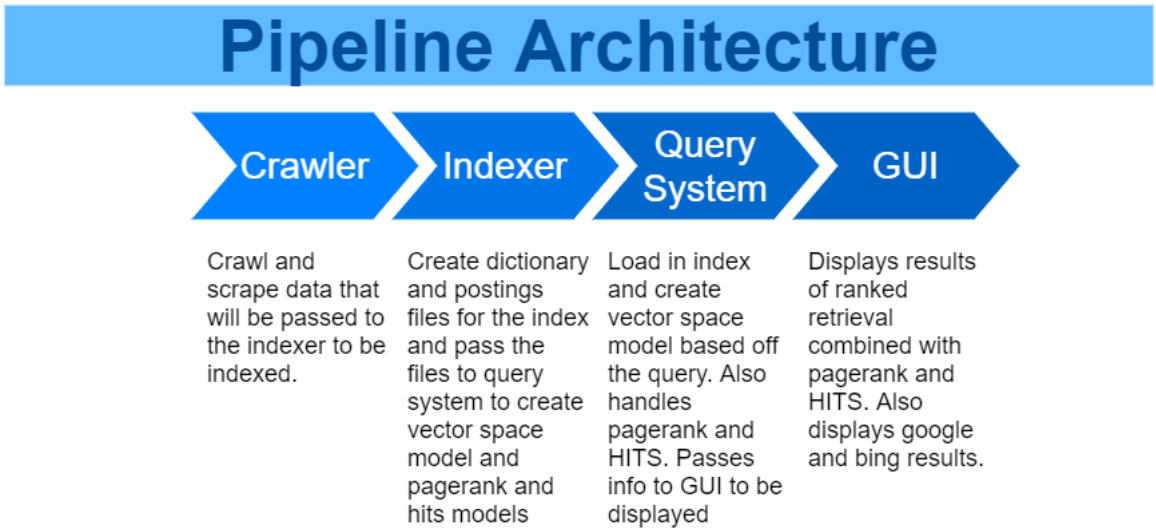
The University of Texas at Dallas
CS 6322
Information Retrieval
Spring 2021
Class Project Report

Project #3: Search engine for Startups

TEAM 03
Students: Abdelmounaim Hafid,
Abhishek Thurlapati, and Dat Quoc Ngo

The Problem



The mission for our search engine is to retrieve relevant links pertaining to the topic of 'Tech Startups'. With regards to our topic, we decided that the best links to grab would be news articles with information on current and somewhat older events in the world of startups. Abhishek Thurlapati was in charge of this task. After crawling and scraping a minimum of 100,000 web pages, we then prepped up our pages for indexing and the creation of relevance models which was done by Abdelmounaim Hafid. After this information is sent to be shown in our gui. The gui creation and the enabling of our queries was done by Dat Quoc Ngo. One thing to note for our team is that we only have three members, so we only did the first three parts of this assignment as confirmed. This assignment has made us realize the limited capacities of local machines

and the importance of cloud data storage and high level distributed data processing. The challenge we faced was in the large quantity of data. We had three main concerns. The first main concern was the ability to perform the task due to the large amounts of power it took. The second concern was using capabilities to up the speed at which we could efficiently maximize the run, but not go over the available memory, causing a crash. The last concern was formatting across multiple platforms. We each used python essentially, but the platforms which we worked on were completely different. Some constraints we had could not be resolved, mainly the high consumption of our resources and the speed at which we could perform the tasks. We chose options that allowed us to be scalable and efficient as we will cover further in the report. One thing we could resolve was the differences in our platforms. The little formatting issues and such were eased with efficient communication regarding what we each needed for our individual parts, and how we can help each other meet in the middle. Another inconvenience was that we all had to switch roles in the middle since two of our members left. That was an issue we struggled through, since we had to re-do some of our work.

