym-identification'], 'dataset_info':
{'features': [{'name': 'id', 'dtype': 'string'}, {'name': 'tokens', 'sequence': 'string'}, {'name': 'labels', 'sequence': {'class_label': {'names':
{'0': 'B-long', '1': 'B-short', '2': 'I-long', '3': 'I-short', '4': 'O'}}}}], 'splits': [{'name': 'train', 'num_bytes': 7792771, 'num_examples':
14006}, {'name': 'validation', 'num_bytes': 952689, 'num_examples': 1717}, {'name': 'test', 'num_bytes': 987712, 'num_examples': 1750}],
'download_size': 2071007, 'dataset_size': 9733172}, 'train-eval-index': [{'config': 'default', 'task': 'token-classification', 'task_id':
'entity_extraction', 'splits': {'eval_split': 'test'}, 'col_mapping': {'tokens': 'tokens', 'labels': 'tags'}}]}, 'createdAt': '2022-03-02T23:29:22
.000Z', 'arxiv': ['2010.14678']}
```


3-Récupération d'un datasetCard (Readme.md) d'un dataset :

```
hf_hub_download(repo_id=repo_id, filename="README.md", repo_type="dataset")
```

params:
"repo_id" correspond à l'id du dataset (auteur/nom_dataset)

results:

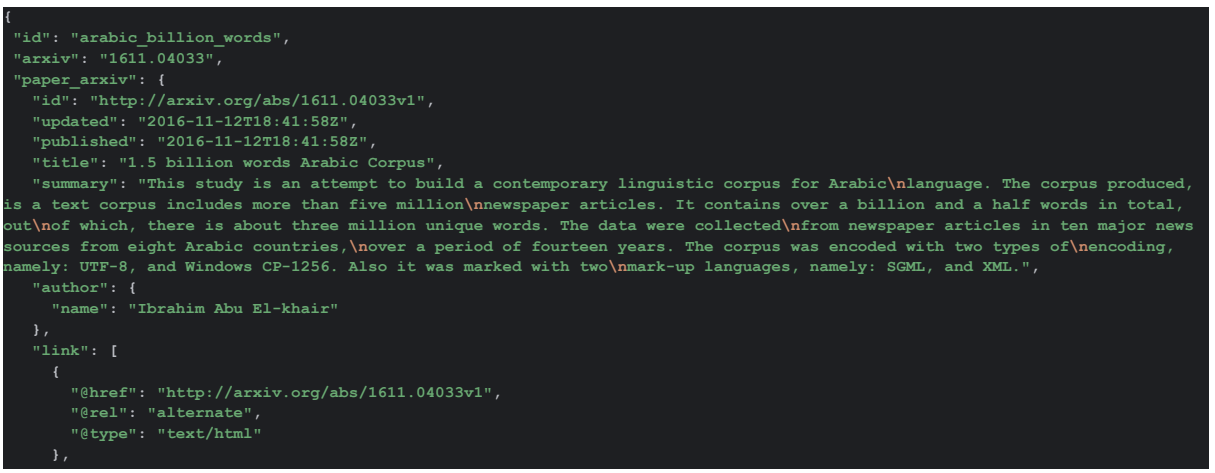
4-Récupération d'un papier arXiv :

Utilisation du numéro d'arxiv dans le datasetInfo

```
url = f"http://export.arxiv.org/api/query?max_results=1&search_query=all:{arxiv}"
data = urllib.request.urlopen(url)
```

params:
"arxiv" correspond au numéro d'arxiv du papier du dataset

results:



```
    {
      "@title": "pdf",
      "@href": "http://arxiv.org/pdf/1611.04033v1",
      "@rel": "related",
      "@type": "application/pdf"
    }
  ],
  "arxiv:primary_category": {
    "@xmlns:arxiv": "http://arxiv.org/schemas/atom",
    "@term": "cs.CL",
    "@scheme": "http://arxiv.org/schemas/atom"
  },
  "category": [
    {
      "@term": "cs.CL",
      "@scheme": "http://arxiv.org/schemas/atom"
    },
    {
      "@term": "cs.DL",
      "@scheme": "http://arxiv.org/schemas/atom"
    },
    {
      "@term": "cs.IR",
      "@scheme": "http://arxiv.org/schemas/atom"
    }
  ]
}
},
```


5-Récupération des informations de citations du datasets (API:Serpapi):

