SukjinMun /
Scholarly_Paper_Crawler

Code Issues Pull requests Actions Projects Security Insights

An automated scholarly literature pipeline that systematically searches, downloads, and analyzes academic papers while extracting key scientific parameters and organizing research data into structured formats for research purposes.

MIT license

3 stars 1 fork 1 watching Branches Activity
Tags

Public repository

.. 1 Branch 0 Tags Go to file t Go to file Add file Code

SukjinMun Update README.md c9bda4f · 4 months ago

arXiv_xml	Create .gitkeep	7 months ago
csv_files	Create .gitkeep	7 months ago
html_parsing	Create .gitkeep	7 months ago
pdf_files	Create .gitkeep	7 months ago
pdf_first_100_sentences	Create .gitkeep	7 months ago
scripts	Delete scripts/.gitkeep	7 months ago
LICENSE	Update LICENSE	7 months ago
README.md	Update README.md	4 months ago
inputs.txt	Add files via upload	7 months ago
webcrawler_paper_search.bat	Add files via upload	7 months ago

Scholarly_Paper_Crawler

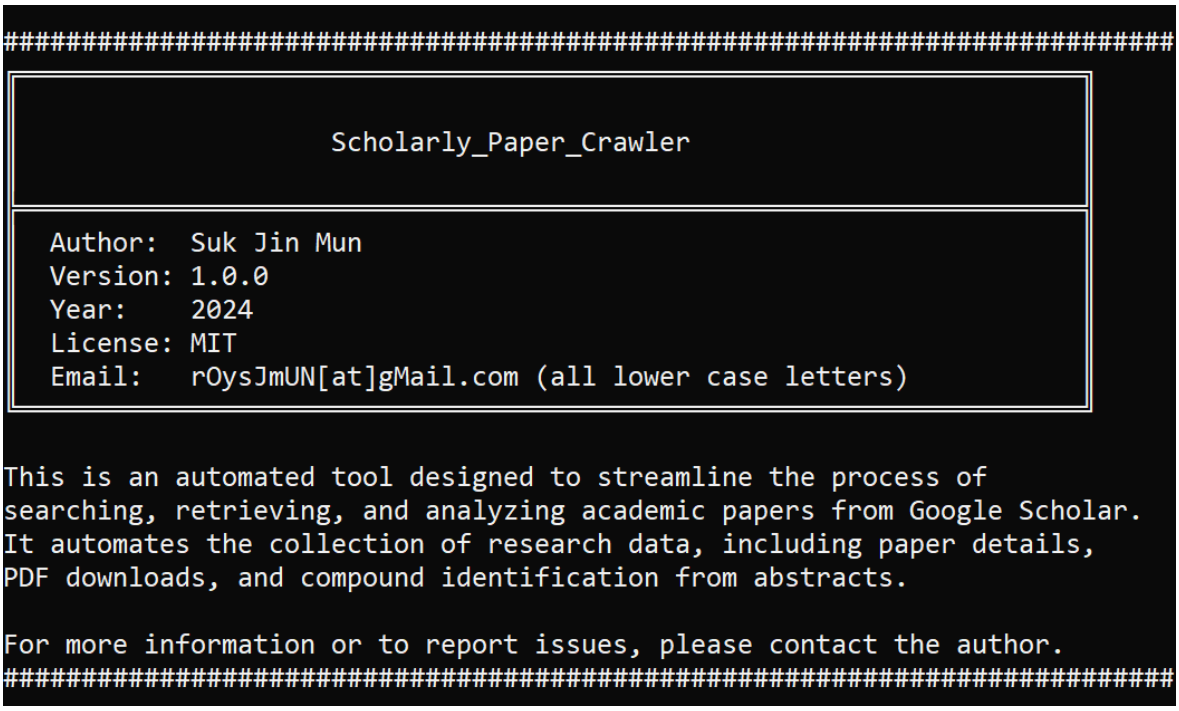


Table of contents

1.0 About

This pipeline automates the search and retrieval of research paper metadata based on specified keywords. It uses Google Scholar to gather information on research papers, including links and author details, and attempts to download PDFs from Sci-Hub. Designed primarily for studying and researching materials, this pipeline is especially useful for material science purposes. However, it is versatile enough to support other research areas as well, as it retrieves PDFs based on the keywords provided. At the end of the process, the pipeline organizes and stores the information in a separate XLSX file for easy reference.

