

School of Computer Science and Engineering

CZ4034 Information Retrieval

Assignment Report

|  |  |
| --- | --- |
| Name | Matriculation No. |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

Table of Contents

Introduction

Nowadays, there are increasingly more people who read news online rather than hard copy newspapers. This may be due to the fact that it is free of charge and it updates instantly. Although users can choose sections that they are generally interested at, it may still take time for users who have specific topics that they want to know. At the same time, searching these topics using search engines like Google may return irrelevant and unreliable information.

[APP NAME] is a web-based application which allows users to explore news reports according to their needs. Unlike conventional news websites where users need to glance through news headlines to find news of their interests, [] provides users with an easier and more convenient way of finding out what is happening around the world.

Among the popular sources of news reports, our team has chosen to build our web application mainly based on The New York Times API. Since The New York Times has been ranked as one of the ten most popular sources of citation in the English Wikipedia, we believe that it is a popular electronic media that people usually refer to, making our web developed useful and helpful for our target audiences.

Crawling

Question 1.1

*How you crawled the corpus (e.g., source, keywords, API, library) and stored them (e.g., whether a record corresponds to a file or a line, meta information like publication date, author name, record ID)*

keyword used in Reddit: subreddit “news”

We crawled 15k news from the New York Times website as the training and testing dataset for model training and 10k from Reddit (subreddit news) as a validation dataset. For New York Times, we used NYT’s API, and for subreddit news, we used pushshiftAPI to crawl news headlines, datetime and sections that they belong to.

A series of Python libraries were used for corpus to be crawled:

* requests
* json
* time
* pandas
* numpy
* datetime
* psaw
* langdetect
* re

All the meta information (headline, datetime, section) was stored into a DataFrame first. Then, data cleaning was performed. In this step, duplicates and non-English headlines were dropped, and at the same time, timestamp was converted to datetime format. Finally, the DataFrame was converted to be stored in a csv file.

Question 1.2

*What kind of information users might like to retrieve from your crawled corpus (i.e., applications), with example queries*

need to find out a word that is common enough and could be under both soft and hard news, so that we can show for example, the user will enter a word ‘corona’ and our application returns all news related to corona. Then the user might choose a soft-news option, our application will remove all coronavirus news and display only maybe ‘corona’ beer news.

Question 1.3

*The numbers of records, words, and types (i.e., unique words) in the corpus*

training dataset:

Indexing and querying

Question 2:

*Perform the following tasks:*

* *Build a simple Web interface for the search engine (e.g., Google)*
* *A simple UI for crawling and incremental indexing of new data would be a bonus (but not compulsory)*
* *Write five queries, get their results, and measure the speed of the querying*

Query 1: Amazon

Speed 1: 0.03483295440673828 s

Result 1:

