Analysis of DeepSeek-Related News Coverage

1. Background

DeepSeek is an emerging AI model serie that has recently gained significant attention. As a student specializing in AI&data, I recognize that it's quite related to our filed because it brings business effects too, since the introduction of DeepSeek-R1 has also sparked discussions about the relationship between model training costs and economic implications. So in order to understand how Deepseek is prescribed in English language media, I collected and analyzed news coverage spending approximately 20 days starting from the release of the DeepSeek R1 model.

To achieve this, I experimented with several web scraping methods. Due to API limitations and restrictions, I opopted to use `pygooglenews` to gather related URLs from Google News, as well as the titles and publication dates.

However, I encountered multiple redirects in these links, requiring a manual approach to obtain the actual URLs. After filtering and verification, I compiled a dataset of **837 valid articles** for this analysis, saved in `deepseek_news_corpus.csv`. You can see the details in the second part, and more details in the two notebooks.

---

2. Data Collection and Preprocessing

The data collection process involved:

- Extracting Google News data using `pygooglenews`.
- Manually retrieving redirected URLs and storing them in structured files.
- Filtering duplicate links and verifying content validity.

Once the dataset was prepared, I conducted **text preprocessing** using `SpaCy`, instead of `nltk`, as I found its lemmatization capabilities more effective, but nltk's lemmatization's kind of "bizarre". Our preprocessing steps included:

- Lowercasing all text at the beginning
- Removing HTML tags and URLs to clean up the content.
- **Lemmatization (I chose the lemmas instead of stemming to ensuring better interpretability).

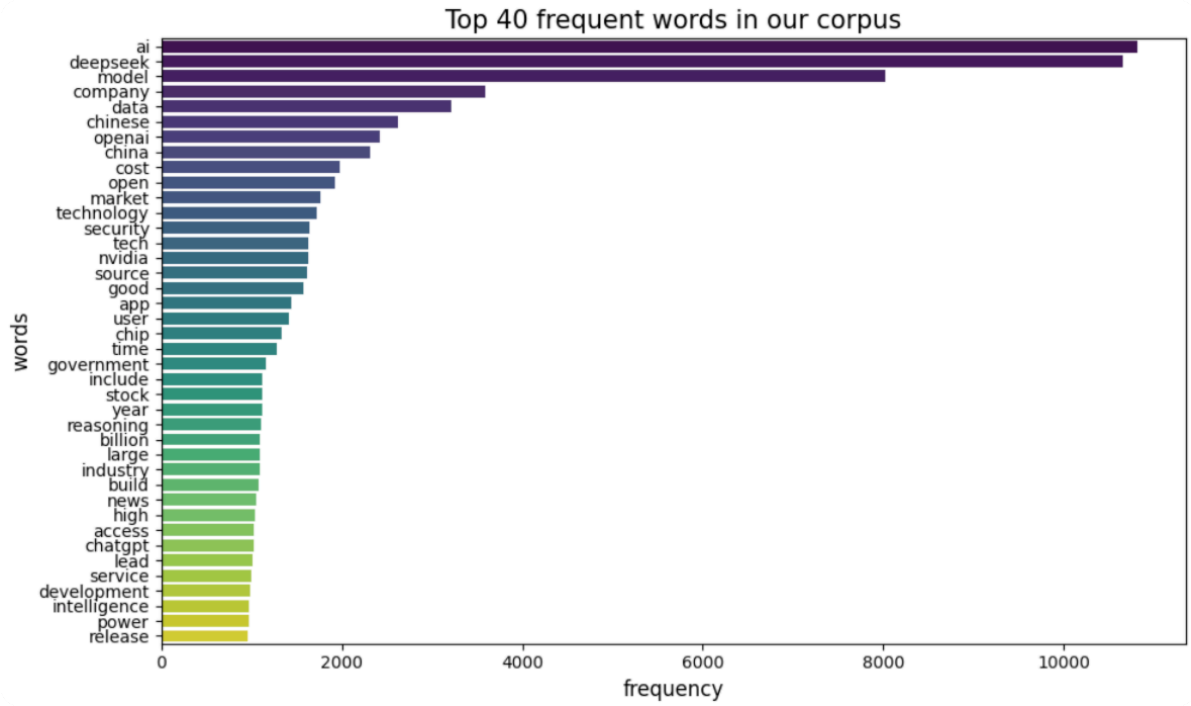
- **Stopword removal**, using both the default `spaCy` stopword list and an extended stopword set from GitHub.
    - I firstly only used the default sw list, but it turns out that there are still many low information words in the results whether it's about the frequency or the Wordcloud, so I searched on GitHub and download a more exhaustive English stopwords list.
- **Filtering non-alphabetic tokens**, as numerical data, while relevant in financial news, lacked sufficient context for meaningful NLP analysis.