Crawling

vxt160630, Abhishek Thurlapati

To perform web crawling/scraping, I used the open source tool Scrapy which is python based. Scrapy is able to provide a scalable and efficient crawler and port the information to a usable format. Using the inbuilt selectors provided within scrapy, I am able to crawl and scrape data at a high level. Scrapy outputs the data into a local csv file which I was able to share with my partner for indexing. To crawl, I had to first find suitable domains to gather the right information from. Startups was a very niche topic in the field of technology, so a lot of websites that I had to obtain from were regular tech pages. A few of them were pure startup news pages, so I used these links as a majority for my list of seeds for scraping. In total I used 8 domains : techcrunch.com, startsavant.com, tech.co, techstartups.com, entrepreneur.com, wired.com, crunchbase.com, and mashable.com. These all had dedicated articles and sections on startups and topics related to them. So, I used several pages from their sites to be able to grab more pure startup articles and not diverge into unrelated topics. Although these domains had a dedicated startup section, it is very easy for a crawler to navigate out of them as well, lowering the quality a little bit of the overall dataset.

My crawler is 100% compliant with the robots.txt files for these pages, even displaying my user agent and delaying a bit before extraction to be polite to the seed domains. I also ensure I do not hit a single domain constantly. This is possible with in built scrapy options that I can alter as needed. In addition to these options, scrapy allowed me to prevent and check for duplicates as a I follow a link. I configured a pipeline to ensure an item being exported is not currently existing. This pipeline will abort any item exportation that is a duplicate.

vxt160630, Abhishek Thurlapati

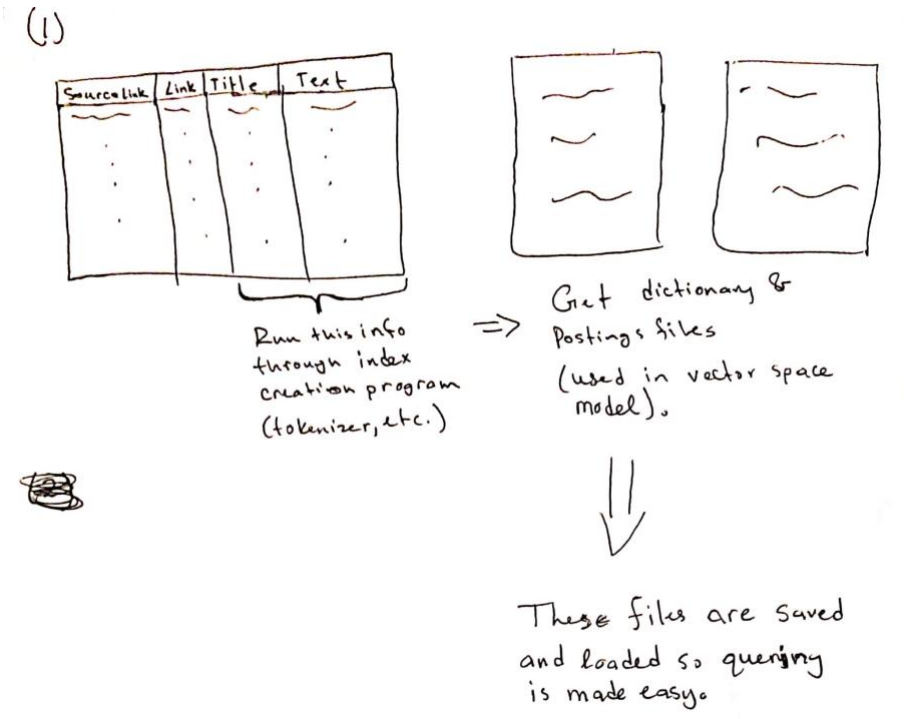
Without duplicates and null value rows, I was able to obtain a total of 100,654 pages from my crawler.

With scrapy, I was also able to parse the exact data I needed by dissecting the html architecture of the webpage. This included Source Link (the link that the current web page was directed from), Link (the current web page), Title (the title of the current page), and Text (the text content of the current page). This information allows my partner to perform and create the relevance models. As stated before, all this information was written in real time, as the data was scraped, to the csv file for easier export and storage.

