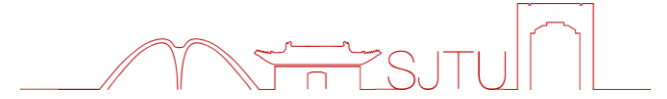




上海交通大学
SHANGHAI JIAO TONG UNIVERSITY



News Being

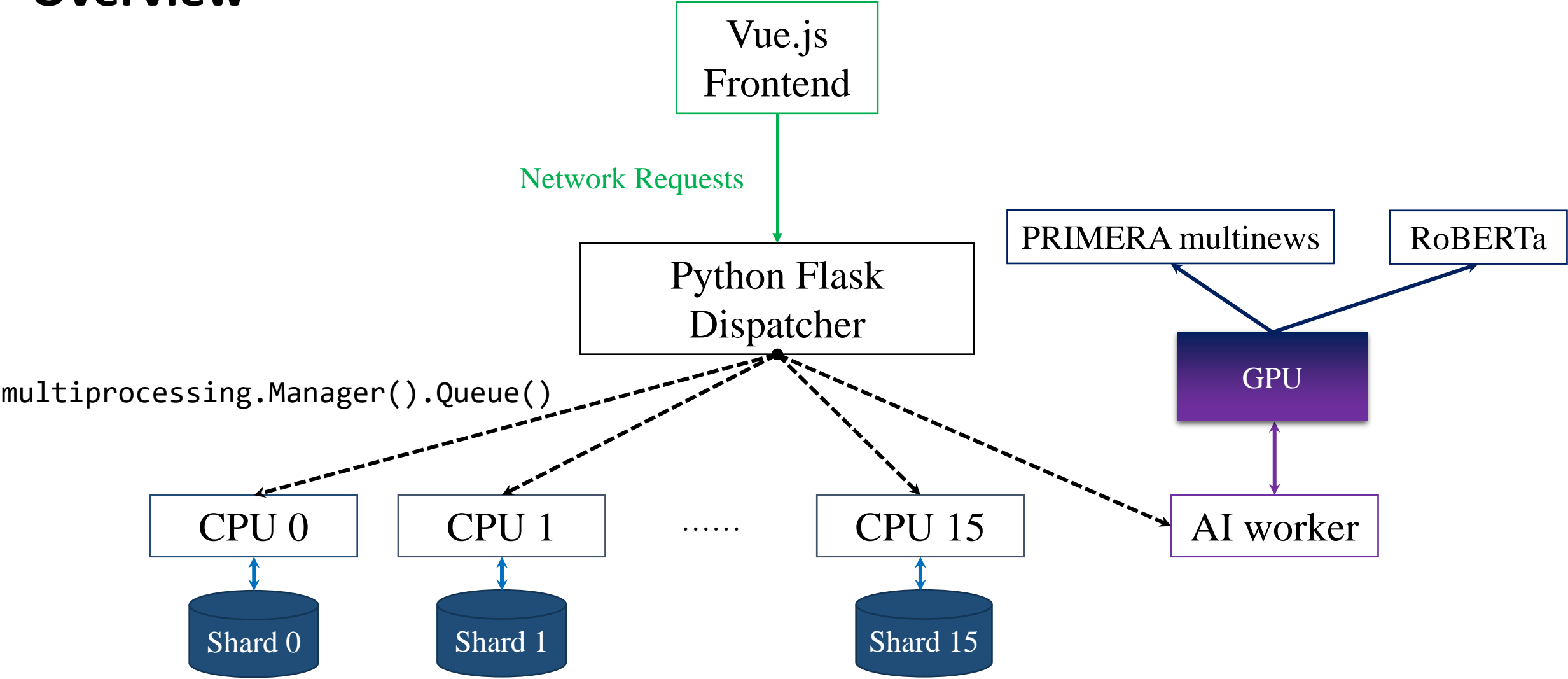
Project for CS7330
Introduction to Web Search and Mining

Zonghan Yang You Lyu Lingyue Fu Yushi Liu

June 17, 2023

饮水思源 · 爱国荣校

Overview



Overview

- **Multiprocessing**
 - Each of 16 workers preprocess and answer queries of 1/16 data.
 - Cross-platform, extensible with simple `config.json` adjust.
- **Fast and Lightweight Search Engine**
 - Process each keywords ~0.5 with cold start;
 - With hot data cache, multiple queries could reach < 0.0001 s level;
 - Reduce 37.5% of disk space usage for index with Gap Compression.
- **Useful AI-aid Search**
 - Summarization, Article Q&A, “New Bing”-like search in less than a second.
- **Friendly User Interface**
 - Beautiful UI built on Ant Design Vue.



