HackerSearch
Data Preparation

Ana Barros - up201806593
João Costa - up201806560
João Martins - up201806436

Dataset

Data Collection

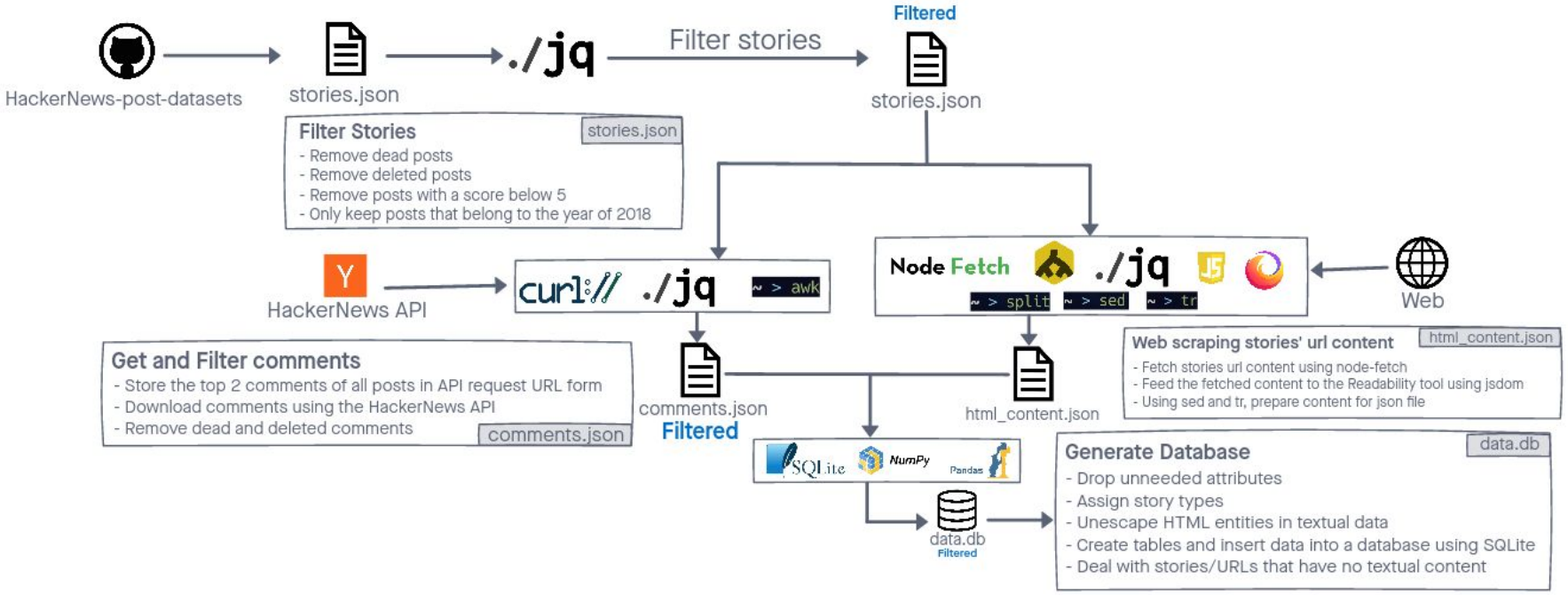
- *stories.json*: Datasets from a GitHub Repository
- *comments.json*: From the HackerNews API
- *html_content.json*: Web scraping using Node JS along with Readability by Mozilla

Dataset Size

Table	Lines	Columns
Story	37857	8
Comment	80612	6
URL	36907	4
Type	4	2
Total	155380	20

data.db (407M)

