Midterm Check - Team 15

Data Science Group Project

Our Project and Research Question

Project Focus

Automation of "See Also" section in Wikipedia articles

Research Question

“How can a system be developed to automate the creation and update of the 'See Also' section in Wikipedia articles by utilizing article features and NLP-generated semantic vectors?”

An Example of the “See Also” Section

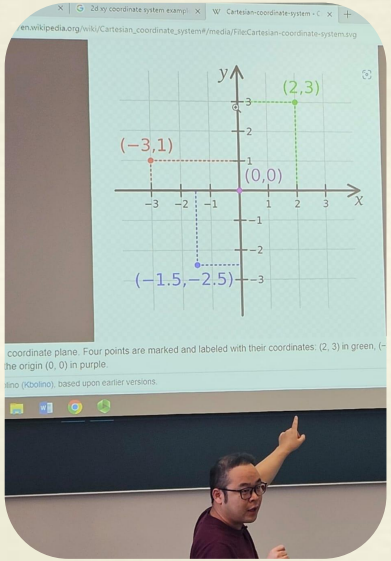
See also

- Earth science – Fields of natural science related to Earth
- Neurophysics – branch of biophysics dealing with the development and use of physical methods to gain information about the nervous system
- Psychophysics – Branch of knowledge relating physical stimuli and psychological perception
- Relationship between mathematics and physics – Study of how mathematics and physics relate to each other
- Science tourism – Travel to notable science locations

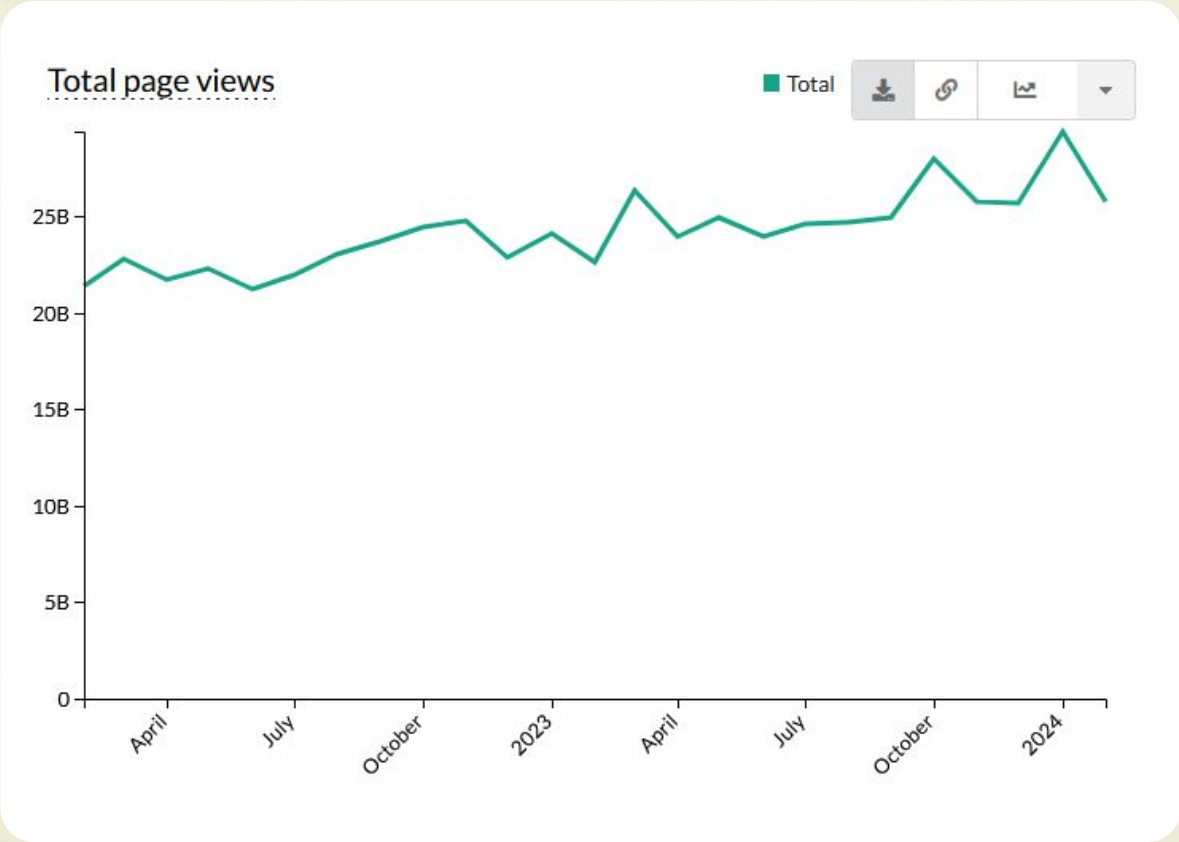
See Also Section in Physics Article

Why Wikipedia?

- **Leading Knowledge Source: Top** Source**, and Largest** Encyclopedia.
- Operates on a **volunteer system**
- **Limited number of volunteers,** particularly in **non-English** languages



Last week, **Dr. Jinming Duan** used a photo from **English Wikipedia** to demonstrate **PCA** in the Visualization module.

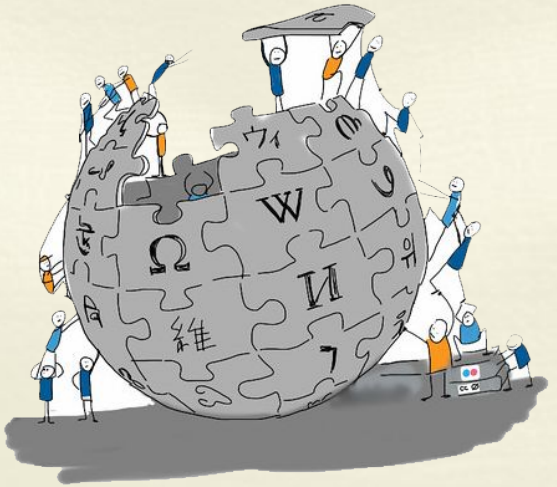


***Source:** Wikimedia Statistics

Wikipedia's **monthly visits** over the **past two years,** with articles read **over 600 billion times** and **growing!** Source: Wikimedia Statistics

Our Project Impact

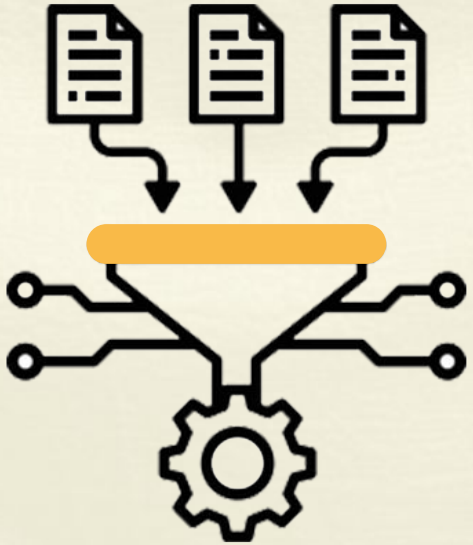
- Many articles **lack** a **"See Also" section**
- Creation and updates need **expert volunteers** and are **time-consuming**
- Our project **enhances reader experience** and **conserves volunteers time**
- Provides **relevant**, **popular**, and **high-quality** articles in "See Also" **for the reader**
- **Automates** creation and updates, s**aving volunteers efforts**



Preparing The Dataset

Preparing The Dataset

- **Challenge Faced:** Tackling an **unprepared** dataset
- **Data Preparation: 3 weeks** of intensive work
- **Key Features:** Identified through **team discussions** and **brainstorming**



Preparing The Dataset

After extensive discussions, we agreed on fetching the following features:

- **Article titles:** limit scope to **60,000 most-clicked** articles for **computational efficiency.**
- Taken from the clickstream dataset.