All data processing was performed and organized on **pandas DataFrames**. The final data frame before the analysis looks like this :

	URL	Title	Date	Content	Status_Code	Validation	preprocessed_content
0	https://www.techtarget.com/whatis/feature/Deep...	DeepSeek explained: Everything you need to know	2025-01-20	In the world of AI, there has been a prevailin...	200	1	ai prevail notion develop leading edge large l...
1	https://venturebeat.com/ai/open-source-deepsee...	Open-source DeepSeek-R1 uses pure reinforcemen...	2025-01-20	Join our daily and weekly newsletters for the ...	200	1	join daily weekly newsletter late update exclu...
2	https://siliconangle.com/2025/01/20/deepseek-o...	DeepSeek open-sources its R1 reasoning model s...	2025-01-20	\nUPDATED 17:12 EST / JANUARY 20 2025\n\n\n\n\...	200	1	update january maria deutscher deepseek today ...
3	https://medium.com/data-science-in-your-pocket...	DeepSeek-R1 vs DeepSeek-R1-Zero. DeepSeek's ne...	2025-01-20	Sign up\nSign in\nSign up\nSign in\nMehul Gupt...	200	1	sign sign sign sign mehul gupta follow data sc...
4	https://analyticsindiamag.com/ai-news-updates/...	DeepSeek Crushes OpenAI o1 with an MIT-License...	2025-01-20	DeepSeek, a Chinese AI research lab backed by ...	200	1	deepseek chinese ai lab high flyer capital man...
...	...	...	...	...	...	...	...
832	https://www.techloy.com/gemini-2-0-flash-is-go...	Gemini 2.0 Flash is Google's latest response t...	2025-02-07	\n\n\nSuccess! Now Check Your Email\n\nTo comp...	200	1	success check email complete subscribe click c...
833	https://newscentral.africa/us-lawmakers-seek-d...	US Lawmakers Seek DeepSeek Ban on Government D...	2025-02-07	A new bill to be introduced in the U.S. Congre...	200	1	introduce congress thursday seek ban deepseek ...
834	https://ipdefenseforum.com/2025/02/prcs-ai-ass...	PRC's AI assistant, DeepSeek, censors informat...	2025-02-07	Voice of America\nUsers of the People's Republ...	200	1	voice america user people republic china prc a...
835	https://www.androidheadlines.com/2025/02/deeps...	DeepSeek and other Chinese AIs might be banned...	2025-02-07	Sign Up! \n\nenvelope_alt\n\n\n\nGet the lates...	200	1	sign late android news inbox day sign receive ...
836	https://www.fool.com/investing/2025/02/07/deep...	The DeepSeek News Makes Apple's Artificial Int...	2025-02-07	Founded in 1993, The Motley Fool is a financia...	200	1	motley fool financial service company dedicate...

