1️⃣

DeepDipper_brainstorm_1

📊📈🔢🎯🤖👀 📟📡💡🔍🗄️‼️⛳🕳️. 🪄 🥸 🪢✨🧩 🔥💥💫⚡🌈💨💦🔥 🛰️ 🕵️

Tools and functionalities

1. Topic modeling: Given the categories/tags for each paper, use topic modeling algorithms like Latent Dirichlet Allocation (LDA) or Non-negative Matrix Factorization (NMF) to identify common themes and topics within the dataset.
2. Text classification: Train a deep neural network to classify papers into different categories based on their abstracts or titles. For example, you could use the categories provided in the dataset as labels, or create your own labels based on specific research topics.
3. Recommendation systems: Use collaborative filtering or content-based filtering techniques to recommend papers to users based on their interests or reading history. For example, could recommend papers to users who have read similar papers in the past, or based on the content of the papers they have already read.
4. Named entity recognition: Use deep learning models to identify and extract named entities (such as people, organizations, and locations) from the abstracts or titles of the papers.
5. Summarization: Use sequence-to-sequence models like encoder-decoder networks or transformer models to generate summaries of the papers based on their abstracts or full text. This could be useful for quickly identifying the main points of a large number of papers.

▼ Citation Network Analysis: Analyze the structure of the citation network to identify important papers or clusters of related papers. This can involve using graph neural networks to learn embeddings for each paper based on its citation network, and then clustering papers based on these embeddings. This can help identify key papers or subfields within the dataset.

For citation network analysis:

1. Graph Convolutional Networks (GCNs): GCNs are a type of neural network designed for graph-structured data, such as citation networks. GCNs can learn to identify important nodes and edges in a graph and can be used for tasks such as node classification, link prediction, and graph classification. You could use a GCN

to predict which papers are likely to be cited by others based on their features (e.g., title, abstract, year, authors).

2. Recurrent Neural Networks (RNNs): RNNs are another type of neural network that are useful for sequential data, such as text. You could use an RNN to model the citation patterns between papers over time. For example, you could train an RNN to predict the next paper that a given paper will cite based on its citation history.

3. It's worth noting that deep learning techniques may not always be the best choice for citation network analysis. Traditional network analysis techniques, such as centrality measures and community detection, can also be effective and may be easier to interpret.

About Graph convolutional networks:

GCNs are a type of neural network designed to operate on graph-structured data. They use a series of convolutional operations to learn node embeddings, which can be used for a variety of tasks, such as node classification, link prediction, and graph classification.

In the context of citation network analysis, you could use a GCN to predict which papers are likely to be cited by others based on their features, such as title, abstract, year, and authors. To do this, you would need to represent the citation network as a graph, with papers as nodes and citations as edges. You would also need to define a feature matrix for each node, which would include the features of the paper as well as any metadata about the node itself (e.g., its degree centrality).

▼ A gpt-style chatbot that allows user to interact with the database, and choose search algos under the hood.

About chatbot gpt-style:

1. Choose a platform: There are several platforms available to build chatbots, such as Dialogflow, Botpress, Rasa, and many more. Choose a platform that suits your needs and expertise.

2. Train a language model: To build a GPT-like chatbot, you need to train a language model. You can either train a model from scratch using a large corpus of data or fine-tune an existing pre-trained language model like GPT-2 or GPT-3. For fine-tuning, you can use a library like Hugging Face's Transformers, which provides pre-trained models and tools for fine-tuning.

3. Build a conversational flow: Define the conversational flow of your chatbot. Think about the user's goal, what information they need, and how they can interact with the chatbot to achieve their goal.

4. Build an API: Build an API that interfaces with your deep learning and machine learning algorithms. The API should take inputs from the chatbot, call the appropriate algorithm, and return the results to the chatbot.
5. Test and deploy: Test your chatbot thoroughly and deploy it to your chosen platform. Monitor its performance and collect user feedback to improve its accuracy and usability over time.

Building a GPT-like chatbot can be a complex process, but it can provide an efficient and user-friendly way for users to extract useful information from your research papers dataset.

- have “influence of researchers”, with profiles for each researchers, listing both their papers and all papers citing researcher, with an overview graph of influence on sectors (# of times cited in each sector).
- add tools to extract text from pdfs and analyze them more thoroughly.