Clickstream data from Wikimedia Dumps

Preparing The Dataset

Then, we used the **Wikipedia APIs** and **web scraping** to collect other features:

Web Scraping

Article introduction

Article Categories

Utilized with the article title for **NLP and vector generation.**

Wikipedia APIs (Xtools, and Views Tool)

Article size

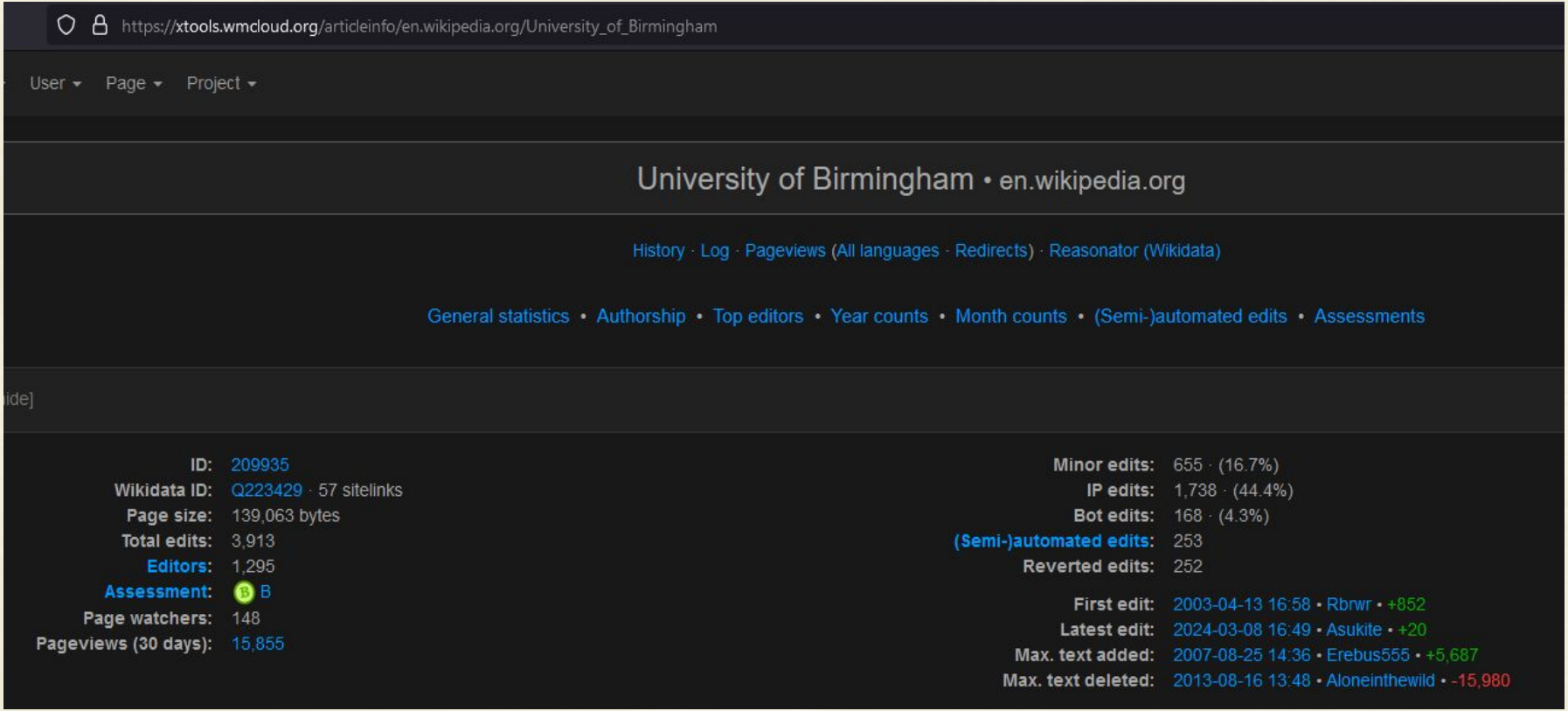
to **exclude** small articles

Article Views

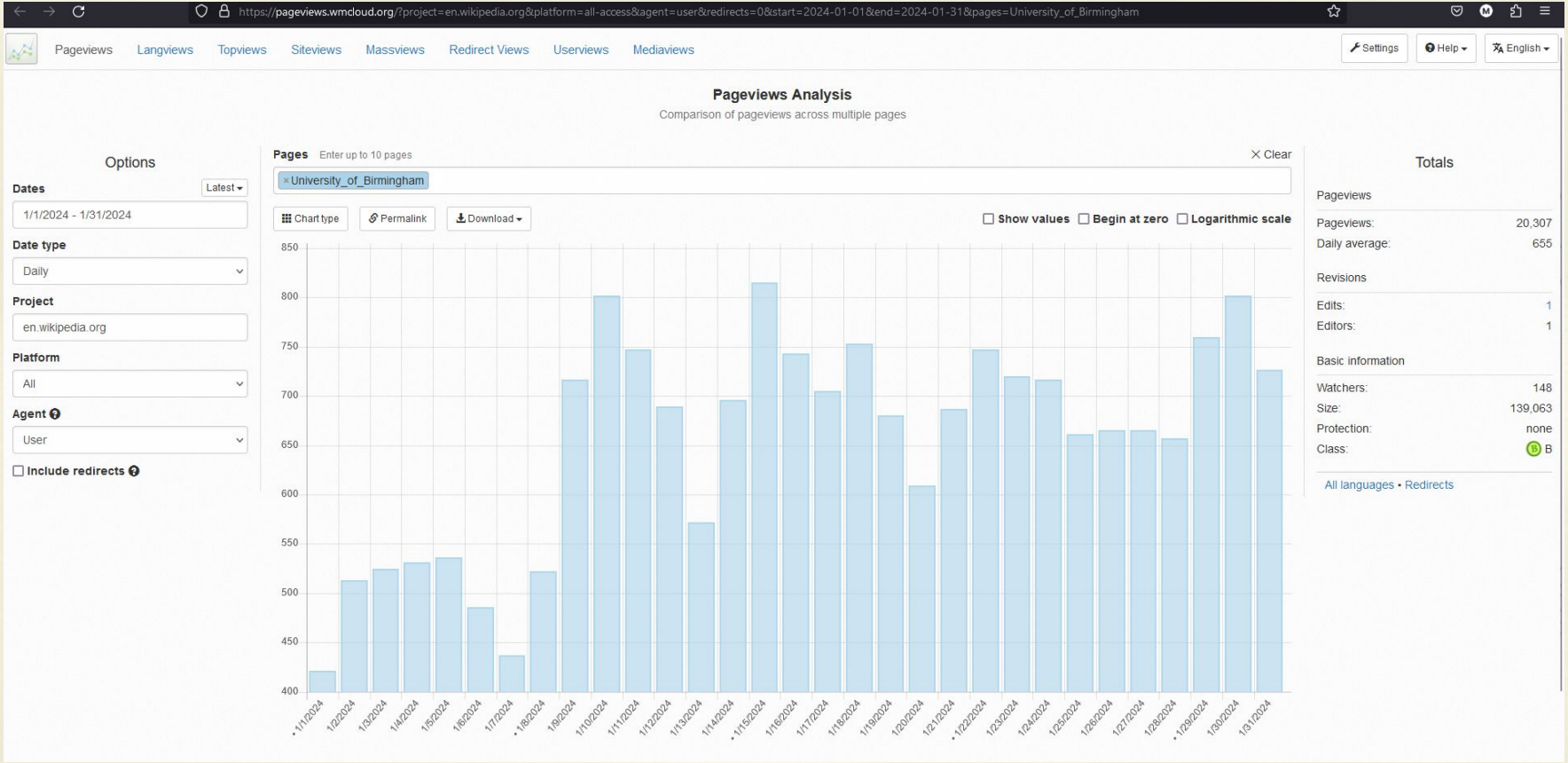
To ensure recommending **popular** articles