```
params = {
    "engine": "google_scholar",
    "num": "1",
    "q": "arXiv:1909.11942",
    "hl": "en",
    "api_key": "5c031d347fae722e2e6576c726ff739ccf6d55165001bb936c95cfa7e15a5994"
}

search = GoogleSearch(params)
```

params:
“q” correspond à l’information que l’on recherche (utilisation de l’ arxiv dans notre cas)
“api_key” correspond à la clé d’api qu’il faut acquérir

results:


```
{
  "search_metadata": {
    "id": "662cecfd3c3fb28aed7b91ce",
    "status": "Success",
    "json_endpoint": "https://serpapi.com/searches/dba6a72319338321/662cecfd3c3fb28aed7b91ce.json",
    "created_at": "2024-04-27 12:18:05 UTC",
    "processed_at": "2024-04-27 12:18:05 UTC",
    "google_scholar_url": "https://scholar.google.com/scholar?q=arXiv%3A1909.11942&hl=en&num=1",
    "raw_html_file": "https://serpapi.com/searches/dba6a72319338321/662cecfd3c3fb28aed7b91ce.html",
    "total_time_taken": 2.79
  },
  "search_parameters": {
    "engine": "google_scholar",
    "q": "arXiv:1909.11942",
    "hl": "en",
    "num": "1"
  },
  "search_information": {
    "organic_results_state": "Results for exact spelling",
    "query_displayed": "arXiv:1909.11942"
  },
  "profiles": {
    "link": "https://scholar.google.com/scholar?lookup=0&q=arXiv:1909.11942&hl=en&num=1&as_sdt=0,5",
    "serpapi_link": "https://serpapi.com/search.json?engine=google_scholar_profiles&hl=en&mauthors=arXiv%3A1909.11942"
  },
  "organic_results": [
    {
      "position": 0,
      "title": "Albert: A lite bert for self-supervised learning of language representations",
      "result_id": "wzWRMzbJrlsJ",
      "link": "https://arxiv.org/abs/1909.11942",
      "snippet": "Increasing model size when pretraining natural language representations often results in improved performance on downstream tasks. However, at some
point further model increases become harder due to GPU/TPU memory limitations and longer training times. To address these problems, we present two parameter-reduction
```

```
techniques to lower memory consumption and increase the training speed of BERT. Comprehensive empirical evidence shows that our proposed methods lead to models that
scale much better compared to the …",
      "publication_info": {
        "summary": "Z Lan, M Chen, S Goodman, K Gimpel… - arXiv preprint arXiv …, 2019 - arxiv.org",
        "authors": [
          {
            "name": "Z Lan",
            "link": "https://scholar.google.com/citations?user=tlDABkgAAAAJ&hl=en&num=1&oi=sra",
            "serpapi_scholar_link": "https://serpapi.com/search.json?author_id=tlDABkgAAAAJ&engine=google_scholar_author&hl=en",
            "author_id": "tlDABkgAAAAJ"
          },
          {
            "name": "M Chen",
            "link": "https://scholar.google.com/citations?user=aRncxakAAAAJ&hl=en&num=1&oi=sra",
            "serpapi_scholar_link": "https://serpapi.com/search.json?author_id=aRncxakAAAAJ&engine=google_scholar_author&hl=en",
            "author_id": "aRncxakAAAAJ"
          },
          {
            "name": "S Goodman",
            "link": "https://scholar.google.com/citations?user=xgZ6V-sAAAAJ&hl=en&num=1&oi=sra",
            "serpapi_scholar_link": "https://serpapi.com/search.json?author_id=xgZ6V-sAAAAJ&engine=google_scholar_author&hl=en",
            "author_id": "xgZ6V-sAAAAJ"
          },
          {
            "name": "K Gimpel",
            "link": "https://scholar.google.com/citations?user=kDHs7DYAAAAJ&hl=en&num=1&oi=sra",
            "serpapi_scholar_link": "https://serpapi.com/search.json?author_id=kDHs7DYAAAAJ&engine=google_scholar_author&hl=en",
            "author_id": "kDHs7DYAAAAJ"
          }
        ]
      },
      "resources": [
        {
          "title": "arxiv.org",
          "file_format": "PDF",
          "link": "https://arxiv.org/pdf/1909.11942.pdf%3E,"
        }
      ],
      "inline_links": {
        "serpapi_cite_link": "https://serpapi.com/search.json?engine=google_scholar_cite&hl=en&q=wzWRMzbJr1sJ",
        "cited_by": {
          "total": 6761,
          "link": "https://scholar.google.com/scholar?cites=6606720413006378435&as_sdt=2005&sciodt=0,5&hl=en&num=1",
          "cites_id": "6606720413006378435",
          "serpapi_scholar_link": "https://serpapi.com/search.json?as_sdt=2005&cites=6606720413006378435&engine=google_scholar&hl=en&num=1"
        },
        "related_pages_link": "https://scholar.google.com/scholar?q=related:wzWRMzbJr1sJ:scholar.google.com/&scioq=arXiv:1909.11942&hl=en&num=1&as_sdt=0,5",
        "serpapi_related_pages_link":
"https://serpapi.com/search.json?as_sdt=0%2C5&engine=google_scholar&hl=en&num=1&q=related%3AwzWRMzbJr1sJ%3Ascholar.google.com%2F",
        "versions": {
          "total": 10,
          "link": "https://scholar.google.com/scholar?cluster=6606720413006378435&hl=en&num=1&as_sdt=0,5",
          "cluster_id": "6606720413006378435",
          "serpapi_scholar_link": "https://serpapi.com/search.json?as_sdt=0%2C5&cluster=6606720413006378435&engine=google_scholar&hl=en&num=1"
        },
        "cached_page_link": "https://scholar.googleusercontent.com/scholar?q=cache:wzWRMzbJr1sJ:scholar.google.com/+arXiv:1909.11942&hl=en&num=1&as_sdt=0,5"
      }
    }
  ]
}
```


6-Récupération des informations de citations du datasets (API:Scholarly):