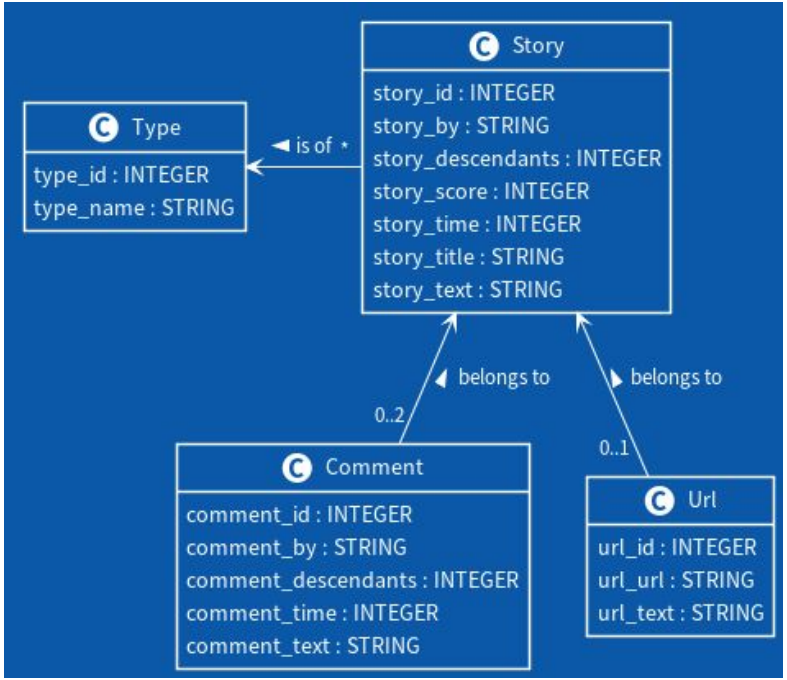
Data processing pipeline



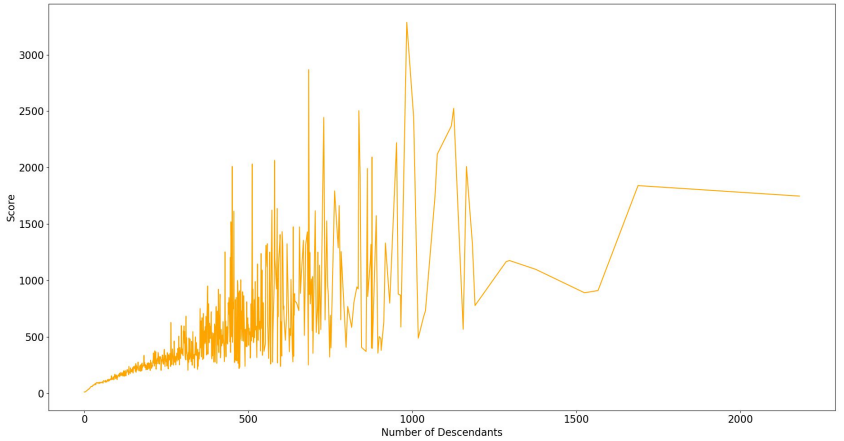
Conceptual Model

Story Types

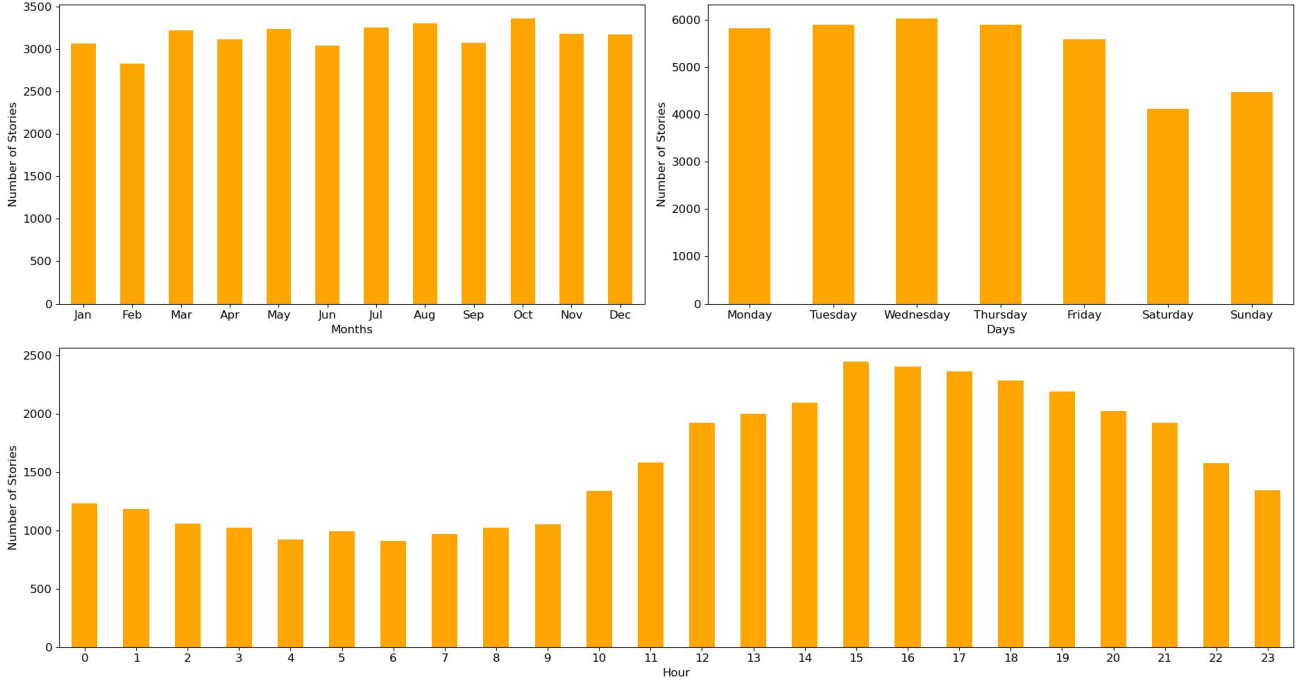
- **LaunchHN**- Stories about new “Y Combinator” backed start-ups.
- **AskHN**- Questions to the website’s community.
- **ShowHN**- Stories that usually want to share a product landing/main page.
- **Normal**- Most time have URLs that link to news, scientific articles or blog posts