2.0 Important Notice

This pipeline performs automated web scraping that may be flagged as bot-like behavior by academic servers. Running it on institutional networks (universities or research institutes) risks IP bans that could affect all users on that network and disrupt legitimate research activities. Instead, use a private environment (home network, personal VPN) to protect both your institution and ensure successful data collection. This approach prevents potential institutional policy violations, avoids administrative issues with IT departments, and maintains uninterrupted access to academic resources for all users. For optimal performance and security, it is strongly recommended to use a reliable VPN service with IP rotation capabilities - this provides an additional layer of protection for both your personal IP and API usage. This helps preventing potential blocks from Google Scholar and other academic servers while maintaining consistent access to the ScraperAPI service. Users must thoroughly review this documentation before using the pipeline, as a complete understanding is essential for successful operation of this pipeline.

3.0 Directory Structure

Directory Structure

```
root_directory/
├── inputs.txt                          # Configuration file for input parameters
├── webcrawler_paper_search.bat         # Batch file to execute the Python script
├── pdf_files/                          # Directory to store downloaded PDF files
│   └── [search_session_name]/          # Subfolder for each search session, named based on input
├── pdf_first_100_sentences/            # Directory to store first 100 sentences from PDFs
│   └── [search_session_name]/          # Subfolder containing text files of sentences
│       └── LastName[Year]_first100.txt  # Text files named after their source PDFs
├── csv_files/                          # Directory to store generated result files
│   └── [search_session_name].xlsx      # Excel file for each search session, named based on input
├── html_parsing/                       # Directory to store raw HTML from Google Scholar searches
│   └── [search_session_name]/          # Subfolder containing raw HTML files for each search result
├── arXiv_xml/                          # Directory to store XML responses from arXiv API
│   └── [search_session_name]/          # Subfolder containing XML files from arXiv queries
└── scripts/                            # Directory for Python scripts
    └── webcrawler_paper_search.py      # Main Python script to run the pipeline
```

3.1 html_parsing

- Stores raw HTML content from Google Scholar search results
- Each search session has its own subfolder
- Helps with debugging and analysis of search results
- Files are named as paper_N_raw.html where N is the result number

3.2 arXiv_xml

- Stores XML responses from arXiv API queries
- Each search session has its own subfolder
- Contains detailed metadata about papers found on arXiv
- Useful for tracking and debugging arXiv search results

3.3 pdf_first_100_sentences

- A directory that stores text files containing the first 100 sentences from each PDF
- Each search session has its own subfolder named after the session

- Text files are named to match their source PDFs (e.g., Duan2021_first100.txt)
- Used for quick content preview and text analysis without opening PDFs
- Files contain extracted and cleaned text from the beginning of each paper

4.0 Pipeline Manual

4.1 Setup

First, go to https://www.scraperapi.com/signup and create a free account. They offer 5,000 free API requests per month. After signing up, you will be directed to https://dashboard.scraperapi.com/. Retrieve your API Key in the API Key section. Then, Enter search parameters in the inputs.txt file. Format example:

```
0> Input API Key
'API KEY YOU OBTAINED FROM SCRAPERAPI.COM'

1> Specify the name for this search session (e.g., "1D_AFM_material_search"):
1D_AFM_material_search

2> Specify the maximum number of searches to perform for each keyword combination (e.g., "10"):
3

3> Provide keywords for the search (e.g., "1D antiferromagnetic chain system"):
one dimensional antiferromagnetic chain system, 1D antiferromagnetic chain system, quasi-1d AFM

4> Search for compound name information? (Y/N) (e.g., "Y"):
Y

5> Search for Spin information? (Y/N) (e.g., "Y"):
Y
```

Note: Processing 45 papers (15 keyword combinations with 3 searches each) typically takes approximately 5 hours due to mandatory waiting periods and rate limiting

4.2 Running the Pipeline

- Execute webcrawler_paper_search.bat to install required dependencies and launch the script automatically.
- The script reads from inputs.txt, performs the search on Google Scholar, attempts to download PDFs from Sci-Hub, and saves results into an Excel file within csv_files/.