[{'author': ['neroo-news-en'], 'subreddit': ['news'], 'title': ['Amazon is selling its cashier-free technology to other retailers'], 'url': ['https://en.neroonews.com/amazon-is-selling-its-cashier-free-technology-to-other-retailers/'], 'created': ['2020-03-10T00:00:00Z'], 'processed\_title': ['amazon selling cashier-free technology retailer'], 'news\_category': ['hard'], 'sentiment': ['positive'], 'id': 'b7bf6804-10f4-47fa-aff3-1dd5f3fca028', '\_version\_': 1662684534040166404}, {'author': ['121kiwi'], 'subreddit': ['news'], 'title': ['Amazon out of powered milk in the US'], 'url': ['https://121private.home.blog/2020/03/09/amazon-out-of-powered-milk-in-the-us/'], 'created': ['2020-03-09T00:00:00Z'], 'processed\_title': ['amazon powered milk u'], 'news\_category': ['hard'], 'sentiment': ['positive'], 'id': '5698a6a6-6781-42d2-ba54-6c4a7f1995af', '\_version\_': 1662684534941941768}, {'author': ['LarryPeru'], 'subreddit': ['news'], 'title': ['Amazon employee tests positive for coronavirus'], 'url': ['https://m.washingtontimes.com/news/2020/mar/3/amazon-employee-tests-positive-coronavirus/'], 'created': ['2020-03-09T00:00:00Z'], 'processed\_title': ['amazon employee test positive coronavirus'], 'news\_category': ['hard'], 'sentiment': ['neutral'], 'id': '6eacc9c8-e557-4837-93e3-48b1cced6076', '\_version\_': 1662684535969546242}, {'author': ['DICHOTOMY-REDDIT'], 'subreddit': ['news'], 'title': ['Pentagon reconsiders Microsoft contract after Amazon protest'], 'url': ['https://abcnews.go.com/Politics/wireStory/pentagon-reconsiders-microsoft-contract-amazon-protest-69581024?cid=clicksource\_4380645\_7\_mobile\_web\_only\_headlines\_headlines\_hed'], 'created': ['2020-03-13T00:00:00Z'], 'processed\_title': ['pentagon reconsiders microsoft contract amazon protest'], 'news\_category': ['hard'], 'sentiment': ['negative'], 'id': '77db48ad-e616-46fb-9bcd-df3cd8fa75eb', '\_version\_': 1662684525541457920}, {'author': ['ChartSearch'], 'subreddit': ['news'], 'title': ['Amazon launches business selling automated checkout to retailers'], 'url': ['https://www.reuters.com/article/us-amazon-com-store-technology/amazon-launches-business-selling-automated-checkout-to-retailers-idUSKBN20W0OD?feedType=RSS&amp;feedName=businessNews'], 'created': ['2020-03-09T00:00:00Z'], 'processed\_title': ['amazon launch business selling automated checkout retailer'], 'news\_category': ['hard'], 'sentiment': ['positive'], 'id': 'c8b79abc-89ff-460f-8c2b-ec5757be2c72', '\_version\_': 1662684535850008576}, {'author': ['NomadFH'], 'subreddit': ['news'], 'title': ['Pentagon reconsiders Microsoft contract after Amazon protest'], 'url': ['https://abcnews.go.com/Politics/wireStory/pentagon-reconsiders-microsoft-contract-amazon-protest-69581024'], 'created': ['2020-03-13T00:00:00Z'], 'processed\_title': ['pentagon reconsiders microsoft contract amazon protest'], 'news\_category': ['hard'], 'sentiment': ['negative'], 'id': '95a17444-b002-4621-9fdf-08da0e825f68', '\_version\_': 1662684526449524736}, {'author': ['121kiwi'], 'subreddit': ['news'], 'title': ['Amazon launches business selling automated checkout to retailers'], 'url': ['https://121private.home.blog/2020/03/09/amazon-launches-business-selling-automated-checkout-to-retailers/'], 'created': ['2020-03-09T00:00:00Z'], 'processed\_title': ['amazon launch business selling automated checkout retailer'], 'news\_category': ['hard'], 'sentiment': ['positive'], 'id': '9fbb4bad-c712-4786-b4b2-4a9f59c60caf', '\_version\_': 1662684535817502720}, {'author': ['MattW224'], 'subreddit': ['news'], 'title': ['Pentagon Asks to Reconsider Awarding Huge Cloud Contract to Amazon'], 'url': ['https://www.nytimes.com/2020/03/12/technology/amazon-jedi-defense-department.html'], 'created': ['2020-03-13T00:00:00Z'], 'processed\_title': ['pentagon asks reconsider awarding huge cloud contract amazon'], 'news\_category': ['hard'], 'sentiment': ['negative'], 'id': '8bfd0e1d-5e19-463e-a009-1fa54081f987', '\_version\_': 1662684526470496256}, {'author': ['121kiwi'], 'subreddit': ['news'], 'title': ['Millions of Amazon and eBay shoppers have data exposed online'], 'url': ['https://121private.home.blog/2020/03/12/millions-of-amazon-and-ebay-shoppers-have-data-exposed-online/'], 'created': ['2020-03-12T00:00:00Z'], 'processed\_title': ['million amazon ebay shopper data exposed online'], 'news\_category': ['hard'], 'sentiment': ['negative'], 'id': '739b7fd2-866c-401a-b0e1-dfd02a480e08', '\_version\_': 1662684529498783744}, {'author': ['Beautifully-Violent'], 'subreddit': ['news'], 'title': ['Amazon tells 750,000 employees to stay home'], 'url': ['https://www.cnbc.com/2020/03/12/amazon-tells-all-employees-to-stay-home-amid-coronavirus-fears.html'], 'created': ['2020-03-12T00:00:00Z'], 'processed\_title': ['amazon tell 750,000 employee stay home'], 'news\_category': ['hard'], 'sentiment': ['positive'], 'id': '8fe9460a-7ccd-48ec-a6c2-90e2575dbc96', '\_version\_': 1662684528106274816}]