Indexing and Relevance:

Abdelmounaim Hafid axh170730

The index was assembled by using a dataset from the crawler. The necessary columns were parsed and then tokenized and sent through the indexer as mentioned in the pipeline architecture diagram. I used my own index creation program to create the dictionary and posting files to be used in creating the vector space model. The web graph was created using a python library called networkx. It contained a node for each link, which is about 100,654 links give or take. The largest number of ingoing and outgoing links was 5, which is a consequence of how far we checked in terms of depth. This library created the web and had built in pagerank and hits algorithms that were used to complete the pagerank and hits parts of the search engine.



The graphs, or, more specifically the nodes, were connected to the index using the links. The links were the main identifier for the documents. The pagerank and hits scores were used as a supplement to the vector space model and are used in secondary ordering of the results. For example, a query will first display results with the highest tf-idf score and then do secondary ordering on pagerank or hits, whichever is chosen. The tf-idf weighting scheme used in this project was:

- ***tf = 1+log(tf) for both query and doc vectors***
- ***idf = log(N/df) where N is the total amount of documents and df is the document frequency. (both query and doc vectors)***
- ***tf-idf = tf*idf → Final score.***

The pagerank graph had the following highest pagerank scores:

- **https://tech.co/news/why-parents-should-worry-kids-facebook-2019-02: 0.00078421549857158**
- **https://tech.co/news/apple-revokes-facebooks-employee-only-ios-apps-2019-01: 0.00078421549857158**
- **https://tech.co/news/influencer-marketing-rise-good-reason-2016-10: 0.00078421549857158**
- **https://tech.co/news/the-key-dos-and-donts-of-influencer-marketing-2016-08: 0.00078421549857158**
- **https://tech.co/news/3-steps-towards-better-influencer-marketing-campaign-2016-02: 0.00078421549857158**

The hits algorithm calculated the following highest scores:

- **https://tech.co/news/5-things-employees-with-startup-founder-ambitions-needs-to-know-2017-04: 0.0004472271914132379**
- **https://tech.co/news/sxsw-startup-night-winners-2017-2017-03: 0.0004472271914132379**
- **https://tech.co/news/creative-ways-brand-influencers-2016-1: 0.0004472271914132379**

The algorithms used were from the package networkx, which creates the actual web graph and calculates pagerank scores and HITS scores.

When working with my team member responsible for the GUI, Dat Quoc Ngo, I made sure I had data structures that could easily be accessed by his own functions that contained all the relevant data that needed to be displayed (ranked retrieval information, pagerank, hits). I made sure that he understood how to access the information and was able to integrate it with his interface in a simple manner. I used about 10 queries to test the relevance model, where half of them

were large queries that were from certain documents to ensure that that specific document was returned at the top of the query results (all of which did so) and the remaining top ranked documents also had some relevance to the query, but not as much. I then tested some general shorter queries to see if documents with relevant results would be returned and most of the queries returned documents that had some information related to the query for example “best american startups” returned results with articles related to growing startups and top startups for women in the US.

Some limitations to this part of the project included the speed of the queries and indexing. For a large amount of links, it would take a long time to index. After indexing and saving the files, loading them in was fast enough within a reasonable range, but not fast enough to compete with modern search engines, understandably so. In order to ensure fast querying, a smaller subset of the crawled and scraped links is used for demonstration purposes.

User Interface and Comparisons with Google and Bing

dqn170000, Dat Quoc Ngo

The GUI interface is designed and created by using PySimpleGUI which is a Python-native package for GUI. Due to limited familiarity with contemporary UI packages (i.e. React.js), PySimpleGUI is utilized to create the native-application UI for the search engine. The package allows to create a flexible UI layout by defining boxes of inputs, buttons, and text outputs. The package also provides an “Event listener” that listens to users inputs and executes functions corresponding to users’ inputs. In the final GUI design, a query-input bar and a “Search” button were placed at the top of the window. Following the bar, there are 5 buttons (Vector-Space, HITS, PageRank, Google, Bing) that perform search algorithms. Finally, there are 5 text-output blocks to display search results generated by clicking the above 5 buttons.