Boolean Query

- **Feature: SortedIndex**
 - Optimize NOT operation using De Morgan's laws.
- **Attempt: C++ cross-language for query speed-up**
 - It's faster but not as expected. We deprecated the C++ implementation due to maintenance cost.

Sqlite Index	In LRU Cache	Language	Time
No	No	Python	188s
Yes	No	Python	1.78s
Yes	No	Python, C++	1.66s
Yes	Yes	Python	<0.0001s

Table 1: Latency Results

Boolean Query

- Gap Compression
 - Polyline Encoding: Gap Compression + Fixed Point Number + Base64 Encoding

Compression	Preprocess(s)	Boolean Query(s)	Ranked Query(s)	Total Size(GB)	Index Size(GB)
No	2160	0.78	0.32	63	18.71
Yes	2250	0.756	0.38	56	11.68

Table 2: Total and Index Size

Ranked Query

- Scoring: BM25 Scoring
- Challenge: SEO pollution

The screenshot displays the News Being search interface. The browser address bar shows the URL `127.0.0.1:8080/search?query=USA president`. The News Being logo is in the top left, and a search bar with the text "USA president" is in the top right. Below the search bar, it states "Ranked search: 2174560 results in 3.9664 second." The results are listed in a table with four entries:

Rank	Source	Timestamp	Snippet	Action
1	www.newser.com	2019-04-22 05:19:14	... USA! USA!	Detail
2	www.seattlepi.com	2019-04-22 08:04:11	... USA, of Eagle River. Brittany Bell, Miss Arizona USA, of Chandler. Adrielle Churchill, Miss Arkansas USA, of Dover. Nicole Michele Johnson, Miss California USA, of Westlake Village. Jessica Hartman, Miss Colorado USA, of Pueblo. Ashley Bickford, Miss Connecticut USA, of Torrington. MacKenzie Green, Miss District of Columbia USA, of Washington, D.C. Julie Citro, Miss Delaware USA, of Wilmington. Megan Clementi, Miss Florida USA, of Orlando. Cassidy Lance, Miss Georgia USA, of Savannah. Renee Mokihana Nobriga, Miss Hawaii USA, of Haleiwa. Jessica Cathleen Heilwinkel, Miss Idaho USA, of Boise. Ashley Bradarich, Miss Illinois USA, of Homer Glen. Allison Biehle, Miss Indiana USA, of North Vernon. Bethany Gerber, Miss Kansas USA, of Winfield. Katherine Anna Connors, Miss Iowa USA, of Bettendorf. Kindra Danielle Clark, Miss Kentucky USA, of Mt. Washington. Sara Brooks, Miss Louisiana USA, of Lafayette. Katie Ashley Whittier, Miss Maine USA, of New Gloucester. Simone Feldman, Miss Maryland USA, of North Potomac. Lacey Wilson, Miss Massachusetts USA, of Boston. Rima Fakh, Miss Michigan USA, of Dearborn. Courtney Basara, Miss Minnesota USA, of Duluth. Breanne Ponder, Miss Mississippi USA, of Mount Olive. Ashley Strohmier, Miss Missouri USA, of Jefferson City. Elizabeth Anseth, Miss Montana USA, of Billings. Belinda Wright, Miss ...	Detail
3	shop.cbssports.com	2019-04-20 06:42:10	Show your sportsmanship and wear an official Team USA Jersey from CBSSports.com Shop. Shop Olympics Team USA Jerseys featuring Team USA Jerseys and Team USA Basketball Jerseys. Get the official staple of Team USA. Our ground shipping is only \$4.99 so have a Team USA Olympic Jersey rushed to your door.	Detail
4	buffalonews.com	2019-04-20 21:05:17	USA's Brody Roybal celebrates his goal against Russia. Opening faceoff between the USA and Russia. Pregame ceremony's prior to the USA and Russia game. USA fans celebrate as they play Russia. USA's Brody Roybal moves the puck against Russia. USA fans celebrate a goal against Russia. USA's Adam Page	

Ranked Query

- Scoring: BM25 Scoring
- Challenge: SEO pollution
 - Solution: Heuristics on the scoring
 - Limit maximum term frequency
 - Weighted Boolean Heuristic
 - Pruning

The screenshot displays the News Being search interface. The browser address bar shows the URL `127.0.0.1:8080/search?query=USA President`. The News Being logo is in the top left, and a 'Search' button is in the top center. A search bar on the right contains the text 'USA President'. Below the search bar, a message states 'Ranked search: 243 results in 1.0028 second.' The results are listed in a table with three entries:

Rank	Source	Timestamp	Snippet	Action
1	dailycaller.com	2019-04-25 10:45:23	... USA! USA! " broke out as President Trump addressed the Conservative Political Action Conference on Friday. The chant broke out several more times during the president's speech to conservatives.	Detail
2	quotes.wsj.com	2019-04-19 10:58:10	... (USA) Ltd. Description Molecular Pharmacology (USA) Ltd. Molecular Pharmacology (USA) Ltd. engages in the development of pharmaceutical products. The firm operates through the Australia and the United States geographical segments. It also commercializes an analgesic and anti-inflammatory drug under the Triptefen brand. The company was founded on May 1, 2002 and is headquartered in Perth, Australia. Key People Molecular Pharmacology (USA) Ltd. Jeffrey David Edwards, 64 President , Chief Executive Officer, CFO & Director OBJ Ltd., Molecular Pharmacology (USA) Ltd. Average Growth Rates Molecular Pharmacology (USA) Ltd. Insider Trading Molecular Pharmacology (USA) Ltd. Ownership Molecular Pharmacology (USA) Ltd.	Detail
3	www.thewrap.com	2019-04-26 12:38:33	NBCU will has announced that development head Mark Stern will be departing and it has appointed USA Network's Bill McGoldrick to the post. He will report to Syfy president Dave Howe. Stern, who headed Syfy's original development for 11 years, will become executive producer for an unnamed new Syfy series premiering in 2014. Meanwhile at USA , Jackie de Crinis has been elevated to the newly created role of executive vice president of original programming. She will report to USA President Chris McCumber. McGoldrick served as the EVP of Original Scripted Programming for USA . He oversaw the development of its scripted dramas, including "Suits" and "Graceland." He also oversaw the network's first comedy projects, which include the upcoming "Sirens" and "Playing House." Following a four-year break from the company to oversee Spike TV's series development and the launch of its first dramas, McGoldrick returned to USA in 2009 as an SVP of original scripted programming. He originally joined the network as an executive assistant in 1998. "USA has benefited from Jackie's great taste and inspired creativity since the moment she joined the network, and it will be great to have her at the helm of the USA development team," said McCumber. ...	Detail

Summarization

- PRIMERA-multinews

News Being

Search

Chat

Ranked Search

Google Assistant

Ranked search: 166103 results in 1.2568 second.

1

[www.geek.com](#) 2019-04-21 18:27:50

...Alexa: Google Assistant is expected to be on one billion devices by the end of January. On Monday, Google announced in a blog post that Google Assistant users grew four times over the past year. Google said it predicts that Google Assistant will be available on one billion devices in a few weeks, which is up from 500 million devices in May 2018. Google Assistant also received a global update for 2019: To date, Google Assistant is now available in nearly 80 countries and 30 languages, up from 14 countries and eight languages last year. Google Assistant is now multilingual too, so customers can speak more than one language to it inside their homes. New Google Assistant features include natural conversation, multiple voice settings, and more smart home compatibility. Google improved the Google Assistant features to make the virtual voice assistant more accessible for different cultures and languages. AI advancements helped improve Google Assistant's voice capabilities. Now, Google can develop new Google Assistant voices in a shorter timeframe, and improve pitch, pace, and pauses for more human-like conversations. The Google Assistant now has eight different types of voices, including Australian and British accents. Google can help people create the ultimate ...

Detail

2

[download.cnet.com](#) 2019-04-25 12:09:07

The Google Assistant app provides another way to launch the Assistant that's already available on your phone. It allows you to quickly

Summarization

– Google Assistant is expected to be available on one billion devices by the end of January, according to a blog post from Google. The company says the number of smart devices has increased more than 600% since last year. The number of Google Assistant users grew four times over the past year, Mashable reports. The new year will find Google Assistant appearing on more than 10,000 smart home devices and in more than 80 countries and 30 languages. According to the Verge, Google Assistant will appear on 500 million devices in the first month of 2019.

Generated in 0.9793 second.

Interactive Q&A

- RoBERTa

News Detail

X

Ask me anything!

What are the rumors?

Ask



RoBERTa in 0.0801 seconds

Cyberpunk 2077 may be delayed for a few more years

Link: <https://screenrant.com/tag/cyberpunk-2077/>

Time: 2019-04-19 19:05:08

Body:

The wait for more news on **Cyberpunk 2077** may finally almost be over, with CD Projekt Red announcing the sci-fi RPG will be showcased at E3 2019.

There are plenty of amazing video games in the pipeline that may see release in 2019. These are Screen Rant's picks of the most exciting of the lot.