3. Statistical and NLP Analysis

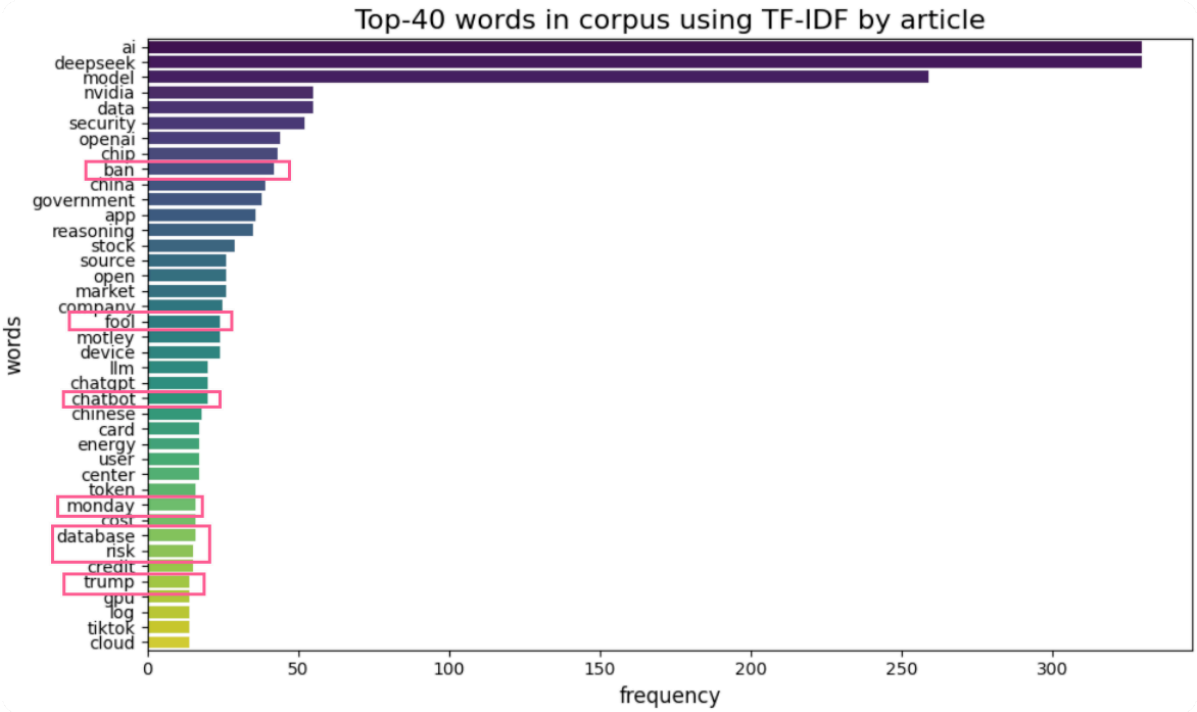
Global Word Frequency Analysis



In order to gain insights into word usage, we:

- Used `Counter` to compute overall word frequencies across all articles.
- Visualized the top 40 most frequent words through bar plots.