Class Engine is created to store all data and indexes generated by Abdelmounaim Hafid in Part 2 and to host 5 functions for 5 search algorithms. Hence, every click by user calls functions of Engine class to execute corresponding search algorithms. The functions return a list of dictionaries (aka hashmap) for Title, Description, and Link. By default, *search* function is executed when users click the “Search” button to

perform the vector-space algorithm that accesses the class's indexes. The resulting *relevance-model* is saved in a class variable that is accessed by either *hits* and pagerank functions when requested. The *hits* and *pagerank* functions access the *relevance-model* to retrieve the relevant docs and sort the docs with the corresponding hubs and pagerank scores. 2 functions *google and bing* simply take query as input and make search requests to 2 search servers. To design the Engine class, I also discussed with Abdelmounaim Hafid who is responsible for Part 2 about the output format of indexes. By agreeing upon the indexes' output format, we were able to simultaneously develop Indexing and GUI.

To test the relevance models or indexes generated by Abdelmounaim Hafid, I and Abdelmounaim Hafid used inspected the scraped documents and retrieved 10 most common terms and 5 least common terms. Once all modules are connected together, we generate 5 query tests: "women in startup", "crypto", "AI", "IoT", and "Entrepreneurs " for testing. These 5 queries are common startup-related terms that our team and I believe should be used for testing when developing relevance models. Descriptions of search results of "crypto" query in the development index set is below on the following page:

dqn170000, Dat Quoc Ngo

Vector-Space	HITS	Pagerank	Google	Bing
The world of cryptocurrencies was weird to begin with, but since its beginnings as a serious form of currency in 2009, it has continued to expand in both value (at the time of writing, one bitcoin is ...	Cryptocurrency is the buzziest tech sector around. It's also one of the least understood. Pair those two facts, and you'll understand what's driving the burgeoning cottage industry of online cryptocur...	If you've never been to South by Southwest (SXSW), you're missing out. Between the cornucopia of live music, the dozens of celebrities walking the streets of Austin, and the lively atmosphere of a cit...	Crypto.com is on a mission to accelerate the world's transition to cryptocurrency. Through the Crypto.com Mobile App and Exchange, you can buy 80+ ...	The global crypto market cap is $2.24T, a 2.46 % increase over the last day. Read more The total crypto market volume over the last 24 hours is $133.87B , which makes a 6.62 % decrease.
Cryptocurrency has created some of the most exciting investment stories in the history of finance. And, like most tales of financial opportunities, a few of them have happy endings, while plenty of th...	HTC has announced the Exodus 1s, a new, cheaper version of its unique blockchain phone.The new HTC Exodus 1s phone is expected to hit the shelves at the end of Q3 (so think September) and should cost ...	Cryptocurrency is the buzziest tech sector around. It's also one of the least understood. Pair those two facts, and you'll understand what's driving the burgeoning cottage industry of online cryptocur...	Education and information about Crypto, Cryptosporidium Infection, Cryptosporidiosis, fact sheets, information for special groups, prevention and control, ...	Explore top cryptocurrencies with Crypto.com, where you can find real-time price, coins market cap, price charts, historical data and currency converter. Bookmark the Price page to get snapshots of the market and track nearly 3,000 coins. Use the social share button on our pages to engage with other crypto enthusiasts.
Cryptocurrency is the buzziest tech sector around. It's also one of the least understood. Pair those two facts, and you'll understand	The world of cryptocurrencies was weird to begin with, but since its beginnings as	The world of cryptocurrencies was weird to begin with, but since its beginnings as a	The crypto module provides cryptographic functionality that includes a	A cryptocurrency, crypto-currency, or crypto is a digital asset designed to work as a medium of