4.3 Checking Output

- The Excel file will contain titles, authors, publication year, keywords, Google Scholar links, and PDF filenames.
- You can explore and analyze the results from the csv_files/ directory.
- Downloaded PDFs can be found in the pdf_files/[search_session_name]/ directory.

5.0 Features

5.1 Spin Detection

The pipeline includes automated detection of quantum mechanical spin values from papers. This feature:

- Extracts physically valid spin values following quantum mechanical principles
- Supports both integer spins (S=0, S=1, S=2) and half-integer spins (S=1/2, S=3/2, S=5/2)
- Validates spin values against quantum mechanical constraints:
  - Only accepts standard quantum spin values (0, 1/2, 1, 3/2, 2, 5/2)
  - Enforces proper denominator (only /2 for half-integer spins)
  - Rejects non-physical spin values
  - Limits maximum spin value to 2 (as per Standard Model constraints)
- Searches for spin values in both paper titles and PDF content
- Reports "No valid quantum spin value found" when no valid spin is detected

The spin values are extracted in formats:

- S = X/Y (with space)
- S=X/Y (without space)
- S=X (integer values) Where X and Y are integer numbers following quantum mechanical constraints.

5.2 Keyword Permutation

The pipeline implements comprehensive keyword combination searching that examines all possible permutations of the input keywords. This feature:

- Generates all possible combinations of input keywords, including individual terms and their permutations
- Searches each combination separately in Google Scholar
- Performs the specified number of searches (max_results) for EACH combination
- Maintains a wait period between combinations to comply with rate limits
- Automatically skips duplicate papers found across different combinations

For example, if keywords "A, B, C" are provided (line 3> of inputs.txt), the pipeline will search for:

1. Individual keywords: "A", "B", "C"
2. Pairs with all permutations: "A B", "A C", "B A", "B C", "C A", "C B"
3. Triples with all permutations: "A B C", "A C B", "B A C", "B C A", "C A B", "C B A"

The total number of combinations grows factorially with the number of keywords:

```
Number of combinations for different keyword counts:
2 keywords: 3 combinations
3 keywords: 15 combinations
4 keywords: 64 combinations
5 keywords: 325 combinations
6 keywords: 1956 combinations
```

Example of search behavior:

Input: 2> Specify the maximum number of searches to perform: 10 3> Provide keywords: A, B, C

Result:

- Will perform 10 searches for "A"
- Will perform 10 searches for "B"
- Will perform 10 searches for "C"
- Will perform 10 searches for "A B"
- Will perform 10 searches for "A C"
- Will perform 10 searches for "B A" ...and so on for all 15 combinations. In such case, a total of 15x10 papers should be processed.

Each combination is treated as a separate search query, with the specified maximum number of results collected for each. This ensures thorough coverage of the literature while avoiding duplicate entries. The script automatically handles:

- Tracking results per combination
- Removing duplicate papers across combinations
- Maintaining proper delays between searches
- Recording which search terms found each paper

6.0 Files Explained

6.1 inputs.txt

- This file contains the parameters for the search pipeline.
- Parameters include the search session name, the maximum number of search results to retrieve, and keywords.