Word frequency via TF-IDF Keyword Extraction for each article

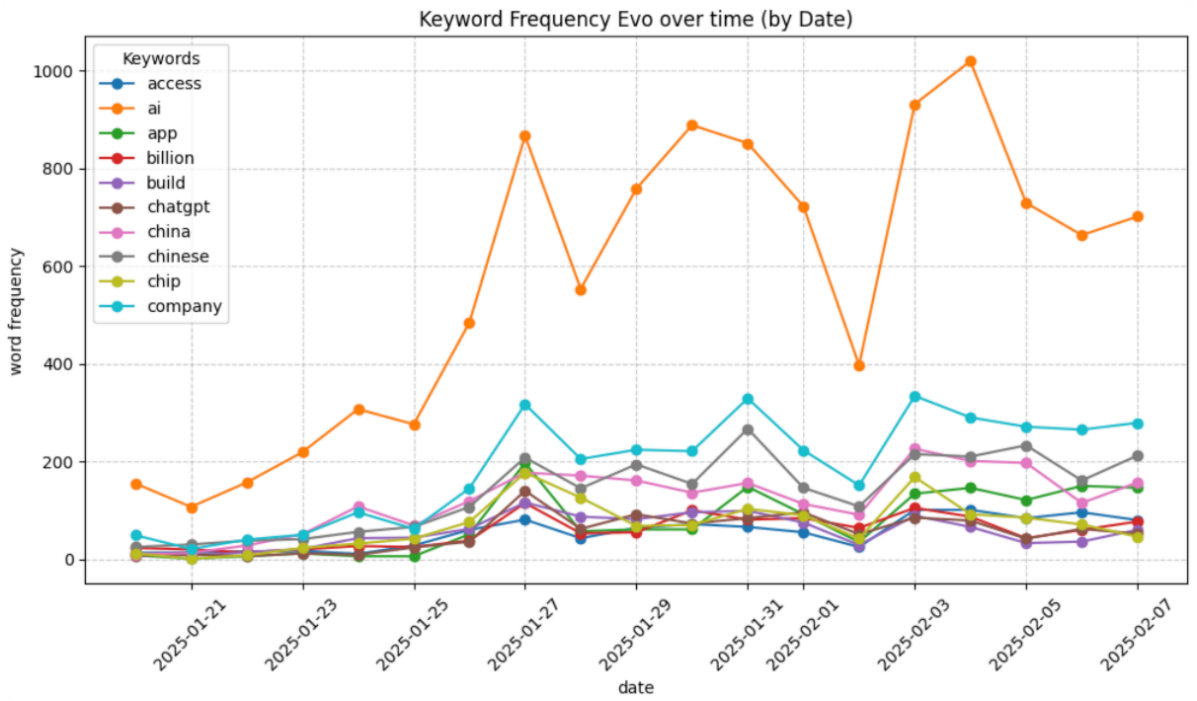


To extract more meaningful and distinctive keywords, I applied **TF-IDF** ：

- Computed the **top 5** keywords for each article.
- Aggregated these keywords to identify the most distinctive words across the dataset.
- Compared results with the first global word frequency analysis and observed notable differences with meaningful words like "ban", .

Time-Series Analysis of Keywords

I also did a examinition keyword frequency over time to detect potential trends. Although **no clear trends** emerged for individual words, I observed a **gradual increase in overall keyword occurrences**, indicating a growing volume of news coverage on *DeepSeek* over time.



Word Cloud Visualization ☁️

What I did about Word Cloud :

- A **classic word cloud** based on global word frequency distribution.

- A **custom DeepSeek-themed word cloud** ✨ which sacrificed some readability...

---

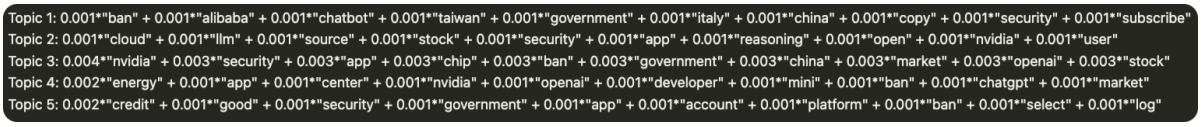
4. Topic Modeling with LDA

Given the volume of news articles, I think it would be interesting if we go beyond the word frequency studies and extract underlying themes using LDA.

The challenges & refinements

- I initially tried the default LDA modeling, which resulted in **high-frequency words appearing across multiple topics**.
- To address this, I **filtered extreme frequency words** using **Gensim's `filter_extremes()` function**.
- I experimented with different numbers of topics (from 3 to 5) and chosed 5.

Using LDA, I identified five key themes in DeepSeek-related news coverage:



To further interpret these topics, I utilized ChatGPT to label the topic groups, then I gave my interpretation based on my personal knowledge about this topic :

1. **Regulatory and Geopolitical Challenges** (Given by GPT)

   > Topics on AI regulation, security risks, and international policies. This theme includes terms like "ban," "Taiwan," "government," and "security," reflecting concerns over DeepSeek's compliance and restrictions in certain regions (e.g., Italy banning the model).

2. **Technological Advancements and Cloud Integration** (Given by GPT)

   > Topics on cloud-based AI development and infrastructure, featuring words such as "cloud," "LLM," and "source."

3. **Market Impact and Competitive Landscape** (Given by GPT)

   > Economic implications, investments, and DeepSeek’s position in the industry. Keywords such as "Nvidia," "stock," "market," and "chip" highlight its influence on the AI ecosystem and competitors like OpenAI.

4. **AI Infrastructure and Energy Consumption** (Given by GPT)

   > Concerns about the cost, scalability, and resource demands of AI models. Keywords such as "GPU," "power," and "industry" suggest discussions on computational efficiency and energy consumption.

5. **Security, Privacy, and Ethical Considerations** (Given by GPT)

   > AI-related risks, privacy debates, and governmental oversight. Words like "credit," "security," "government," "platform," and "account" reflect concerns about data security and the potential risks of interacting with AI models online.