Query 2: Facebook

Speed 2: 0.05533623695373535 s

Result 2:

[{'author': ['Juanita-Collins'], 'subreddit': ['news'], 'title': ['Buy Negative Facebook Reviews'], 'url': ['https://buy5stareviews.com/product/buy-negative-facebook-reviews/'], 'created': ['2020-03-09T00:00:00Z'], 'processed\_title': ['buy negative facebook review'], 'news\_category': ['soft'], 'sentiment': ['neutral'], 'id': '3adcad0d-5edf-4fa1-9f71-d3d8dc6a6c86', '\_version\_': 1662684535105519616}, {'author': ['Al-Andalusia'], 'subreddit': ['news'], 'title': ['Australia sues Facebook, alleges breach of user data'], 'url': ['https://www.reuters.com/article/us-australia-facebook/australia-sues-facebook-alleges-breach-of-user-data-idUSKBN20W0IW'], 'created': ['2020-03-09T00:00:00Z'], 'processed\_title': ['australia sue facebook alleges breach user data'], 'news\_category': ['hard'], 'sentiment': ['negative'], 'id': '1507bc06-2792-4724-bc7d-c548024a0bf0', '\_version\_': 1662684535780802560}, {'author': ['121kiwi'], 'subreddit': ['news'], 'title': ['Surge of Virus Misinformation Stumps Facebook and Twitter'], 'url': ['https://121private.home.blog/2020/03/09/surge-of-virus-misinformation-stumps-facebook-and-twitter-2/'], 'created': ['2020-03-09T00:00:00Z'], 'processed\_title': ['surge virus misinformation stump facebook twitter'], 'news\_category': ['hard'], 'sentiment': ['neutral'], 'id': 'e737848a-d85c-4bb0-92f7-bbf5633a71fa', '\_version\_': 1662684535001710592}, {'author': ['NotsoPG'], 'subreddit': ['news'], 'title': ['Zambia: boy arrested for allegedly defaming president on Facebook'], 'url': ['https://www.theguardian.com/world/2020/mar/11/zambia-boy-arrested-allegedly-defaming-president-facebook'], 'created': ['2020-03-12T00:00:00Z'], 'processed\_title': ['zambia boy arrested allegedly defaming president facebook'], 'news\_category': ['hard'], 'sentiment': ['negative'], 'id': 'c81c8ec8-30e0-4c6f-ab47-86ff7aa7648c', '\_version\_': 1662684529748344832}, {'author': ['neroo-news-en'], 'subreddit': ['news'], 'title': ['Facebook Bans Face Mask Ads Amid Coronavirus Outbreak'], 'url': ['https://en.neroonews.com/facebook-bans-face-mask-ads-amid-coronavirus-outbreak/'], 'created': ['2020-03-09T00:00:00Z'], 'processed\_title': ['facebook ban face mask ad amid coronavirus outbreak'], 'news\_category': ['hard'], 'sentiment': ['negative'], 'id': 'ddd4ef16-f0ea-42c2-9877-cdd0eec13bcc', '\_version\_': 1662684535132782592}, {'author': ['121kiwi'], 'subreddit': ['news'], 'title': ['Australia sues Facebook, alleges breach