what's driving the burgeoning cottage industry of online cryptocur...	a serious form of currency in 2009, it has continued to expand in both value (at the time of writing, one bitcoin is ...	serious form of currency in 2009, it has continued to expand in both value (at the time of writing, one bitcoin is ...	set of wrappers for OpenSSL's hash, HMAC, cipher, decipher, sign, and verify functions .	exchange wherein individual coin ownership records are stored in a ledger existing in a form of a computerized database using strong cryptography to secure transaction records, to control the creation of additional coins, and to verify the transfer of coin ownership.
Opera has announced a major new update to its desktop browser today, which includes several key future-proofing features.Along with a new dark mode, Opera’s desktop browser gains a built-in cryptocurr...	Cryptocurrency has created some of the most exciting investment stories in the history of finance. And, like most tales of financial opportunities, a few of them have happy endings, while plenty of th...	Cryptocurrency has created some of the most exciting investment stories in the history of finance. And, like most tales of financial opportunities, a few of them have happy endings, while plenty of th...	Crypto token. A blockchain account can provide functions other than making payments, for example in decentralized applications or smart contracts. In this case, ...	View crypto prices and charts, including Bitcoin, Ethereum, XRP, and more. Earn free crypto. Market highlights including top gainer, highest volume, new listings, and most visited, updated every 24 hours.
HTC has announced the Exodus 1s, a new, cheaper version of its unique blockchain phone.The new HTC Exodus 1s phone is expected to hit the shelves at the end of Q3 (so think September) and should cost ...	Opera has announced a major new update to its desktop browser today, which includes several key future-proofing features.Along with a new dark mode, Opera’s desktop browser gains a built-in cryptocurr...	Opera has announced a major new update to its desktop browser today, which includes several key future-proofing features.Along with a new dark mode, Opera’s desktop browser gains a built-in cryptocurr...	Cryptocurrencies WallStreetBets Forum Members Targeted in Telegram Cryptocurrency Scam by Brandon Kochkodin relates to WallStreetBets Forum Members ...	Crypto definition is - a person who adheres or belongs secretly to a party, sect, or other group. How to use crypto in a sentence.

If you've never been to South by Southwest (SXSW), you're missing out. Between the cornucopia of live music, the dozens of celebrities walking the streets of Austin, and the lively atmosphere of a cit...	If you've never been to South by Southwest (SXSW), you're missing out. Between the cornucopia of live music, the dozens of celebrities walking the streets of Austin, and the lively atmosphere of a cit...	HTC has announced the Exodus 1s, a new, cheaper version of its unique blockchain phone.The new HTC Exodus 1s phone is expected to hit the shelves at the end of Q3 (so think September) and should cost ...	Mar 11, 2019 ... CRYPTO Official Trailer (2019) Kurt Russell, Luke Hemsworth Movie HDSubscribe to Rapid Trailer For All The Latest Movie Trailers!	Crypto.com exchange is powered by CRO, with deep liquidity, low fees and best execution prices, you can trade major cryptocurrencies like Bitcoin,Ethereum on our platform with the best experience
			The latest Tweets from Bloomberg Crypto (@crypto). A look at how cryptocurrencies and blockchain are reshaping our world from Bloomberg @ business.	View the full list of all active cryptocurrencies. Rank Name Symbol Market Cap Price Circulating Supply Volume(24h) % 1h % 24h % 7d
			Today's Cryptocurrency Prices by Market Cap. The global crypto market cap is $2.32T, a 1.42% increase over the last day. Read ...	Crypto (CTO) is a cryptocurrency . Crypto has a current supply of 13,742,738.4040553. The last known price of Crypto is 0.00433539 USD and is up 0.00 over the last 24 hours.
			Overview ▾. Package crypto collects common cryptographic constants. Index ▹. Index ▾. func RegisterHash(	Our Cryptocurrency News feed is a one stop shop destination on all the latest news in crypto. Cryptocurrency News today play an important role