CD Projekt Red CEO Adam Kiciński states the studio has lofty ambitions for **Cyberpunk 2077** to be as refined as Rockstar's Red Dead Redemption 2.

CD Projekt Red is aware of player concerns regarding their choice to make **Cyberpunk 2077** in first-person but has no plans to change the perspective.

CD Projekt Red has officially revealed gameplay for **Cyberpunk 2077**, with a stream of the game over on Twitch that took over the platform.

Cyberpunk 2077 and The Witcher developer CD Projekt RED has apologized after making an inappropriate joke on the **Cyberpunk 2077** Twitter page.

New screenshots from CD Projekt RED's upcoming **Cyberpunk 2077** are the closest thing to gameplay fans of the game have seen so far.

Cyberpunk 2077 developer CD Projekt RED explains why character customization is a part of the game, in spite of it being first person.

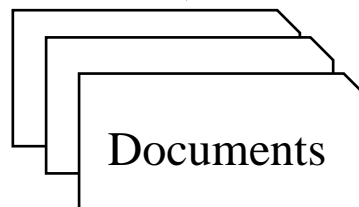
“New Bing” News Being

User: Which year did Trump win the election?

QA: What are the key words?

Generated Keywords: Donald Trump win the election

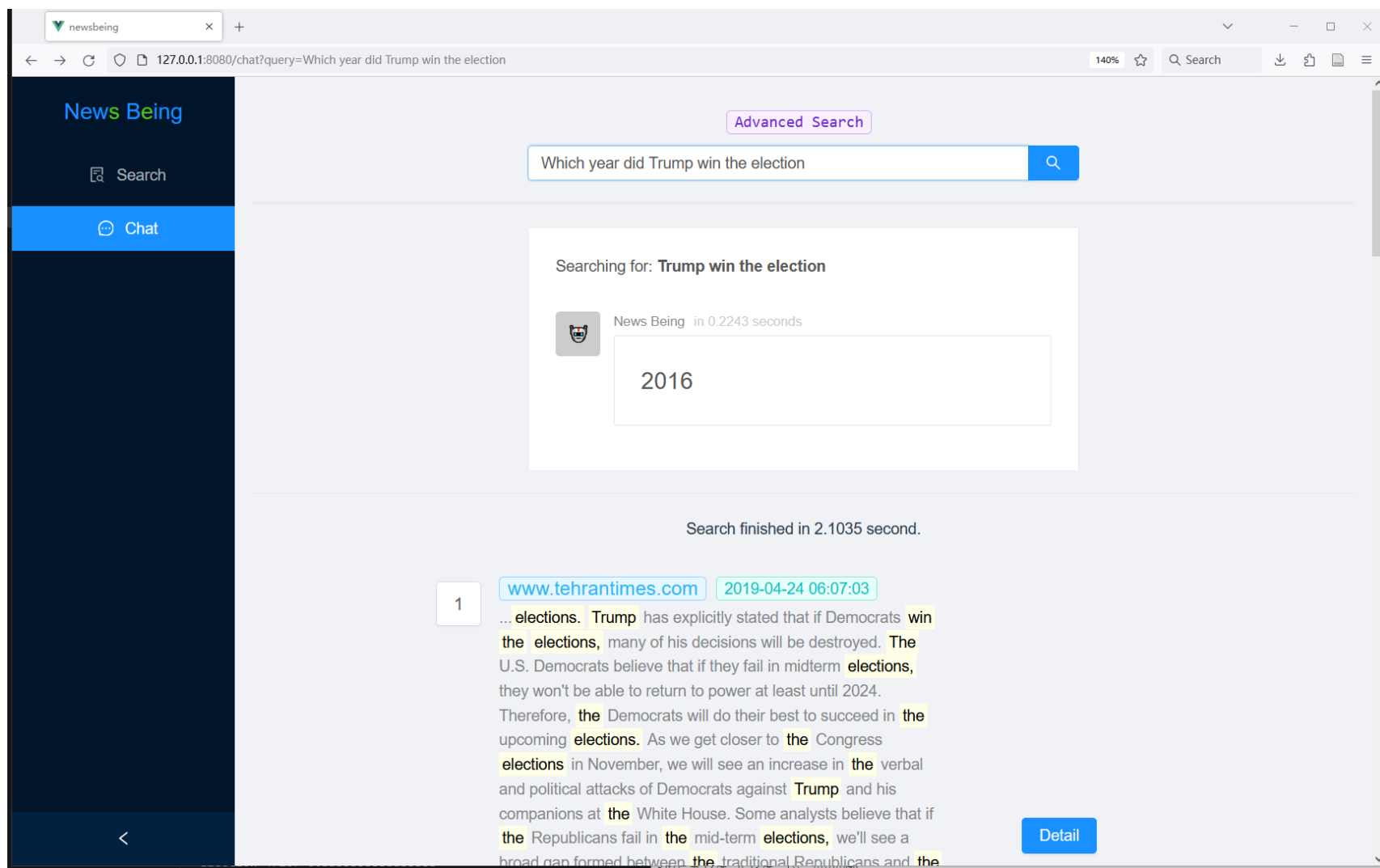
Ranked Search



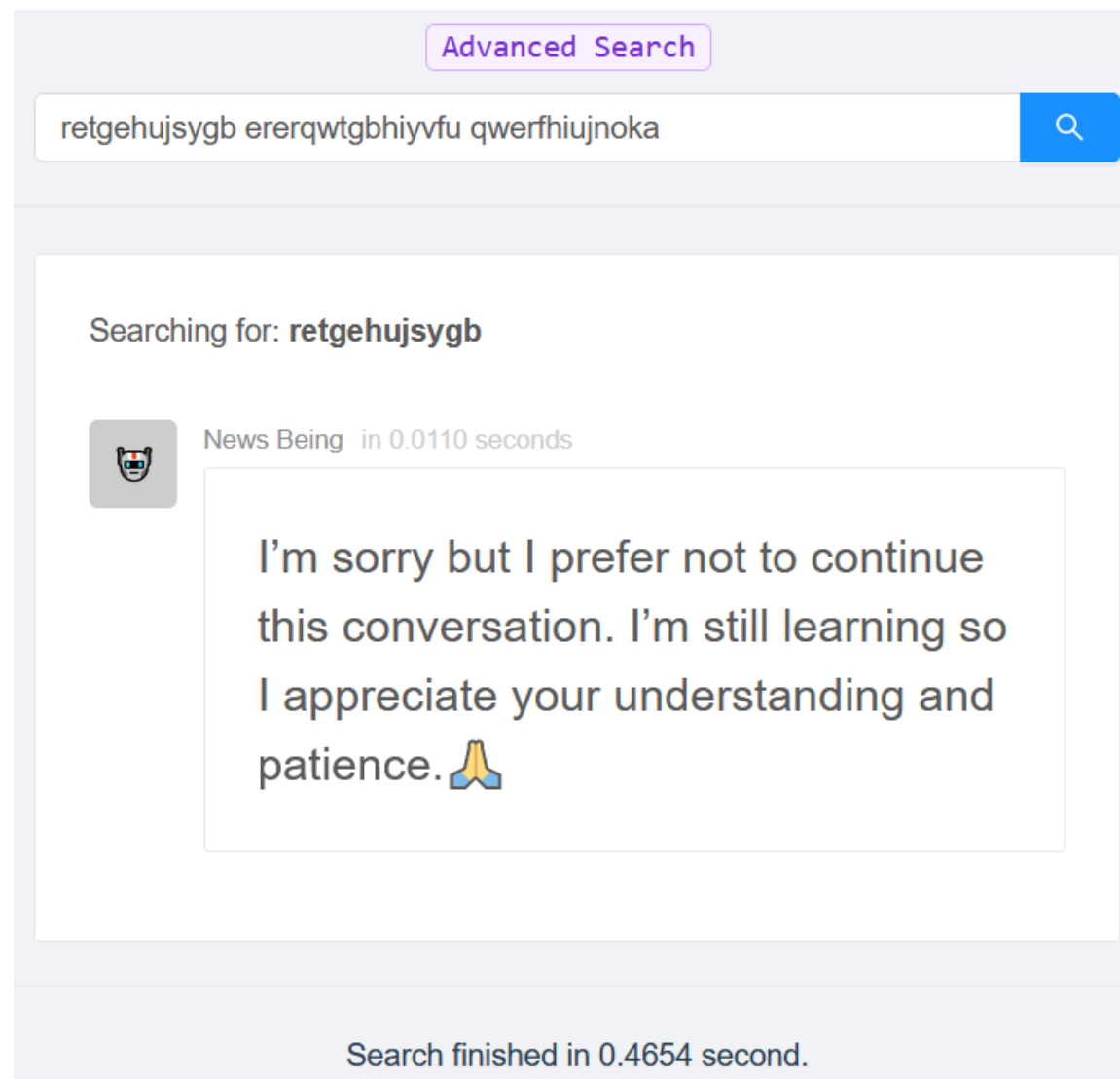
QA: Which year did Trump win the election?

 : 2016

“New Bing” News Being



One more thing...





Thank you!

饮水思源 爱国荣校