of user data'], 'url': ['https://121private.home.blog/2020/03/09/australia-sues-facebook-alleges-breach-of-user-data/'], 'created': ['2020-03-09T00:00:00Z'], 'processed\_title': ['australia sue facebook alleges breach user data'], 'news\_category': ['hard'], 'sentiment': ['negative'], 'id': '8e697cff-e476-4b23-9174-7e07f1e32722', '\_version\_': 1662684535830085635}, {'author': ['121kiwi'], 'subreddit': ['news'], 'title': ['Facebook Contractors Must Work in Offices During Coronavirus Pandemic'], 'url': ['https://121private.home.blog/2020/03/13/facebook-contractors-must-work-in-offices-during-coronavirus-pandemic-2/'], 'created': ['2020-03-13T00:00:00Z'], 'processed\_title': ['facebook contractor must work office during coronavirus pandemic'], 'news\_category': ['hard'], 'sentiment': ['negative'], 'id': '01046ff9-5b8c-4ba2-ae91-388581442ad7', '\_version\_': 1662684526105591808}, {'author': ['IAmTheJudasTree'], 'subreddit': ['news'], 'title': ['Facebook uncovers Russian-led troll network based in west Africa'], 'url': ['https://www.theguardian.com/technology/2020/mar/13/facebook-uncovers-russian-led-troll-network-based-in-west-africa'], 'created': ['2020-03-13T00:00:00Z'], 'processed\_title': ['facebook uncovers russian-led troll network based west africa'], 'news\_category': ['hard'], 'sentiment': ['negative'], 'id': '4f3f9b60-d633-4ec9-80e4-d3a3401de5c5', '\_version\_': 1662684526145437696}, {'author': ['roseindianet'], 'subreddit': ['news'], 'title': ['Corona Virus Scare Continues as Facebook Cancels SXSW Participation'], 'url': ['https://www.newstrackindia.com/newsdetails/2020/3/11/2-Corona-Virus-Scare-Continues-as-Facebook-Cancels-SXSW-Participation.html'], 'created': ['2020-03-11T00:00:00Z'], 'processed\_title': ['corona virus scare continues facebook cancel sxsw participation'], 'news\_category': ['soft'], 'sentiment': ['neutral'], 'id': '4e5d69c2-af01-4cd7-9b02-1dcfe33226b9', '\_version\_': 1662684532703232000}, {'author': ['mynameisalex1'], 'subreddit': ['news'], 'title': ['Google, Facebook Contractors Still Expected On Site Amid Outbreak While Full-Time Staff Works Remote'], 'url': ['https://gizmodo.com/google-facebook-contractors-still-expected-on-site-ami-1842307005'], 'created': ['2020-03-13T00:00:00Z'], 'processed\_title': ['google facebook contractor still expected on site amid outbreak while full-time staff work remote'], 'news\_category': ['hard'], 'sentiment': ['neutral'], 'id': '8f357209-3fdc-4fcf-8a8a-d333826e3b22', '\_version\_': 1662684526615199746}]