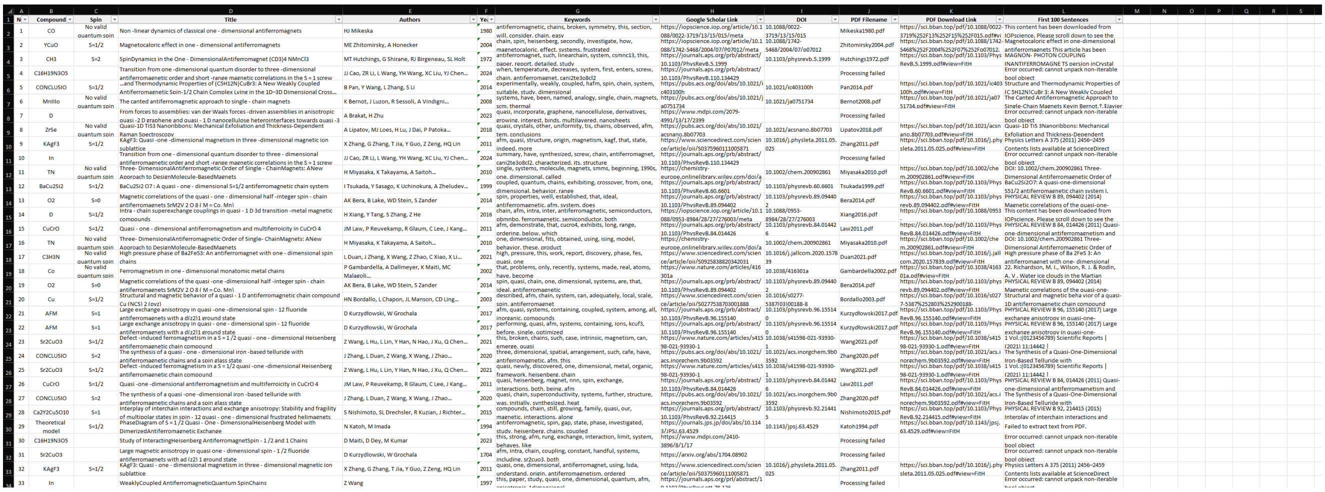
6.2 webcrawler_paper_search.bat

- A batch file that automates the execution of the webcrawler_paper_search.py script.
- It installs required Python packages and runs the main script.

6.3 pdf_files\

- A directory that stores PDFs associated with the search results.
- Each search session creates a subfolder based on the session name.
- The script attempts to download PDFs from Sci-Hub.
- PDFs are named using the format: LastName[Year].pdf (e.g., Duan2021.pdf)