Article Quality Type

To sort the final list according to quality rank



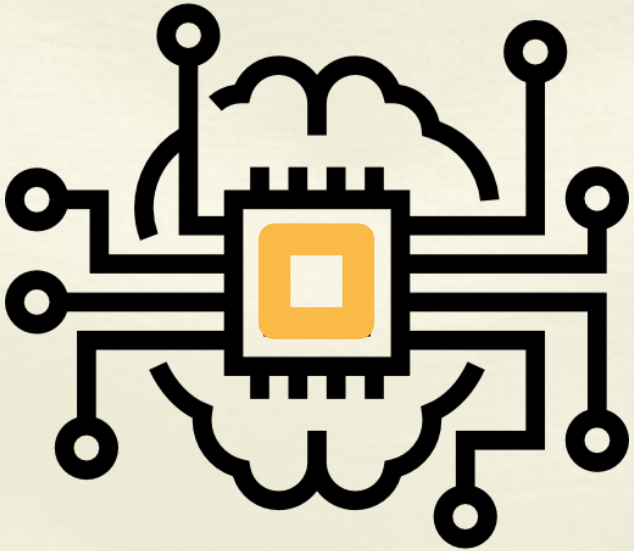
University of Birmingham Article - Xtools



University of Birmingham Article - Views Tool

Embeddings Generation

- Researched leading **ML - NLP models**.
- Chose **Google's BERT** for:
  - **State-of-the-Art** Performance.
  - **Deep** Contextual Understanding.
  - Availability of **Pre-trained Models.**
- Vector extraction: around **7 hours.**



```
Processing time: 26626.30 seconds
```

```
[ ]  1 new_df
```

	title	first_paragraph	article_categories	text	bert_0	bert_1	bert_2
0	Malcolm_Brogdon	Malcolm Moses Adams Brogdon (born December 11,...	1992 births, 21st-century African-American spo...	Malcolm_Brogdon Malcolm Moses Adams Brogdon (b...	-0.766978	0.500399	-0.265357
1	Thomas_Kinkade	William Thomas Kinkade III (January 19, 1958 –...	1958 births, 2012 deaths, 20th-century America...	Thomas_Kinkade William Thomas Kinkade III (Jan...	-0.485413	0.378151	0.136504
2	Frank_Gore	Franklin Gore Sr. (born May 14, 1983) is an Am...	1983 births, American football running backs, ...	Frank_Gore Franklin Gore Sr. (born May 14, 198...	-0.705500	0.552352	-0.777939
3	Drug_Enforcement_Administration	The Drug Enforcement Administration (DEA) is a...	1973 establishments in Washington, D.C., Drug ...	Drug_Enforcement_Administration The Drug Enfor...	-0.530161	0.554903	-0.081502

Proof of Running Time - Executed on Google Colab

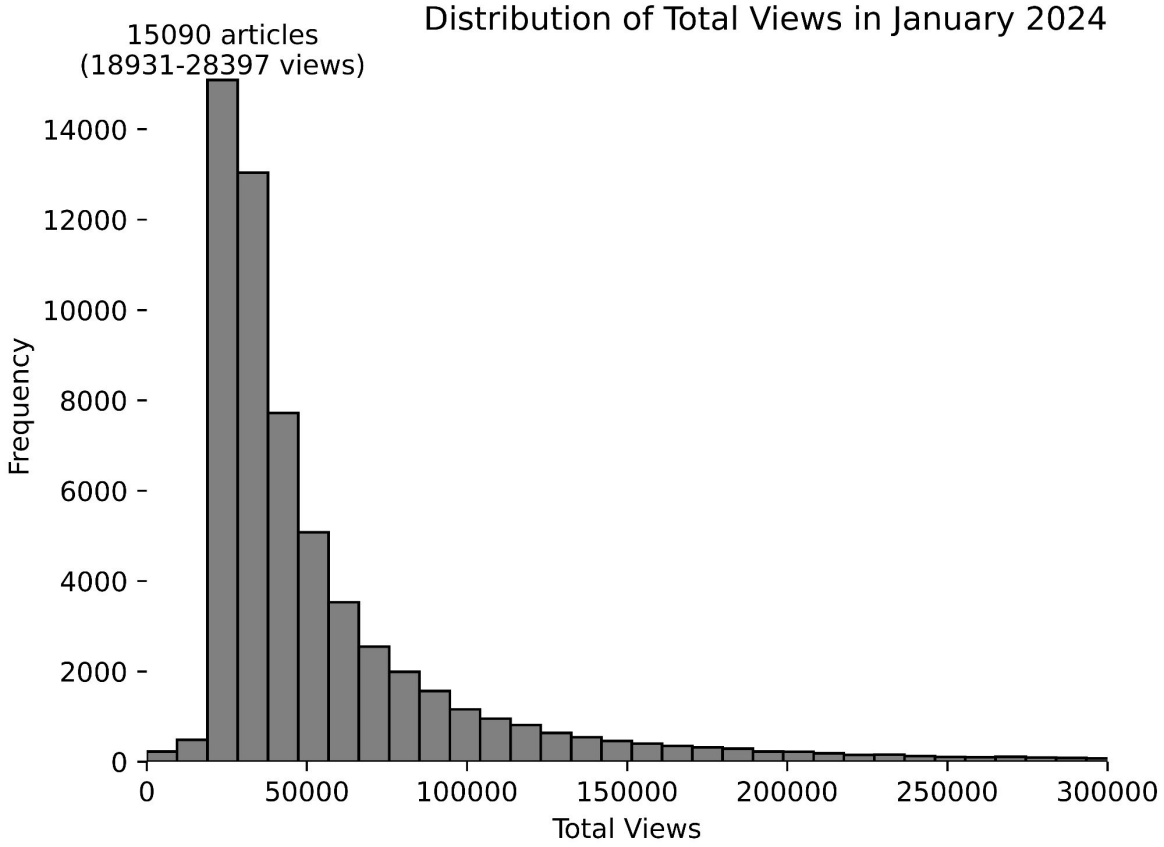
```
In [5]: data
```

Out[5]:

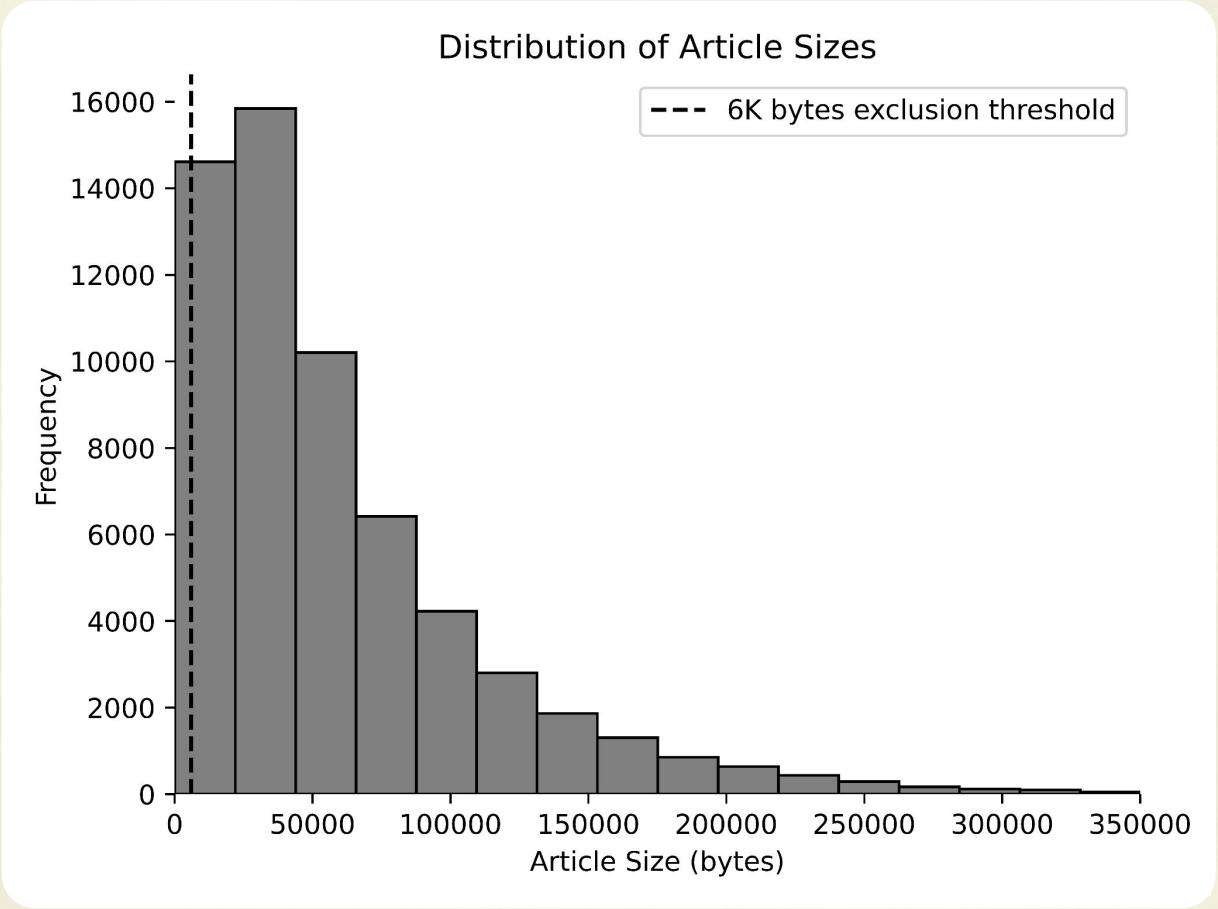
	title	size	total_views	Introduction	article_quality	article_categories	bert_0	bert_1	bert_2	bert_3	...
1	Saltburn_(film)	69335	7450496	Saltburn is a 2023 black comedy psychological ...	B	['2020s American films', '2020s British films'...	-0.447147	0.665674	0.293427	-0.021366	...
3	Griselda_Blanco	22518	5360838	Griselda Blanco Restrepo (February 15, 1943 – ...	C	['1943 births', '2012 deaths', '20th-century C...	-0.302390	-0.095012	-0.879526	-0.183938	...
4	XXXTentacion	239001	7723810	Jahseh Dwayne Ricardo Onfroy (January 23, 1998...	B	['1998 births', '2018 deaths', '21st-century A...	-0.038375	0.158305	0.200293	-0.055052	...
5	Jeffrey_Epstein	286715	4402725	Jeffrey Edward Epstein (EP-steen; January 20,...	B	['1953 births', '2000s controversies in the Un...	-0.603641	0.218759	-0.318667	0.037199	...
6	Deaths_in_2024	146551	4143166	The following notable deaths occurred in 2024....	List	['2024 deaths', 'Articles with Dutch-language ...	-0.322070	0.320208	-0.204003	-0.184792	...