			h Hash, f func() hash.Hash) ...	in the awareness and expansion of of the crypto industry, so don't miss out on all the buzz and stay in the known on all the Latest Cryptocurrency News.
			Mar 11, 2021 ... The company provides shareholders with exposure to digital currency mining as well as a portfolio of crypto-coins. The blockchain companies ...	See our list of cryptocurrency exchanges Ranked by volume Binance Coinbase Pro Huobi Kraken Bithumb Bitfinex And many more

Look at the above table, our search results by Vector-Space, HITS, and PageRank are relevant to the query input in terms of context or the word frequency. However, when looking at search results by Google and Bing, the results are more relevant to financial topics that attract people since crypto is gradually considered as a digital currency. This observation means that Google and Bing search engines are ads-biased rather than context-biased. However, the Google and Bing search engines return more exact-matching results than our search algorithms especially in titles.

The first ten search results of each search algorithm are retrieved for demonstration. This decision follows the default number of search results returned by both Google and Bing. This helps reduce the display latency. 10 search results is not a big number. However, 10 is a reasonable number of results for users to digest in a glance. While users digest the 10 results, the engine loads the next results concurrently. Despite the fact that the concurrent loading is useful in reality, this setting is not implemented in this project.

Three query examples and their first results in the development index set are:

		Vector-Space	HITS	Page-Rank	Google	Bing
Crypto	Title	What the Heck Is an ICO Anyway?	Cryptocurrency Enthusiasts Aren't Making Money the Way You'd Expect	Meet the Featured Speakers at SXSW 2018	Crypto.com | The Best Place to Buy, Sell, and Pay with Cryptocurrency	Cryptocurrency Prices, Charts And Market ... - CoinMarketCap
	Description	The world of cryptocurrencies was weird to begin with, but since its beginnings as a serious form of currency in 2009, it has continued to expand in both value (at the time of writing, one bitcoin is ...	Cryptocurrency is the buzziest tech sector around. It's also one of the least understood. Pair those two facts, and you'll understand what's driving the burgeoning cottage industry of online cryptocur...	if you've never been to South by Southwest (SXSW), you're missing out. Between the cornucopia of live music, the dozens of celebrities walking the streets of Austin, and the lively atmosphere of a cit...	Crypto.com is on a mission to accelerate the world's transition to cryptocurrency. Through the Crypto.com Mobile App and Exchange, you can buy 80+ ...	The global crypto market cap is $2.24T, a 2.46 % increase over the last day. Read more The total crypto market volume over the last 24 hours is $133.87B, which makes a 6.62 % decrease.
women in startup	Title	Tech.Co Top Stories: 15 Best Companies for Women	Tech.Co's Gadget Guide to Unique Graduation Parties	5miles Classifieds App Looks To Replace Craigslist	Products | Women Who Startup	50 Women-Led Startups That Are Crushing Tech
	Description	Did you miss out on one (or	For many, graduation day has	For years, we have been a	Women Who Startup is	As the founder of Women