6.4 csv_files\



N	Compound	Spin	Title	Authors	Year	Keywords	DOI	PDF Filename
1	CO	No valid quantum spin	Non -linear dynamics of classical one - dimensional antiferromagnets	HJ Mikeska	1980	antiferromagnetic, chains, broken, symmetry, this, section, will, consider, chain, easy	10.1088/0022-3719/13/15/015	Mikeska1980.pdf
2	YCuO	S=1/2	Magnetocaloric effect in one - dimensional antiferromagnets	ME Zhitomirsky, A Honecker	2004	chain, spin, heisenberg, secondly, investigate, how, magnetocaloric, effect, systems, frustrated	10.1088/1742-5468/2004/07/p07012	Zhitomirsky2004.pdf
3	CH3	S=2	SpinDynamics in the One - DimensionalAntiferromagnet (CD3)4 NMnCl3	MT Hutchings, G Shirane, RJ Birgeneau, SL Holt	1972	antiferromagnet, such, linearchain, system, csmnc13, this, paper, report, detailed, study	10.1103/physrevb.5.1999	Hutchings1972.pdf
4	C16H19N3O5		Transition from one -dimensional quantum disorder to three -dimensional antiferromagnetic order and short -range magnetic correlations in the S = 1 screw	JJ Cao, ZR Li, L Wang, YH Wang, XC Liu, YJ Chen...	2024	when, temperature, decreases, system, first, enters, screw, chain, antiferromagnet, cani2te3o8cl2		Processing failed
5	CONCLUSIO	S=1/2	...and Thermodynamic Properties of (C5H12N)CuBr3: A New Weakly Coupled Antiferromagnetic Spin-1/2 Chain Complex Lying in the 1D–3D Dimensional Cross...	B Pan, Y Wang, L Zhang, S Li	2014	experimentally, weakly, coupled, hafm, spin, chain, system, suitable, study, dimensional	10.1021/ic403100h	Pan2014.pdf
6	MnIIIo	No valid quantum spin	The canted antiferromagnetic approach to single - chain magnets	K Bernot, J Luzon, R Sessoli, A Vindigni...	2008	systems, have, been, named, analogy, single, chain, magnets, scm, thermal	10.1021/ja0751734	Bernot2008.pdf
7	D		From forces to assemblies: van der Waals forces -driven assemblies in anisotropic quasi -2 D graphene and quasi - 1 D nanocellulose heterointerfaces towards quasi -3	A Brakat, H Zhu	2023	quasi, incorporate, graphene, nanocellulose, derivatives, growing, interest, binds, multilayered, nanosheets		Processing failed
8	ZrSe	No valid quantum spin	Quasi-1D TiS3 Nanoribbons: Mechanical Exfoliation and Thickness-Dependent Raman Spectroscopy	A Lipatov, MJ Loes, H Lu, J Dai, P Patoka...	2018	quasi, crystals, other, uniformity, tis, chains, observed, afm, tem, conclusions	10.1021/acsnano.8b07703	Lipatov2018.pdf
9	KAgF3	S=1/2	KAgF3: Quasi -one -dimensional magnetism in three -dimensional magnetic ion sublattice	X Zhang, G Zhang, T Jia, Y Guo, Z Zeng, HQ Lin	2011	afm, quasi, structure, origin, magnetism, kagf, that, state, indeed, more	10.1016/j.physleta.2011.05.025	Zhang2011.pdf
10	In		Transition from one - dimensional quantum disorder to three - dimensional antiferromagnetic order and short -range magnetic correlations in the S = 1 screw	JJ Cao, ZR Li, L Wang, YH Wang, XC Liu, YJ Chen...	2024	summary, have, synthesized, screw, chain, antiferromagnet, cani2te3o8cl2, characterized, its, structure		Processing failed
11	TN	No valid quantum spin	Three- DimensionalAntiferromagnetic Order of Single - ChainMagnets: ANew Approach to DesignMolecule-BasedMagnets	H Miyasaka, K Takayama, A Saitoh...	2010	single, systems, molecule, magnets, smms, beginning, 1990s, one, dimensional, called	10.1002/chem.200902861	Miyasaka2010.pdf
12	BaCu2Si2	S=1/2	BaCu2Si2 O7 : A quasi - one - dimensional S=1/2 antiferromagnetic chain system	I Tsukada, Y Sasago, K Uchinokura, A Zheludev...	1999	coupled, quantum, chains, exhibiting, crossover, from, one, dimensional, behavior, range	10.1103/physrevb.60.6601	Tsukada1999.pdf
13	O2	S=0	Magnetic correlations of the quasi - one - dimensional half -integer spin - chain antiferromagnets SrM2V 2 O 8 (M = Co, Mn)	AK Bera, B Lake, WD Stein, S Zander	2014	spin, properties, well, established, that, ideal, antiferromagnetic, afm, system, does	10.1103/physrevb.89.094402	Bera2014.pdf
14	D	S=1/2	Intra - chain superexchange couplings in quasi - 1 D 3d transition -metal magnetic compounds	H Xiang, Y Tang, S Zhang, Z He	2016	chain, afm, intra, inter, antiferromagnetic, semiconductors, pbmnbo, ferromagnetic, semiconductor, both	10.1088/0953-8984/28/27/276003	Xiang2016.pdf
15	CuCrO	S=1/2	Quasi - one - dimensional antiferromagnetism and multiferroicity in CuCrO 4	JM Law, P Reuvekamp, R Glaum, C Lee, J Kang...	2011	afm, demonstrate, that, cucro4, exhibits, long, range, ordering, below, which	10.1103/physrevb.84.014426	Law2011.pdf
16	TN	No valid quantum spin	Three- DimensionalAntiferromagnetic Order of Single- ChainMagnets: ANew Approach to DesignMolecule-BasedMagnets	H Miyasaka, K Takayama, A Saitoh...	2010	one, dimensional, fits, obtained, using, ising, model, behavior, these, product	10.1002/chem.200902861	Miyasaka2010.pdf