Final Dataset

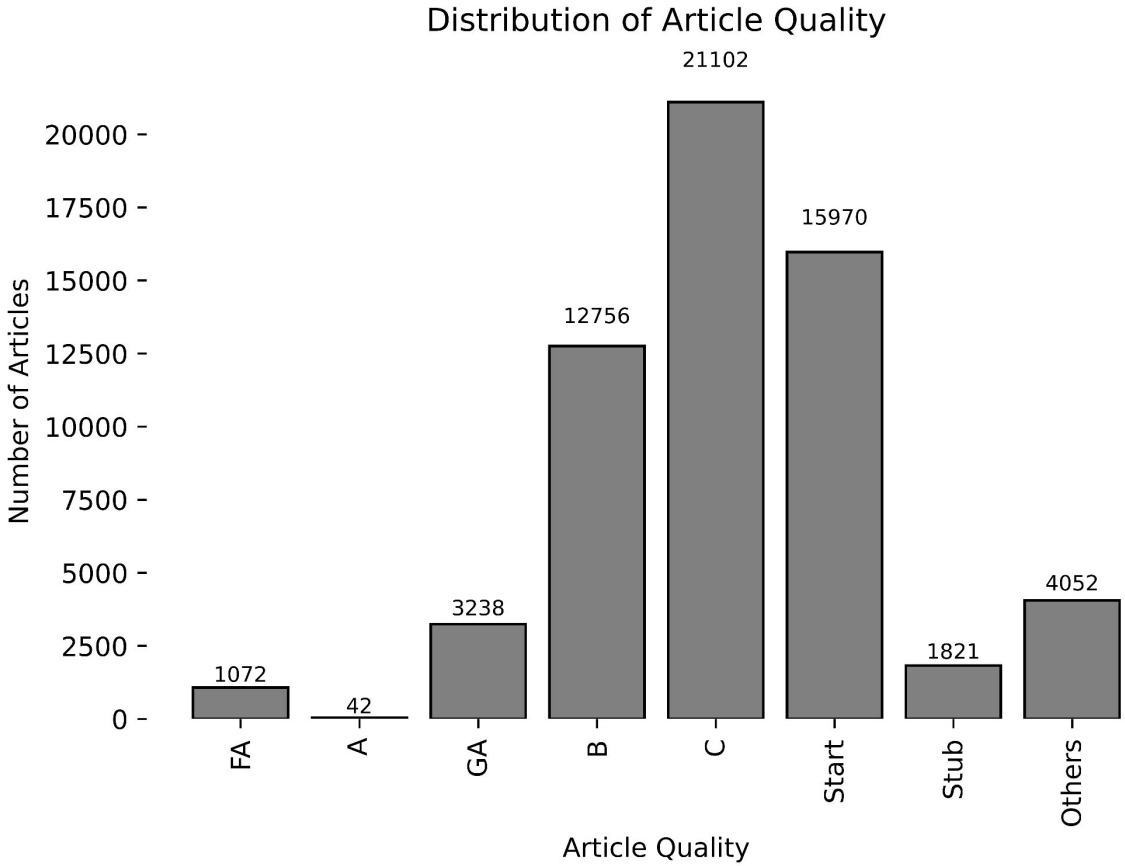
Exploratory Data Analysis



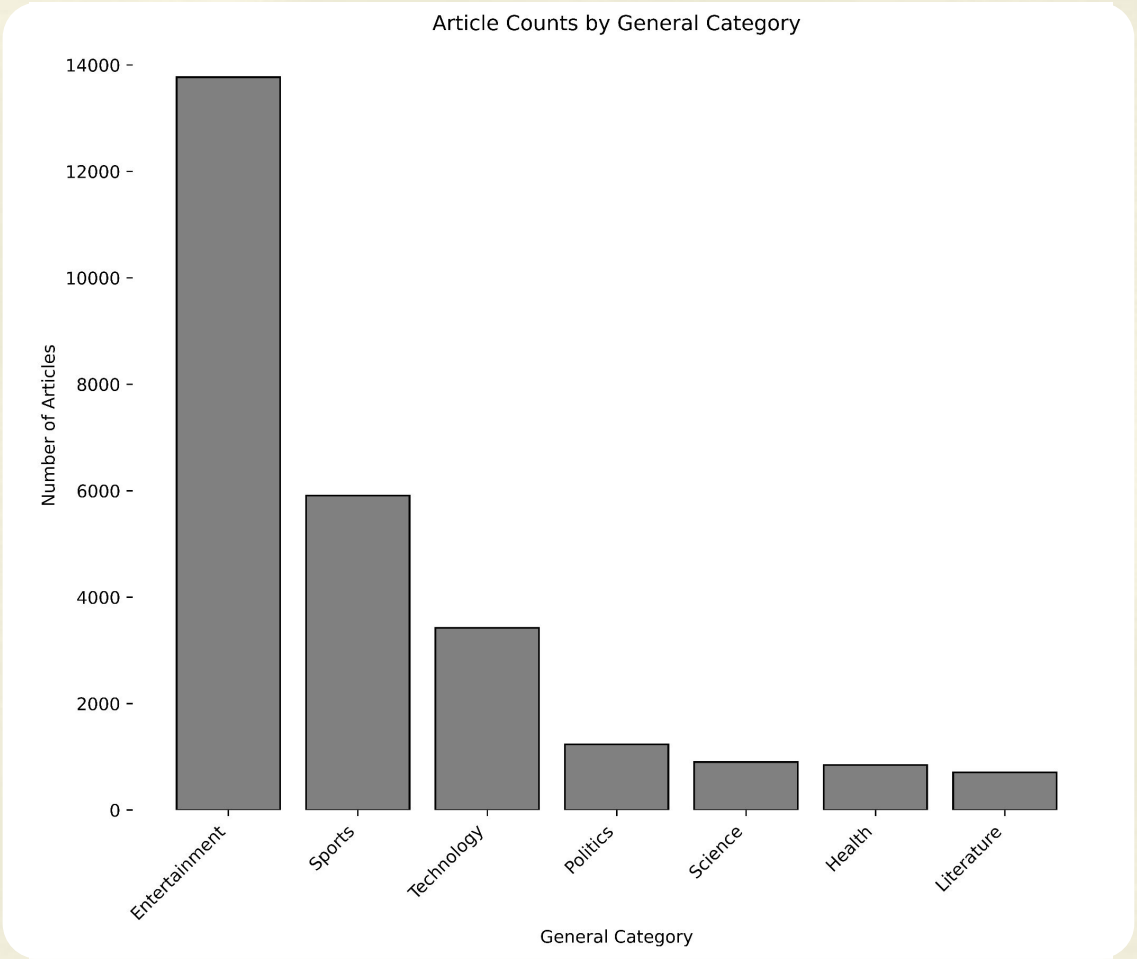
Views Distribution: Assessed to gauge the popularity of articles, confirming our selections resonate with reader interests.