Data available - to aggregate

Use Kaggle available dataset

Get citation count through API?

https://sites.google.com/site/teamcitations/data-sources

Other possible datasets:

- arXiv (downloaded from Kaggle, 2.2m lines, 3.66Go, up to today)
- arXiv cross-reference and citations (downloaded from Kaggle, 224Mb)
- arXiv AI_ML focused since 2013 (json 161Mb
- JMLR (scraped on website, 3.3k lines, up to today)

▼ CrossRef metadata records (to be dl’ed in torrents, 120m+, up to jan 21)

https://academictorrents.com/details/e4287cb7619999709f6e9db5c359dda17e93d515

Crossref citation data (77m publi, 1.4m citations)

https://doi.org/10.6084/m9.figshare.6741422

▼ Crossref DOI records (to be dl’ed torrents, 112m+, up to mar 20):

https://academictorrents.com/details/0c6c3fbfdc13f0169b561d29354ea8b188eb9d63

dump containing the number of incoming citations to each bibliographic entity (77m public, 2.2Gb, 0.7Gb zipped, identified by DOI)
https://doi.org/10.6084/m9.figshare.21494286.v1

▼ SciHub database (to be dl’ed, 88m+, think it is 30Gb but link doesn’t work anymore)

https://sci-hub.ru/database

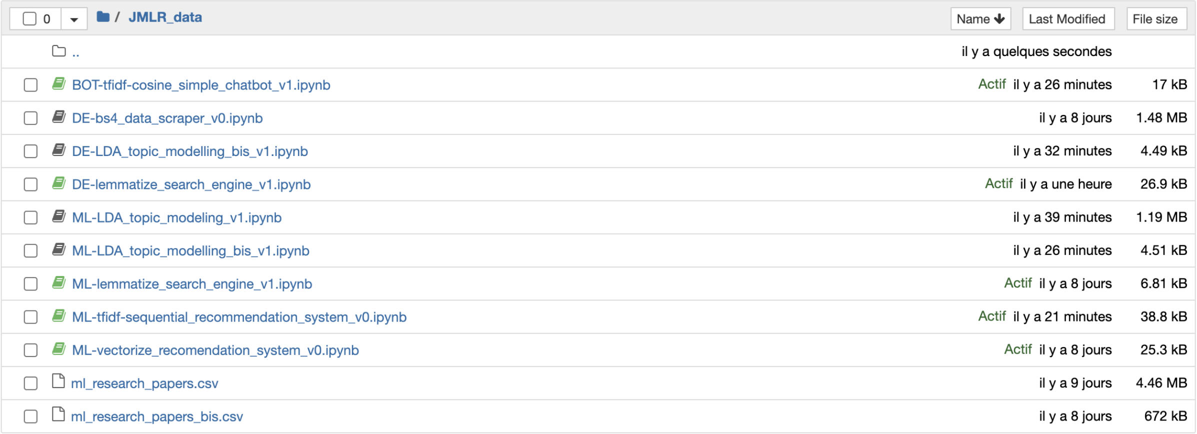
▼ OpenCitations Metadata (to be dl’ed, 90m biblio, 282m authors, 36.3 GB, 8.2Gb zipped

https://doi.org/10.6084/m9.figshare.21747461

▼ OpenCitations citation data (to be dl’ed, 77m biblio, 1.4Bn citations, 238Gb, 37.5Gb zipped

Available notebooks and algos worksheets (as of 05.03.2023):