17	C3H3N	No valid quantum spin	High pressure phase of Ba2FeS3: An antiferromagnet with one - dimensional spin chains	L Duan, J Zhang, X Wang, Z Zhao, C Xiao, X Li...	2021	high, pressure, this, work, report, discovery, phase, fes, quasi, one	10.1016/j.jallcom.2020.157839	Duan2021.pdf
18	Co	No valid quantum spin	Ferromagnetism in one - dimensional monatomic metal chains	P Gambardella, A Dallmeyer, K Maiti, MC Malagoli...	2002	that, problems, only, recently, systems, made, real, atoms, have, become	10.1038/416301a	Gambardella2002.pdf
19	O2	S=0	Magnetic correlations of the quasi -one -dimensional half -integer spin - chain antiferromagnets SrM2V 2 O 8 (M = Co, Mn)	AK Bera, B Lake, WD Stein, S Zander	2014	spin, quasi, chain, one, dimensional, systems, are, that, ideal, antiferromagnetic	10.1103/physrevb.89.094402	Bera2014.pdf
20	Cu	S=1/2	Structural and magnetic behavior of a quasi - 1 D antiferromagnetic chain compound Cu (NCS) 2 (pyz)	HN Bordallo, L Chapon, JL Manson, CD Ling...	2003	described, afm, chain, system, can, adequately, local, scale, spin, antiferromagnet	10.1016/s0277-5387(03)00188-8	Bordallo2003.pdf
21	AFM	S=1	Large exchange anisotropy in quasi -one -dimensional spin - 12 fluoride antiferromagnets with a d(z2)1 ground state	D Kurzydłowski, W Grochala	2017	afm, quasi, systems, containing, coupled, system, among, all, inorganic, compounds	10.1103/physrevb.96.155140	Kurzydłowski2017.pdf
22	AFM	S=1	Large exchange anisotropy in quasi - one - dimensional spin - 12 fluoride antiferromagnets with a d(z2)1 ground state	D Kurzydłowski, W Grochala	2017	performing, quasi, afm, systems, containing, ions, kcuf3, before, single, optimized	10.1103/physrevb.96.155140	Kurzydłowski2017.pdf
23	Sr2CuO3	S=1/2	Defect -induced ferromagnetism in a S = 1 /2 quasi - one - dimensional Heisenberg antiferromagnetic chain compound	Z Wang, L Hu, L Lin, Y Han, N Hao, J Xu, Q Chen...	2021	this, broken, chains, such, case, intrinsic, magnetism, can, emerge, quasi	10.1038/s41598-021-93930-1	Wang2021.pdf
24	CONCLUSIO	S=2	The synthesis of a quasi - one - dimensional iron -based telluride with antiferromagnetic chains and a spin glass state	J Zhang, L Duan, Z Wang, X Wang, J Zhao...	2020	three, dimensional, spatial, arrangement, such, cafe, have, antiferromagnetic, afm, this	10.1021/acs.inorgchem.9b03592	Zhang2020.pdf
25	Sr2CuO3	S=1/2	Defect -induced ferromagnetism in a S = 1/2 quasi -one -dimensional Heisenberg antiferromagnetic chain compound	Z Wang, L Hu, L Lin, Y Han, N Hao, J Xu, Q Chen...	2021	quasi, newly, discovered, one, dimensional, metal, organic, framework, heisenberg, chain	10.1038/s41598-021-93930-1	Wang2021.pdf
26	CuCrO	S=1/2	Quasi -one -dimensional antiferromagnetism and multiferroicity in CuCrO 4	JM Law, P Reuvekamp, R Glaum, C Lee, J Kang...	2011	quasi, heisenberg, magnet, nnn, spin, exchange, interactions, both, being, afm	10.1103/physrevb.84.014426	Law2011.pdf
27	CONCLUSIO	S=2	The synthesis of a quasi -one -dimensional iron -based telluride with antiferromagnetic chains and a spin glass state	J Zhang, L Duan, Z Wang, X Wang, J Zhao...	2020	quasi, chain, superconductivity, systems, further, structure, was, initially, synthesized, heat	10.1021/acs.inorgchem.9b03592	Zhang2020.pdf
28	Ca2Y2Cu5O10	S=1	Interplay of interchain interactions and exchange anisotropy: Stability and fragility of multipolar states in spin - 12 quasi - one - dimensional frustrated helimagnets	S Nishimoto, SL Drechsler, R Kuzian, J Richter...	2015	compounds, chain, still, growing, family, quasi, our, magnetic, interactions, along	10.1103/physrevb.92.214415	Nishimoto2015.pdf
29	Theoretical model	S=1/2	PhaseDiagram of S = 1 /2 Quasi - One - DimensionalHeisenberg Model with DimerizedAntiferromagnetic Exchange	N Katoh, M Imada	1994	antiferromagnetic, spin, gap, state, phase, investigated, study, heisenberg, chains, coupled	10.1143/jpsj.63.4529	Katoh1994.pdf
30	C16H19N3O5		Study of InteractingHeisenberg AntiferromagnetSpin - 1 /2 and 1 Chains	D Maiti, D Dey, M Kumar	2023	this, strong, afm, rung, exchange, interaction, limit, system, behaves, like		Processing failed
31	Sr2CuO3		Large magnetic anisotropy in quasi one - dimensional spin - 1 /2 fluoride antiferromagnets with ad (z2) 1 ground state	D Kurzydlowski, W Grochala	1704	afm, intra, chain, coupling, constant, handful, systems, including, sr2cuo3, both		Processing failed
32	KAgF3	S=1/2	KAgF3: Quasi - one - dimensional magnetism in three - dimensional magnetic ion sublattice	X Zhang, G Zhang, T Jia, Y Guo, Z Zeng, HQ Lin	2011	quasi, one, dimensional, antiferromagnet, using, lsda, understand, origin, antiferromagnetism, ordered	10.1016/j.physleta.2011.05.025	Zhang2011.pdf
33	In		WeaklyCoupled AntiferromagneticQuantum SpinChains	Z Wang	1997	this, paper, study, quasi, one, dimensional, quantum, afm, anisotropic, 1dimensional		Processing failed