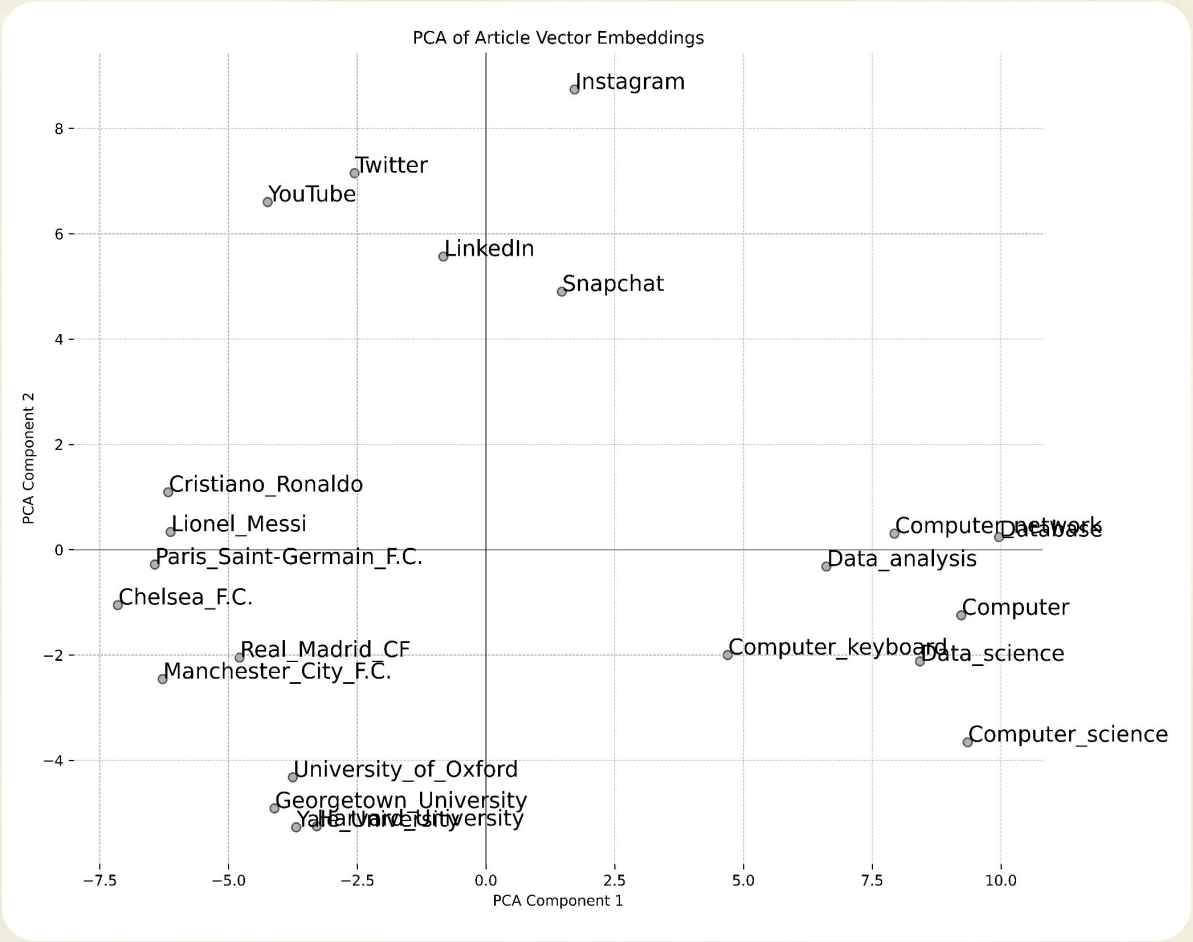
Article Size Distribution: Analyzed to later exclude articles under 6KB, ensuring the recommendations are of substantial content



Quality Distribution: Evaluated with a focus on ascending quality, prioritizing 'Featured Article' (FA) as the highest quality, to ensure recommendations are high quality.

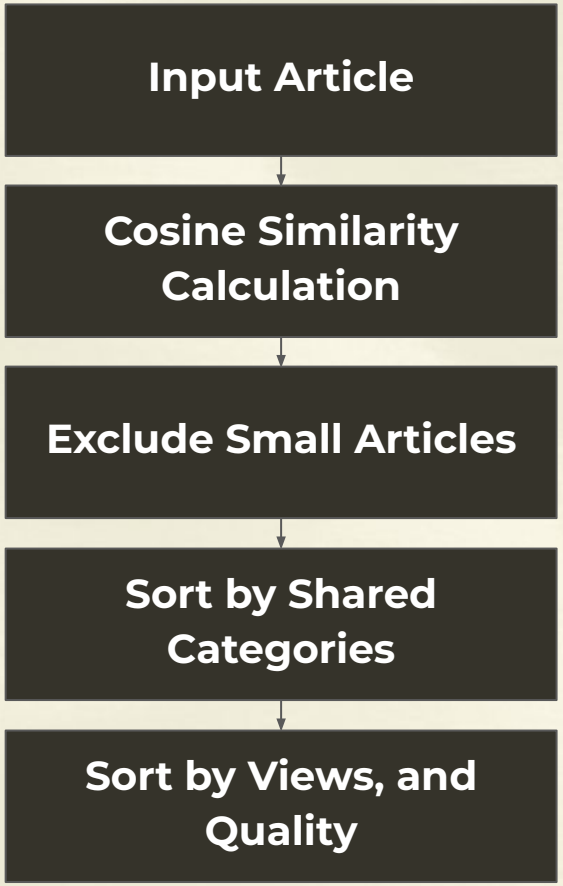


Categories Distribution: Explored to guarantee a diverse selection across various subjects, enhancing the general applicability of our automated system.



BERT Embeddings Visualization: Employed PCA to visualize BERT embeddings, ensuring our system accurately groups and recommends related articles, enhancing the 'See Also' section's relevance.

How The System Works?



Start with an **article** as **the input**

Calculate cosine similarity between **the input article** and **others using BERT embeddings**

Remove articles **under 6,000 bytes** to ensure content depth, and **select top 20**

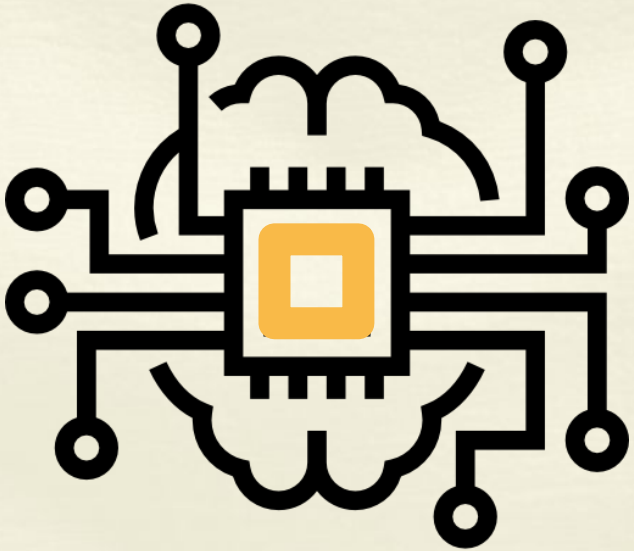
Sort the remaining articles by the **number of shared categories** with the input article and **select the top 10.**

Sort these 10 articles **by total views**, **select top 5**, and sort them **by quality**

How The System Works

Cosine Similarity

- **High-Dimensional Suitability: Ideal** for comparing **vectors** in **high-dimensional spaces.**
- **Efficiency:** Less computationally intensive, crucial **for large datasets.**
- **Content-Length Neutral: Normalizes vector lengths,** allowing **fair comparison** across **varying** article **lengths**.



```
    filtered_indices = [i for i in top_20_indices if data.iloc[i]['size'] >= 6000]

    selected_features = ["title", "size", "total_views", "Introduction", "article_quality", "article_categories"]

    recommendations_df = data.iloc[filtered_indices][:top_n][selected_features].copy()
    recommendations_df['Similarity Score'] = similarities[filtered_indices][:top_n]

    current_article_info = pd.DataFrame({
        'title': [data.iloc[current_article_index]['title']],
        'size': [data.iloc[current_article_index]['size']],
        'total_views': [data.iloc[current_article_index]['total_views']],
        'Introduction': [data.iloc[current_article_index]['Introduction']],
        'article_quality': [data.iloc[current_article_index]['article_quality']],
        'article_categories': [data.iloc[current_article_index]['article_categories']],
        'Similarity Score': [np.nan]
    })

    return pd.concat([current_article_info, recommendations_df], ignore_index=True)

current_article_name = 'Data_science'
recommendations_df = recommend_articles_with_info(current_article_name, data, top_n=20)
recommendations_df
```