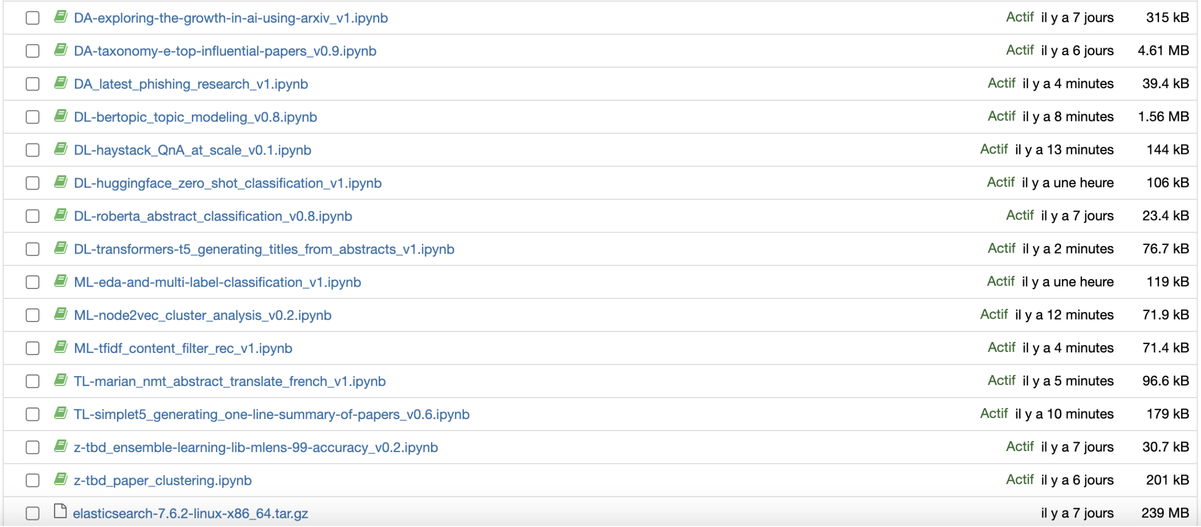
- *JMLR-focused*



☐ 0 ▾ / JMLR_data

Name ↓		Last Modified	File size
..		il y a quelques secondes	
☐ BOT-tfidf-cosine_simple_chatbot_v1.ipynb	Actif	il y a 26 minutes	17 kB
☐ DE-bs4_data_scraper_v0.ipynb		il y a 8 jours	1.48 MB
☐ DE-LDA_topic_modelling_bis_v1.ipynb		il y a 32 minutes	4.49 kB
☐ DE-lemmatize_search_engine_v1.ipynb	Actif	il y a une heure	26.9 kB
☐ ML-LDA_topic_modeling_v1.ipynb		il y a 39 minutes	1.19 MB
☐ ML-LDA_topic_modelling_bis_v1.ipynb		il y a 26 minutes	4.51 kB
☐ ML-lemmatize_search_engine_v1.ipynb	Actif	il y a 8 jours	6.81 kB
☐ ML-tfidf-sequential_recommendation_system_v0.ipynb	Actif	il y a 21 minutes	38.8 kB
☐ ML-vectorize_recomendation_system_v0.ipynb	Actif	il y a 8 jours	25.3 kB
☐ ml_research_papers.csv		il y a 9 jours	4.46 MB
☐ ml_research_papers_bis.csv		il y a 8 jours	672 kB

- *arXiv-focused*



☐ DA-exploring-the-growth-in-ai-using-arxiv_v1.ipynb	Actif	il y a 7 jours	315 kB
☐ DA-taxonomy-e-top-influential-papers_v0.9.ipynb	Actif	il y a 6 jours	4.61 MB
☐ DA_latest_phishing_research_v1.ipynb	Actif	il y a 4 minutes	39.4 kB
☐ DL-bertopic_topic_modeling_v0.8.ipynb	Actif	il y a 8 minutes	1.56 MB
☐ DL-haystack_QnA_at_scale_v0.1.ipynb	Actif	il y a 13 minutes	144 kB
☐ DL-huggingface_zero_shot_classification_v1.ipynb	Actif	il y a une heure	106 kB
☐ DL-roberta_abstract_classification_v0.8.ipynb	Actif	il y a 7 jours	23.4 kB
☐ DL-transformers-t5_generating_titles_from_abstracts_v1.ipynb	Actif	il y a 2 minutes	76.7 kB
☐ ML-eda-and-multi-label-classification_v1.ipynb	Actif	il y a une heure	119 kB
☐ ML-node2vec_cluster_analysis_v0.2.ipynb	Actif	il y a 12 minutes	71.9 kB
☐ ML-tfidf_content_filter_rec_v1.ipynb	Actif	il y a 4 minutes	71.4 kB
☐ TL-marian_nmt_abstract_translate_french_v1.ipynb	Actif	il y a 5 minutes	96.6 kB
☐ TL-simplet5_generating_one-line-summary-of-papers_v0.6.ipynb	Actif	il y a 10 minutes	179 kB
☐ z-tbd_ensemble-learning-lib-mlens-99-accuracy_v0.2.ipynb	Actif	il y a 7 jours	30.7 kB
☐ z-tbd_paper_clustering.ipynb	Actif	il y a 6 jours	201 kB
☐ elasticsearch-7.6.2-linux-x86_64.tar.gz		il y a 7 jours	239 MB