Query 3: Covid

Speed 3: 0.04396462440490723 s

Result 3:

[{'author': ['fallguy19'], 'subreddit': ['news'], 'title': ['COVID-19'], 'url': ['http://www.cidrap.umn.edu/infectious-disease-topics/covid-19'], 'created': ['2020-03-13T00:00:00Z'], 'processed\_title': ['covid-19'], 'news\_category': ['soft'], 'sentiment': ['neutral'], 'id': '433fb4cd-2483-4fa5-8b9f-18e701ec2427', '\_version\_': 1662684525893779456}, {'author': ['thehopinion'], 'subreddit': ['news'], 'title': ['Do you know all this about the Covid-19?'], 'url': ['https://www.thehopinion.com/life/health/covid-19/'], 'created': ['2020-03-11T00:00:00Z'], 'processed\_title': ['do know covid-19'], 'news\_category': ['soft'], 'sentiment': ['neutral'], 'id': 'cf200510-ca63-404a-b773-9ddbc99e3539', '\_version\_': 1662684530843058176}, {'author': ['walyc'], 'subreddit': ['news'], 'title': ['COVID-19 tests'], 'url': ['https://globalnews.ca/news/6662348/kingston-lab-covid-19/'], 'created': ['2020-03-12T00:00:00Z'], 'processed\_title': ['covid-19 test'], 'news\_category': ['soft'], 'sentiment': ['neutral'], 'id': '6f1aabc7-3b25-4776-8cae-d8fffa0564ad', '\_version\_': 1662684528042311681}, {'author': ['annieme'], 'subreddit': ['news'], 'title': ['Broadway has suspended all performances through April 12 due to COVID-19'], 'url': ['https://www.broadway.com/announcement/covid-19-update/'], 'created': ['2020-03-12T00:00:00Z'], 'processed\_title': ['broadway suspended performance april due covid-19'], 'news\_category': ['soft'], 'sentiment': ['neutral'], 'id': '1fa9d8c7-2b01-4958-865e-76f5fa10e024', '\_version\_': 1662684527963668485}, {'author': ['121kiwi'], 'subreddit': ['news'], 'title': ['COVID-19 and the campaign trail'], 'url': ['https://121private.home.blog/2020/03/13/covid-19-and-the-campaign-trail/'], 'created': ['2020-03-13T00:00:00Z'], 'processed\_title': ['covid-19 campaign trail'], 'news\_category': ['hard'], 'sentiment': ['neutral'], 'id': 'a544c0a0-9e25-4fa2-bae1-7c690205674b', '\_version\_': 1662684526801846273}, {'author': ['121kiwi'], 'subreddit': ['news'], 'title': ['Box Response to Covid-19'], 'url': ['https://121private.home.blog/2020/03/08/box-response-to-covid-19-2/'], 'created': ['2020-03-09T00:00:00Z'], 'processed\_title': ['box response covid-19'], 'news\_category': ['soft'], 'sentiment': ['positive'], 'id': '1407d24e-336f-4026-a6a5-5764425b0702', '\_version\_': 1662684536206524421}, {'author': ['nerdburg'], 'subreddit': ['news'], 'title': ['Comcast Announces Comprehensive COVID-19 Response to Help Keep Americans Connected to the Internet'], 'url': ['https://corporate.comcast.com/covid-19'], 'created': ['2020-03-13T00:00:00Z'], 'processed\_title': ['comcast announces comprehensive covid-19 response help keep american connected internet'], 'news\_category': ['hard'], 'sentiment': ['positive'], 'id': '0a6ead0f-5f27-46e9-9af2-387db4216ad8', '\_version\_': 1662684525332791296}, {'author': ['neroo-news-en'], 'subreddit': ['news'], 'title': ['COVID-19 Survivor Recovered Without Visiting Doctor'], 'url': ['https://en.neroonews.com/covid-19-survivor-recovered-without-visiting-doctor/'], 'created': ['2020-03-13T00:00:00Z'], 'processed\_title': ['covid-19 survivor recovered without visiting doctor'], 'news\_category': ['soft'], 'sentiment': ['negative'], 'id': '78d1b5b3-122f-4167-a058-255933cb9977', '\_version\_': 1662684526488322048}, {'author': ['Frocharocha'], 'subreddit': ['news'], 'title': ['CES Asia Postponed Amid COVID-19 Concerns'], 'url': ['http://www.cesasia.cn/ces-asia-postponed-amid-covid-19-concerns/'], 'created': ['2020-03-09T00:00:00Z'], 'processed\_title': ['ce asia postponed amid covid-19 concern'], 'news\_category': ['hard'], 'sentiment': ['neutral'], 'id': '5e399f65-5c83-4070-8fe7-9dfcb6d3717b', '\_version\_': 1662684535062528000}, {'author': ['HariSeldon1067'], 'subreddit': ['news'], 'title': ['Researchers Prepare COVID-19 Diagnostic For Human Testing'], 'url': ['https://medicalxpress.com/news/2020-03-covid-diagnostic-human.html'], 'created': ['2020-03-13T00:00:00Z'], 'processed\_title': ['researcher prepare covid-19 diagnostic for human testing'], 'news\_category': ['soft'], 'sentiment': ['neutral'], 'id': '7272f595-26a2-4b81-8ea0-9eb57c9d20f6', '\_version\_': 1662684524098617345}]

Query 4: Trump

Speed 4: 0.04787254333496094 s

Result 4:

[{'author': ['BreakingDC'], 'subreddit': ['news'], 'title': ['Trump to declare national emergency to combat coronavirus'], 'url': ['https://www.nbcnews.com/politics/donald-trump/trump-hold-friday-afternoon-press-conference-coronavirus-n1157981'], 'created': ['2020-03-13T00:00:00Z'], 'processed\_title': ['trump declare national emergency combat coronavirus'], 'news\_category': ['hard'], 'sentiment': ['negative'], 'id': 'a24a241a-17e8-482e-8077-51baf4c8681b', '\_version\_': 1662684526372978688}, {'author': ['luck3d'], 'subreddit': ['news'], 'title': ['Trump tests POSITIVE for Covid-19.'], 'url': ['https://worldhealthwebsite.now.sh/donald-trump-virus'], 'created': ['2020-03-13T00:00:00Z'], 'processed\_title': ['trump test positive covid-19'], 'news\_category': ['hard'], 'sentiment': ['negative'], 'id': '37f4da27-a8e3-4898-a1ba-9e046fccaed7', '\_version\_': 1662684524098617344}, {'author': ['INTJ-XP'], 'subreddit': ['news'], 'title': ["Trump and the 'New Hoax'"], 'url': ['https://www.factcheck.org/2020/03/trump-and-the-new-hoax/'], 'created': ['2020-03-09T00:00:00Z'], 'processed\_title': ["trump 'new hoax"], 'news\_category': ['hard'], 'sentiment': ['neutral'], 'id': '7af02f7a-b372-4c1b-b15b-3ae12f8641a2', '\_version\_': 1662684534925164548}, {'author': ['121kiwi'], 'subreddit': ['news'], 'title': ['Trump Declares National Emergency'], 'url': ['https://121private.home.blog/2020/03/13/trump-declares-national-emergency/'], 'created': ['2020-03-13T00:00:00Z'], 'processed\_title': ['trump declares national emergency'], 'news\_category': ['hard'], 'sentiment': ['negative'], 'id': '8f552153-fb86-4391-bd94-52298918218a', '\_version\_': 1662684525882245120}, {'author': ['121kiwi'], 'subreddit': ['news'], 'title': ['Trump Declares National Emergency'], 'url': ['https://121private.home.blog/2020/03/13/trump-declares-national-emergency-2/'], 'created': ['2020-03-13T00:00:00Z'], 'processed\_title': ['trump declares national emergency'], 'news\_category': ['hard'], 'sentiment': ['negative'], 'id': 'b7f08627-daaa-4f0a-8ce4-76a96d3f512a', '\_version\_': 1662684524911263744}, {'author': ['Ibeenjamin'], 'subreddit': ['news'], 'title': ['Trump refuses to be tested for Coronavirus'], 'url': ['https://www.nytimes.com/2020/03/12/us/politics/trump-brazil-coronavirus.html'], 'created': ['2020-03-13T00:00:00Z'], 'processed\_title': ['trump refuse tested coronavirus'], 'news\_category': ['hard'], 'sentiment': ['negative'], 'id': 'ac6df21f-f7e1-4b3a-a40f-838d213d67ba', '\_version\_': 1662684526025900032}, {'author': ['Pure\_Golden'], 'subreddit': ['news'], 'title': ['Brazilian official who met Trump testing for coronavirus - paper'], 'url': ['https://www.reuters.com/article/us-health-coronavirus-brazil-trump/brazilian-official-who-met-trump-testing-for-coronavirus-paper-idUSKBN20Z071'], 'created': ['2020-03-12T00:00:00Z'], 'processed\_title': ['brazilian official met trump testing coronavirus paper'], 'news\_category': ['hard'], 'sentiment': ['neutral'], 'id': '2a26d0c1-e7fb-4edc-ad89-481b67ff84a2', '\_version\_': 1662684528806723584}, {'author': ['Rare\_Pupper\_Dog'], 'subreddit': ['news'], 'title': ["Trump's Response to Coronavirus causes his supporters to deny it exists"], 'url': ['https://www.gq.com/story/trump-authoritarian-coronavirus'], 'created': ['2020-03-09T00:00:00Z'], 'processed\_title': ["trump 's response coronavirus cause supporter deny exists"], 'news\_category': ['hard'], 'sentiment': ['negative'], 'id': '34f2f794-cdab-4c08-a981-f1a06862b9bb', '\_version\_': 1662684536141512706}, {'author': ['traegeryyc'], 'subreddit': ['news'], 'title': ['After 163 weeks in office, Trump gets his defining crisis'], 'url': ['https://www.cbc.ca/news/world/trump-defining-crisis-1.5496203'], 'created': ['2020-03-13T00:00:00Z'], 'processed\_title': ['after week office trump get defining crisis'], 'news\_category': ['hard'], 'sentiment': ['neutral'], 'id': 'acaa359b-6e7f-4302-9516-7b53b43b26be', '\_version\_': 1662684526616248321}, {'author': ['King\_ofthe\_Wasteland'], 'subreddit': ['news'], 'title': ['Trump declares national emergency over coronavirus.'], 'url': ['https://thehill.com/homenews/administration/487473-trump-declares-national-emergency-over-coronavirus'], 'created': ['2020-03-13T00:00:00Z'], 'processed\_title': ['trump declares national emergency coronavirus'], 'news\_category': ['hard'], 'sentiment': ['negative'], 'id': 'fad73395-d9a9-4529-9054-895fba348048', '\_version\_': 1662684525610663938}]