- This directory contains Excel files generated from each search session.
- Each session creates a file named after the session with detailed metadata including:
  - Paper titles, authors, and year
  - Extracted compound information
  - Quantum spin values (following physical constraints)
  - Keywords and links
  - First 100 sentences from PDFs

6.5 scripts\

- The main Python script that handles reading inputs, querying Google Scholar, downloading PDFs from Sci-Hub, and

README MIT license

7.0 API Management

7.1 Free Trial Accounts

- Initial free credits: 5,000 requests
- Trial duration: 7 days from account creation
- Credits are one-time only and do not renew
- After 7 days or using all credits (whichever comes first):
  - Remaining credits expire
  - API key stops working
  - Must upgrade to paid plan to continue

7.2 Paid Accounts

- API keys remain active as long as account is in good standing
- Credits reset monthly based on subscription plan

- The key itself doesn't expire unless manually revoked

7.3 Usage Monitoring

1. Through ScraperAPI Dashboard (https://www.scraperapi.com/dashboard):
    - View current credit balance
    - Monitor daily/monthly usage statistics
    - Check plan status and renewal date
    - Track remaining trial days (for free accounts)
2. Through Console Messages: Common API Status Messages:

```
"Error in API request: 401 Client Error: Unauthorized"
  → API key is invalid or credits depleted
"Error in API request: 429 Too Many Requests"
  → Temporary rate limit, script will automatically retry
"Success"
  → API key is valid and has available credits
```

7.4 Best Practices

- Each Google Scholar search page uses 1 credit
- Typically 10 results per page
- Example: 100 papers might use 10-20 credits

For Free Trial:

- Plan searches carefully within 7-day window
- Start with smaller searches to test the system
- Save important searches for when familiar with tool

For All Users:

- Check credit balance before starting large search sessions
- Monitor console output for API-related error messages
- Keep track of credit usage for planning
- Consider upgrading if you need regular access
- Save API usage statistics for future planning

8.0 Notes

- The script saves raw HTML from Google Scholar searches for analysis
- ArXiv is used as an alternative source when Sci-Hub download fails
- The script implements exponential backoff and retry mechanisms for robust web scraping
- Added better error handling and logging for debugging purposes
- The script uses cloudscraper to bypass DDoS protection
- Downloads have mandatory delays (60 seconds minimum) between requests to comply with rate limits
- Spin values are validated against quantum mechanical principles before being included in the output
- Due to the factorial growth of combinations, users should be mindful when using more than 4 keywords, as this can lead to very long execution times. Each combination requires its own set of searches with appropriate waiting periods to avoid overloading the search servers.
- This project is licensed under the MIT License - see the LICENSE file for details.

9.0 Known Issues and Future Updates

Ver 1.0.0

Issues

- Current compound name and spin value extraction algorithms exhibit reduced accuracy. This indicates the pipeline requires further refinement of the extraction logic.
- The pipeline currently lacks the capability to extract and classify material properties such as crystal structure (single crystal/powder form), synthesis methods, and dimensional characteristics.
- PDF retrieval success rate is inconsistent due to intermittent connectivity issues with Sci-Hub and arXiv APIs, which requires more robust fallback mechanisms.

Future Updates

Releases

No releases published

Packages

No packages published

Languages

Python 98.8% Batchfile 1.2%