		five) of Tech.Co's articles from the last week? It's okay. We want to let you know that we forgive you; we totally get it. You're a busy person, with a lot of things on you...	come and gone and it's time to start adulting. Right? Of course not! The summer before you put on your suit should be about digesting all that you've accomplished through ...	slave to Craigslist. Whether it be looking for a job or selling a couch, when all else failed, we had to turn to this haven for weirdos and creepers. The website has become a...	a Learning platform for a Global Community of Women Entrepreneurs and Innovators. Headquartered in Denver, Colorado. Learn more	Who Tech, I've launched one of the largest global programs, the Women Startup Challenge, to disrupt a culture and economy that has made it exceedingly difficult for women ...
IoT	Title	The Importance of Investing in IoT for Your Startup	The Internet of Things Is Perfect for Company Security	The Internet of Things Is Perfect for Company Security	Internet of things - Wikipedia	What Is the Internet of Things (IoT)? - Oracle
	Description	By 2020, the number of connected smart devices interacting with one another through the Internet of Things (IoT) will reach nearly 20 billion.Yes, that is a lot. However, deployment and managea	When you think of the Internet of Things, or IoT, you likely picture everything from smart homes to comprehensive entertainment systems – technology geared for consumers, not for companie	When you think of the Internet of Things, or IoT, you likely picture everything from smart homes to comprehensive entertainment systems – technology geared for consumers, not for companie	The Internet of things (IoT) describes the network of physical objects—a.k.a. "things"—that are embedded with sensors, software, and other technologies for the ...	IoT applications can collect data concerning the scope of an outage and deploy resources to help utilities recover from outages with greater speed. Healthcare. IoT asset

dqn170000, Dat Quoc Ngo

		bility ...	s. However, n...	s. However, n...		monitoring provides multiple benefits to the healthcare industry. Doctors, nurses, and orderlies often need to know the exact location of patient-assistance assets such as ...

Discussion

dqn170000, Dat Quoc Ngo
Abhishek Thurlapati, vxt160630
Abdelmounaim Hafid axh170730

Since 2 members who were responsible for X1 and X3 parts dropped the course, we had a long discussion about changing roles. The result was that we had to restart research about crawling and designing GUI. The final team was not familiar with contemporary front-end packages. Hence, our GUI design was not user-friendly as expected. Another issue we had intensively discussed was the crawling task. The crawling function was run well to crawl links up to 100,000 links. While crawling was the biggest issue that the team faced. We were able to see that there were many difficulties and intricacies that were looked over and had to be resolved, some of which were very time consuming. A helpful resource would be access to a more powerful UTD system in which we could deploy our code for faster performance. This allows us to play around with the details of the tasks and experiment to learn better the material at hand from a more practical experience.

Conclusion

Dat Quoc Ngo, dqn17000

The search engine for “Startup” provides a better understanding of distributed systems for search engines. Through the project, I observe that web crawling is the most difficult and time-consuming part due to the huge amount of data generated daily every day. To replicate a distributed system locally, we have to look for advanced packages such as scrappy and multiprocessing/threading to distribute web-crawling resources. The web-

crawling is quite scary as it is able to scrape private information in mass that the scraped information could be leaked and be an issue for data privacy. The algorithms for indexing, hits, and pagerank are easier to be implemented than I expected. However, to scale up the algorithms are challenging. In our deployment, each query search takes 2 seconds of latency. In reality, search engines like Google or Bing may serve millions of search requests at a moment that could lead to huge latency if scaling is not considered carefully.

vxt160630, Abhishek Thurlapati

In conclusion, I believe this was a great learning experience for us on the capabilities of a real search engine, and the amount of work and power that goes into creating and maintaining them. The web crawling took me much longer than I had anticipated. Free tools were a bit scarce, and documentation was lacking. I believe this side of the tech market is very underdeveloped in the community aspect. However, I can understand the reasoning behind this. Web crawling is a superpower that should be used responsibly. As we get more and more digitized (fast forwarded with the recent pandemic), we need to be cautious for cyber attacks. Web crawling and scraping is an extremely easy way to get private information, but it can also be used for research and analysis. It is indeed a double-edged sword.

Abdelmounaim Hafid, axh170730

For my final remarks, the search engine project was a project that gave us an opportunity to not only understand how a search engine worked but give us a better understanding of the difficulties and intricacies of the search engine. It gives a good perspective of what companies like Google were able to create and why its search engine is the top search engine used today. If I were to change something, I would probably change the subject, because it was difficult to find pages related strictly to startups. In terms of the indexing, I would utilize better big data tools to speed up indexing several times. I am satisfied with the speed of the querying and result display on the GUI. The project was a valuable learning experience and many of the skills and ideas learned will be used in the future.