Query 5: Singapore

Speed 5: 0.030919790267944336 s

Result 5:

[{'author': ['smithn1905'], 'subreddit': ['news'], 'title': ['Get the Best Singapore Vacation Packages'], 'url': ['https://legendstraveler.com/destinations/asia/singapore/'], 'created': ['2020-03-12T00:00:00Z'], 'processed\_title': ['get best singapore vacation package'], 'news\_category': ['soft'], 'sentiment': ['positive'], 'id': '811f6e6b-b965-4c41-ac82-c00c401334f0', '\_version\_': 1662684529493540864}, {'author': ['121kiwi'], 'subreddit': ['news'], 'title': ['Singapore charges visitors for coronavirus treatment after imported Indonesian cases'], 'url': ['https://121private.home.blog/2020/03/10/singapore-charges-visitors-for-coronavirus-treatment-after-imported-indonesian-cases/'], 'created': ['2020-03-10T00:00:00Z'], 'processed\_title': ['singapore charge visitor coronavirus treatment imported indonesian case'], 'news\_category': ['hard'], 'sentiment': ['negative'], 'id': '855eaff0-1373-42d9-a4fb-51bca95b72bc', '\_version\_': 1662684533929017344}, {'author': ['picoglobal'], 'subreddit': ['news'], 'title': ['Singapore Garden at the International Horticultural Exhibition 2019 Beijing China'], 'url': ['https://www.pico.com/en/work/specialties/singapore-garden-at-the-international-horticultural-exhibition-2019-beijing-china/?utm\_source=Social%20Reddit'], 'created': ['2020-03-10T00:00:00Z'], 'processed\_title': ['singapore garden international horticultural exhibition beijing china'], 'news\_category': ['hard'], 'sentiment': ['negative'], 'id': '73a4d676-321a-4681-b740-f7fbc5b11858', '\_version\_': 1662684533719302144}, {'author': ['121kiwi'], 'subreddit': ['news'], 'title': ['Padlet (YC W13) Is Hiring Engineers in San Francisco and Singapore'], 'url': ['https://121private.home.blog/2020/03/10/padlet-yc-w13-is-hiring-engineers-in-san-francisco-and-singapore/'], 'created': ['2020-03-10T00:00:00Z'], 'processed\_title': ['padlet yc w13 is hiring engineer san francisco singapore'], 'news\_category': ['hard'], 'sentiment': ['negative'], 'id': 'c4efebee-e0ab-4a99-8b14-0a975cc96b68', '\_version\_': 1662684534532997120}, {'author': ['dondeyo'], 'subreddit': ['news'], 'title': ['Singapore Wins Praise For Its COVID-19 Strategy. The U.S. Does Not'], 'url': ['https://www.npr.org/sections/goatsandsoda/2020/03/12/814522489/singapore-wins-praise-for-its-covid-19-strategy-the-u-s-does-not'], 'created': ['2020-03-12T00:00:00Z'], 'processed\_title': ['singapore win praise for it covid-19 strategy the u.s doe not'], 'news\_category': ['soft'], 'sentiment': ['positive'], 'id': '425785bf-699b-4e82-ad7a-f4d4f06e61d8', '\_version\_': 1662684528793092097}, {'author': ['robertacordovan'], 'subreddit': ['news'], 'title': ['buy property singapore'], 'url': ['http://realestatecompany.aircus.com/'], 'created': ['2020-03-08T00:00:00Z'], 'processed\_title': ['buy property singapore'], 'news\_category': ['soft'], 'sentiment': ['neutral'], 'id': '93bd10cb-75e8-4289-9d78-0987b9cd225d', '\_version\_': 1662684536208621571}, {'author': ['mishi1999'], 'subreddit': ['news'], 'title': ['Facebook Shuts Down Offices in London and Singapore'], 'url': ['https://www.youtube.com/watch?v=vA6eJBJoIM0&amp;feature=share'], 'created': ['2020-03-08T00:00:00Z'], 'processed\_title': ['facebook shuts down office london singapore'], 'news\_category': ['hard'], 'sentiment': ['neutral'], 'id': 'a00ca13b-f1af-419f-a5d6-4d2e5f6faaf6', '\_version\_': 1662684536449794050}, {'author': ['leavebeautifulgalaxy'], 'subreddit': ['news'], 'title': ['Singapore cybersecurity firm raises $1m from SGInnovate, Atlas, Entrepreneur First'], 'url': ['https://www.techinasia.com/right-hand-raises-1m'], 'created': ['2020-03-13T00:00:00Z'], 'processed\_title': ['singapore cybersecurity firm raise 1m sginnovate atlas entrepreneur first'], 'news\_category': ['hard'], 'sentiment': ['positive'], 'id': '3bd49536-9424-45c6-aa82-c6406ac04d81', '\_version\_': 1662684527725641729}]