Data characterization

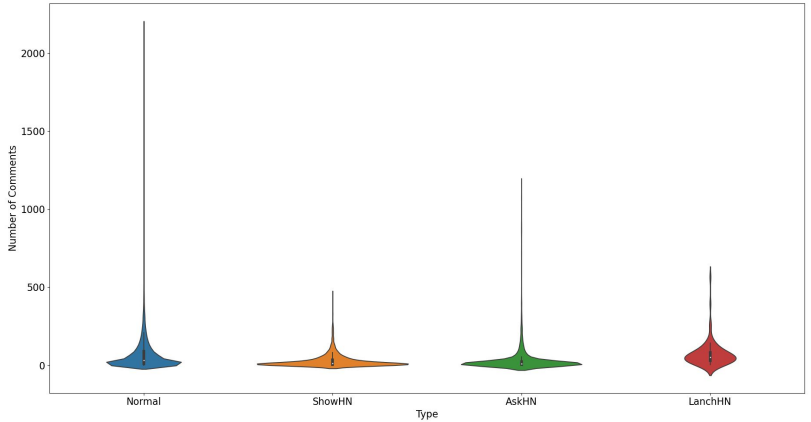


Data characterization (pt. 2)



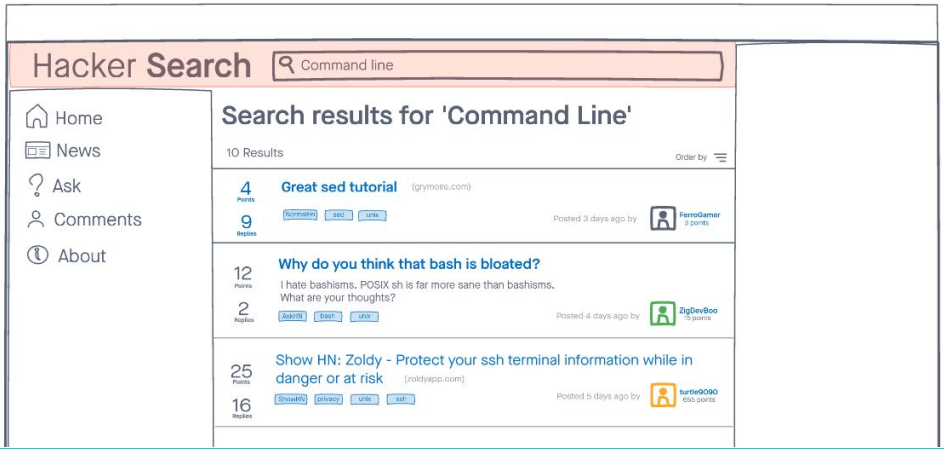
Data characterization (pt. 3)

hour	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
0	104.891	99.8994	93.6436	102.891	98.5217	105.28	121.572
2	115.845	96.096	104.261	92.6728	99.0994	124.173	93.3566
4	131.031	107.213	116.503	110.093	105.221	122.679	103.409
6	134.109	109.726	108.98	123.79	112.349	91.4455	106.189
8	98.7068	110	104.616	110.923	98.1111	101.698	87.1223
10	94.141	103.762	120.435	106.185	110.316	88.3214	96.8435
12	164.841	132.729	139.285	91.8	124.221	101.158	90.7542
14	137.424	113.747	109.421	109.014	106.36	118.276	106.948
16	115.669	122.174	125.081	121.752	123.114	87.9619	112.259
18	123.655	146.433	132.13	144.881	110.706	102.25	120.6
20	137.332	127.382	139.892	126.074	110.192	115.514	125.711
22	136.644	124.107	128.171	136.53	126.517	110.299	112.492
	127.954	127.134	133.407	129.951	125.954	106.725	113.585
	122.198	138.219	135.772	127.336	128.687	139.231	121.034
	142.456	158.071	123.406	127.559	132.428	122.767	111.126
	121.187	132.258	121.995	111.618	120.32	122.736	123.979
	144.262	143.116	133.51	129.654	110.971	113.859	143.62
	142.049	129.59	120.515	122.064	121.161	117.818	108.186
	124.257	119.65	102.45	107.351	107.397	104.256	127.088
	109.188	107.679	134.06	119.818	118.94	106.129	118.583
	106.867	106.589	120.237	113.422	99.6643	113.52	108.684
	100.097	108.253	99.0732	110.155	96.5179	86.0741	112.079
	112.333	100.325	111.749	121.08	93.8936	105.059	119.43
	127.605	116.052	138.953	119.692	91.3121	108.469	123.274



Future work

- Search for items fitting a description, e.g.: searching for ``*cli work with JSON*'' leading to results like ***jq***
- Filtering news related content from the results, e.g.: searching for ``*github*'' leading to version control system usage/guides instead of update news and policies changes of GitHub
- Ranking based on popularity
- Ranking based on relevance
- Context dependent word meaning



Possible improvements

- Collect all comments instead of just the top two comments.
- Improve web scrapping part of the pipeline to handle all websites (e.g.: YouTube.com)
- Process binary data from URLs (e.g. PDFs)
- Find real descendants count of comments
- Collect all types of posts