Step 1 - Input article

	title	size	total_views	Introduction	article_quality	article_categories	Similarity Score
0	Data_science	23985	41348	Data science is an interdisciplinary academic ...	C	['Computational fields of study', 'Computer oc...	NaN
1	Data	21180	45251	In common usage data (US: ; UK:) is a collect...	C	['Data', 'Data management', 'Statistical data']	0.798207
2	Bioinformatics	135562	26881	Bioinformatics ( ) is an interdisciplinary fie...	Unknown	['Bioinformatics']	0.783384
3	Methodology	97488	27475	In its most common sense, methodology is the s...	Start	['Methodology']	0.781497
4	Bayesian_inference	65403	23862	Bayesian inference (BAY-zee-ən or BAY-zhən) ...	B	['Bayesian inference', 'Logic and statistics',...	0.773080
5	Deep_learning	180889	52094	Deep learning is the subset of machine learnin...	C	['Artificial neural networks', 'Deep learning']	0.770569
6	Computer_science	78268	88087	Computer science is the study of computation, ...	C	['Computer science', 'Formal sciences']	0.760875
7	Root_cause_analysis	30816	24353	In science and engineering, root cause analysi...	C	['Problem solving', 'Quality control tools']	0.750081
8	Branches_of_science	39847	46239	The branches of science, also referred to as s...	B	['Branches of science', 'Scientific disciplines']	0.748793
9	Data_structure	17005	28170	In computer science, a data structure is a dat...	C	['Data structures']	0.746238
10	Data_analysis	87968	32325	Data analysis is the process of inspecting, cl...	B	['Big data', 'Computational fields of study', ...	0.744758
11	Cognition	48637	31442	Cognition is the "mental action or process of ...	C	['Cognition', 'Cognitive psychology', 'Cogniti...	0.744027
12	Research	66088	44184	Research is "creative and systematic work unde...	B	['Ethics', 'Methodology', 'Research', 'Scienti...	0.742864
13	Science	168040	147213	Science is a rigorous, systematic endeavor tha...	B	['Main topic articles', 'Observation', 'Science']	0.741779

Step 2 - Cosine Similarity Calculation

	title	size	total_views	Introduction	article_quality	article_categories	Similarity Score
0	Data_science	23985	41348	Data science is an interdisciplinary academic ...	C	['Computational fields of study', 'Computer oc...	NaN
1	Data	21180	45251	In common usage data (US: ; UK:) is a collect...	C	['Data', 'Data management', 'Statistical data']	0.798207
2	Bioinformatics	135562	26881	Bioinformatics ( ) is an interdisciplinary fie...	Unknown	['Bioinformatics']	0.783384
3	Methodology	97488	27475	In its most common sense, methodology is the s...	Start	['Methodology']	0.781497
4	Bayesian_inference	65403	23862	Bayesian inference (BAY-zee-ən or BAY-zhən) ...	B	['Bayesian inference', 'Logic and statistics',...	0.773080
5	Deep_learning	180889	52094	Deep learning is the subset of machine learnin...	C	['Artificial neural networks', 'Deep learning']	0.770569
6	Computer_science	78268	88087	Computer science is the study of computation, ...	C	['Computer science', 'Formal sciences']	0.760875
7	Root_cause_analysis	30816	24353	In science and engineering, root cause analysi...	C	['Problem solving', 'Quality control tools']	0.750081
8	Branches_of_science	39847	46239	The branches of science, also referred to as s...	B	['Branches of science', 'Scientific disciplines']	0.748793
9	Data_structure	17005	28170	In computer science, a data structure is a dat...	C	['Data structures']	0.746238
10	Data_analysis	87968	32325	Data analysis is the process of inspecting, cl...	B	['Big data', 'Computational fields of study', ...	0.744758
11	Cognition	48637	31442	Cognition is the "mental action or process of ...	C	['Cognition', 'Cognitive psychology', 'Cogniti...	0.744027
12	Research	66088	44184	Research is "creative and systematic work unde...	B	['Ethics', 'Methodology', 'Research', 'Scienti...	0.742864
13	Science	168040	147213	Science is a rigorous, systematic endeavor tha...	B	['Main topic articles', 'Observation', 'Science']	0.741779

Step 3 - Exclude Small Articles, and Select top 20

	title	size	total_views	Introduction	article_quality	article_categories	Similarity Score	NSC
0	Data_science	23985	41348	Data science is an interdisciplinary academic ...	C	['Computational fields of study', 'Computer oc...	NaN	4
10	Data_analysis	87968	32325	Data analysis is the process of inspecting, cl...	B	['Big data', 'Computational fields of study', ...	0.744758	2
18	Natural_language_processing	54071	50731	Natural language processing (NLP) is an interd...	C	['Computational fields of study', 'Computation...	0.730104	1
1	Data	21180	45251	In common usage data (US: ; UK:) is a collect...	C	['Data', 'Data management', 'Statistical data']	0.798207	0
2	Bioinformatics	135562	26881	Bioinformatics ( ) is an interdisciplinary fie...	Unknown	['Bioinformatics']	0.783384	0
3	Methodology	97488	27475	In its most common sense, methodology is the s...	Start	['Methodology']	0.781497	0
4	Bayesian_inference	65403	23862	Bayesian inference (BAY-zee-ən or BAY-zhən) ...	B	['Bayesian inference', 'Logic and statistics',...	0.773080	0
5	Deep_learning	180889	52094	Deep learning is the subset of machine learnin...	C	['Artificial neural networks', 'Deep learning']	0.770569	0
6	Computer_science	78268	88087	Computer science is the study of computation, ...	C	['Computer science', 'Formal sciences']	0.760875	0
7	Root_cause_analysis	30816	24353	In science and engineering, root cause analysi...	C	['Problem solving', 'Quality control tools']	0.750081	0
8	Branches_of_science	39847	46239	The branches of science, also referred to as s...	B	['Branches of science', 'Scientific disciplines']	0.748793	0