```
# Définir le terme de recherche
search_query = 'arXiv:1909.11942'
# Effectuer la recherche
search_results = scholarly.search_pubs(search_query)
```

results:


```
{'container_type': 'Publication', 'source': <PublicationSource.PUBLICATION_SEARCH_SNIPPET: 'PUBLICATION_SEARCH_SNIPPET'>, 'bib': {'title': 'Albert:
A lite bert for self-supervised learning of language representations', 'author': ['Z Lan', 'M Chen', 'S Goodman', 'K Gimpel'], 'pub_year': '2019',
'venue': 'arXiv preprint arXiv …', 'abstract': 'Increasing model size when pretraining natural language representations often results in improved
performance on downstream tasks. However, at some point further model increases become harder due to GPU/TPU memory limitations and longer training
times. To address these problems, we present two parameter-reduction techniques to lower memory consumption and increase the training speed of BERT.
 Comprehensive empirical evidence shows that our proposed methods lead to models that scale much better compared to the'}, 'filled': False,
'gsrank': 1, 'pub_url': 'https://arxiv.org/abs/1909.11942', 'author_id': ['tlDABkgAAAAJ', 'aRncxakAAAAJ', 'xgZ6V-sAAAAJ', 'kDHs7DYAAAAJ'],
'url_scholarbib': '/scholar?hl=en&q=info:wzWRMzbJr1sJ:scholar.google.com/&output=cite&scirp=0&hl=en', 'url_add_sclib':
'/citations?hl=en&xsrf=&continue=/scholar%3Fq%3DarXiv:1909.11942%26hl%3Den%26as_sdt%3D0,
33&citilm=1&update_op=library_add&info=wzWRMzbJr1sJ&ei=pqgrZvjdE5SCy9YP29Cc0AY&json=', 'num_citations': 6757, 'citedby_url':
'/scholar?cites=6606720413006378435&as_sdt=5,33&sciodt=0,33&hl=en', 'url_related_articles': '/scholar?q=related:wzWRMzbJr1sJ:scholar.google
.com/&scioq=arXiv:1909.11942&hl=en&as_sdt=0,33', 'eprint_url': 'https://arxiv.org/pdf/1909.11942.pdf%3E,'}
```

7-Récupération des papiers de recherche qui citent le datasets (API:Scholarly):

```
citedbyUrl = "/scholar?cites=6606720413006378435&as_sdt=2005&sciodt=0,5&hl=en"
cites_id_match = re.search(r'cites=(\d+)', url)
results = scholarly.search_citedby(cites_id_match.group(1))
```

params:
“citedbyUrl” correspond à l’url de citation récupéré avec Scholarly
“cites_id_match.group(1)” correspond à l’id de citation : 6606720413006378435

results**:**

Une liste de publications comme le résultat juste au dessus

8-Récupération des descriptions contenu dans les datasetCards(Readme.md) :

Utilisation de mots clés présent dans les titres pour récupérer les descriptions correspondantes

mots clés **:**

```
description_keywords =\
["Description", "description", "Summary", "summary", "Detail", "detail", "Dataset", "dataset"]
```

matching **:** 40298/90540 datasets, soit **44,5%**