Question 3:

*Explore some innovations for enhancing the indexing and ranking. Explain why they are important to solve specific problems, illustrated with examples. ( you could describe them as possible extensions of your system or future work if no time or cannot implement)*

Right now, we have added several tabs to our website that represent the different categories the news could belong to. We classified our news into the following categories: hard news, soft news, positive news, neutral news, and negative news. Each category is represented by a tab in our website. The user can click on a tab to look at news that belong to the corresponding category. For example, if the user wants to look at only hard news, he/she can just click on the “hard news” tab, and only hard news will be shown.

In addition, we also have another interesting feature. If the user wants to read a news, he/she can click on that link. The moment the user clicks on the link that leads to the actual article, our system will check if that article belongs to the soft or hard news category, and increment the count accordingly. For example, if the user clicks on a soft news article, the soft news count will be incremented. For each query the user makes, the system will compare the two counts and show results belonging to the category with the higher count, ie the type of article the user more often clicks on. For example, if hard count is greater than soft count, the system will display just the results belonging to the hard news category. Of course, the user can click on the “All” tab to view all results, or any of the other tabs to view any of the other categories.

The problem that this solves is that depending on interests and demographics, different groups of people prefer different kinds of news. Allowing them to use the tabs to filter the type of news they want makes it easier for them to narrow down on certain news. Figuring out what they tend to like and giving it to them straight away also reduces the amount of work they have to do, making it more user friendly. For example, working adults tend to prefer reading hard news, while teenagers tend to prefer reading soft news.

Classification

Question 4:

*Perform the following tasks:*

* *Motivate the choice of your classification approach in relation with the state of the art*
* *Discuss whether you had to preprocess data and why*
* *Build an evaluation dataset by manually labeling 10% of the collected data (at least 1,000 records) with an inter-annotator agreement of at least 80%*
* *Provide evaluation metrics such as precision, recall, and F-measure and discuss results*
* *Discuss performance metrics, e.g., records classified per second, and scalability of the system*
* *A simple UI for visualizing classified data would be a bonus (but not compulsory)*

Question 5:

*Explore some innovations for enhancing classification. Explain why they are important to solve specific problems, illustrated with examples.*

log down all query histories of the user and update user preference such as soft-news, so next time when the user enters some keywords the application will rank soft-news results before hard-news results.

why they are important: so when a user wants to know some promotions regarding ‘corona’ beer, he will get what he wants on the top of the returning results instead of the bottom part or even on the next page.

<https://www.youtube.com/watch?v=wnDz0Oat4Ms>

Conclusion