Step 4 - Sort by the Number of Shared Categories (NSC), and Select Top 10

	title	size	total_views	Introduction	article_quality	article_categories	Similarity Score	NSC
0	Data_science	23985	41348	Data science is an interdisciplinary academic ...	C	['Computational fields of study', 'Computer oc...	NaN	4
1	Computer_science	78268	88087	Computer science is the study of computation, ...	C	['Computer science', 'Formal sciences']	0.760875	0
2	Deep_learning	180889	52094	Deep learning is the subset of machine learnin...	C	['Artificial neural networks', 'Deep learning']	0.770569	0
3	Natural_language_processing	54071	50731	Natural language processing (NLP) is an interd...	C	['Computational fields of study', 'Computation...	0.730104	1
4	Branches_of_science	39847	46239	The branches of science, also referred to as s...	B	['Branches of science', 'Scientific disciplines']	0.748793	0
5	Data	21180	45251	In common usage data (US: ; UK:) is a collect...	C	['Data', 'Data management', 'Statistical data']	0.798207	0

Step 5/a - Sort by the Number of Views, and Select Top 5

	title	size	total_views	Introduction	article_quality	article_categories	Similarity Score	NSC	quality_score
0	Data_science	23985	41348	Data science is an interdisciplinary academic ...	C	['Computational fields of study', 'Computer oc...	NaN	4	5
1	Branches_of_science	39847	46239	The branches of science, also referred to as s...	B	['Branches of science', 'Scientific disciplines']	0.748793	0	4
2	Computer_science	78268	88087	Computer science is the study of computation, ...	C	['Computer science', 'Formal sciences']	0.760875	0	5
3	Deep_learning	180889	52094	Deep learning is the subset of machine learnin...	C	['Artificial neural networks', 'Deep learning']	0.770569	0	5
4	Natural_language_processing	54071	50731	Natural language processing (NLP) is an interd...	C	['Computational fields of study', 'Computation...	0.730104	1	5
5	Data	21180	45251	In common usage data (US: ; UK:) is a collect...	C	['Data', 'Data management', 'Statistical data']	0.798207	0	5

Step 5/b - Sort by quality rank, Final Result

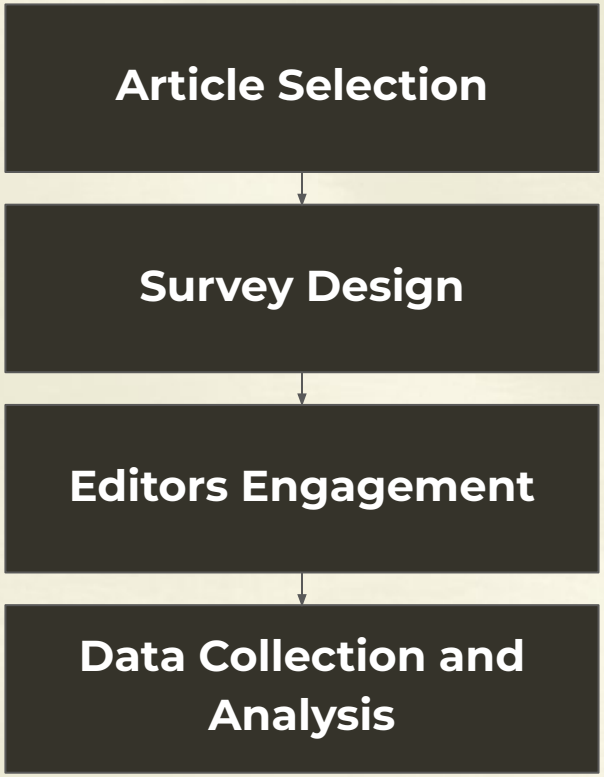
Our System List	Current See Also List
Branches_of_science	Open Data Science Conference
Computer_science	Scientific Data
Deep_learning	Women in Data
Natural_language_processing	Python (programming language)
Data	R (programming language)

Our Generated List VS The Current Manual one - Data Science Article

System Evaluation

System Evaluation

- **Initial Feedback** from some **Wikipedia admins** are **encouraging.**
- Collaborated with **Wikimedia Foundation Research Team** to design **an evaluation methodology**, surveying **top 5000 Wikipedians.**



Tested system's **effectiveness** with **200 diverse articles**

Created surveys with **anonymous** 'See Also' lists **(automated List A, manual List B)** via **Google API**

Engaging 5,000 top Wikipedia editors for diverse feedback via mass messaging

Analyzing survey responses **to compare automated vs. manual 'See Also' preferences**, identifying system's **efficiency** and **improvement areas**

Evaluation Methodology

List A	List B
Branches of science	Open Data Science Conference
Computer science	Scientific Data
Deep learning	Women in Data
Natural language processing	Python (programming language)
Data	R (programming language)

Which list of 'See Also' articles do you prefer that contains articles you are more likely to click on for related information? *

- (●) List A
- ( ) List B

Survey Example - Part 1

Please select the reason(s) for your choice. What aspects influenced your preference *

- [ ] Relevance of articles to the main topic
- [ ] Variety/diversity of articles presented
- [ ] Popularity of articles presented
- [ ] Other...

If you have any additional comments, feedback, or reasons for your preference that were not covered in the previous options, please share them with us. This could include suggestions for improvement, specific features you liked or disliked, or any other thoughts on the 'See Also' sections provided. This question is optional, but your insights would be invaluable to us.

Long-answer text

Survey Example - Part 2

Conclusion

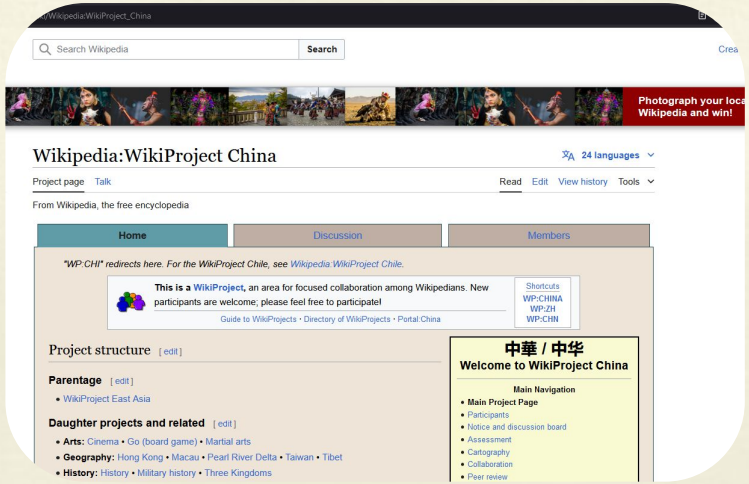
Conclusion

- **Research Question Achievement:** Successfully explored utilizing article features and NLP generated vectors for "See Also" automation.
- **Early opinions** from Wikipedia admins are **encouraging**, **BUT the full effectiveness are not assessed yet.**
- **Surveys planned for Easter;** outcomes will be detailed in the final report.

Challenges & Future Directions

Challenges & Future Directions

- **Model Assessment Challenges: Time-consuming** surveys are required for model evaluation..
- **Wikimedia Foundation's Recommendation:** Emphasizes the importance of **expert evaluations**.
- **Alternative NLP Models:** Besides **BERT**, other models like **RoBERTa**, and **DistilBERT** offer potential improvements or variations.
- **Future Directions:** Explore other NLP models, refine evaluation processes, and seek continuous feedback from Wikipedia's expert community.



WikiProject China

Questions?