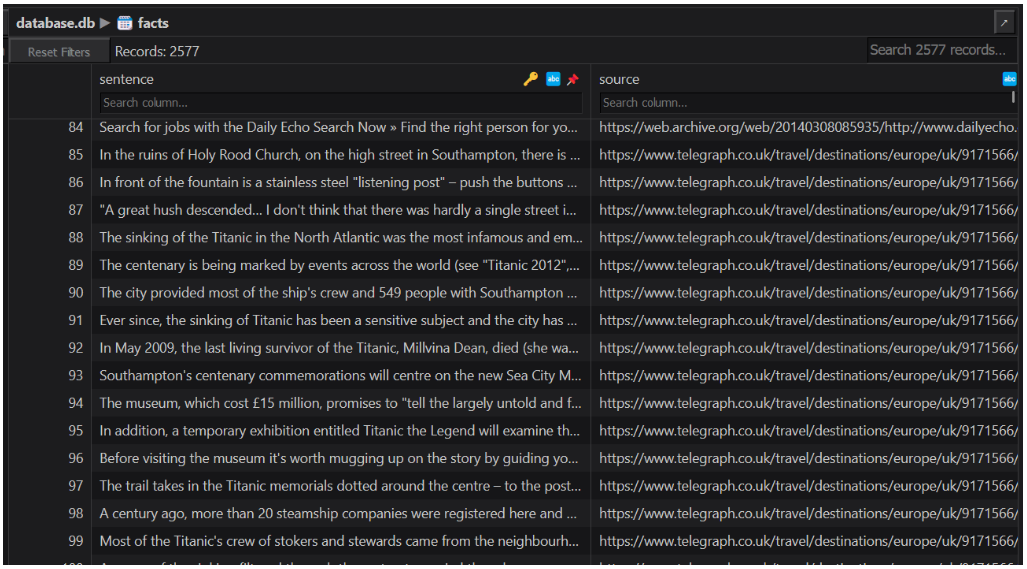
Jordan Frimpter JEF180001
Henry Kim HTK180000
CS 4395.001 Human Language Technology
Dr. Mazidi

Web Crawler Discussion

Knowledge Base

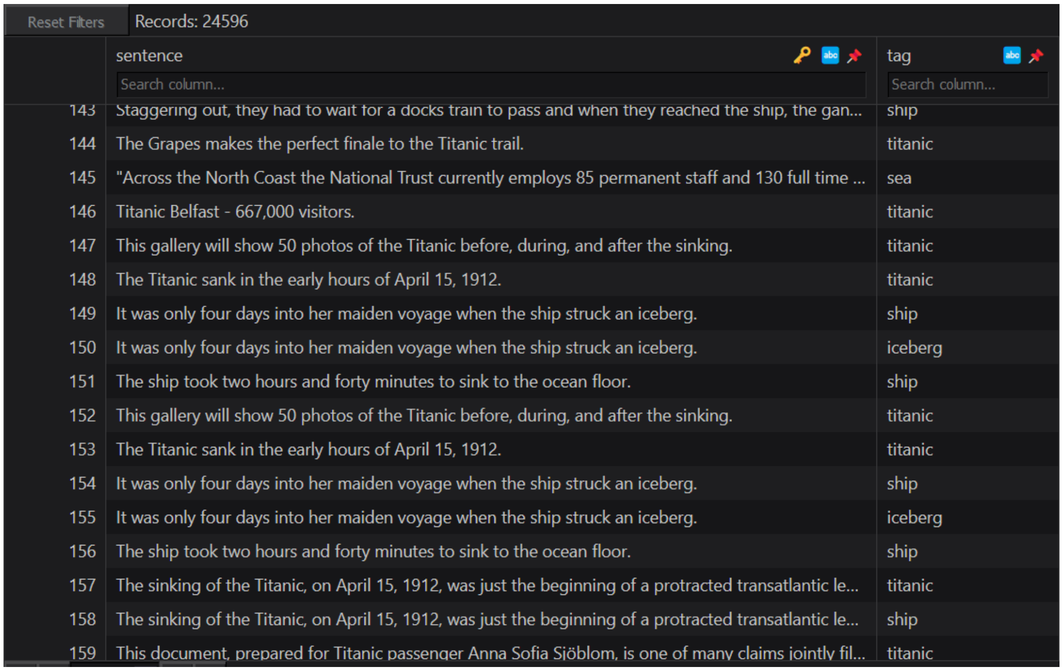
A knowledge base generated by the web scraper code on the subject of the Titanic is implemented by using the sqlite3 Python library to create a simple relational database file. The database contains two tables: facts and tags.



Records: 2577

	sentence	source
84	Search for jobs with the Daily Echo Search Now » Find the right person for yo...	https://web.archive.org/web/20140308085935/http://www.dailyecho.
85	In the ruins of Holy Rood Church, on the high street in Southampton, there is ...	https://www.telegraph.co.uk/travel/destinations/europe/uk/9171566/
86	In front of the fountain is a stainless steel "listening post" – push the buttons ...	https://www.telegraph.co.uk/travel/destinations/europe/uk/9171566/
87	"A great hush descended... I don't think that there was hardly a single street i...	https://www.telegraph.co.uk/travel/destinations/europe/uk/9171566/
88	The sinking of the Titanic in the North Atlantic was the most infamous and em...	https://www.telegraph.co.uk/travel/destinations/europe/uk/9171566/
89	The centenary is being marked by events across the world (see "Titanic 2012",...	https://www.telegraph.co.uk/travel/destinations/europe/uk/9171566/
90	The city provided most of the ship's crew and 549 people with Southampton ...	https://www.telegraph.co.uk/travel/destinations/europe/uk/9171566/
91	Ever since, the sinking of Titanic has been a sensitive subject and the city has ...	https://www.telegraph.co.uk/travel/destinations/europe/uk/9171566/
92	In May 2009, the last living survivor of the Titanic, Millvina Dean, died (she wa...	https://www.telegraph.co.uk/travel/destinations/europe/uk/9171566/
93	Southampton's centenary commemorations will centre on the new Sea City M...	https://www.telegraph.co.uk/travel/destinations/europe/uk/9171566/
94	The museum, which cost £15 million, promises to "tell the largely untold and f...	https://www.telegraph.co.uk/travel/destinations/europe/uk/9171566/
95	In addition, a temporary exhibition entitled Titanic the Legend will examine th...	https://www.telegraph.co.uk/travel/destinations/europe/uk/9171566/
96	Before visiting the museum it's worth mugging up on the story by guiding yo...	https://www.telegraph.co.uk/travel/destinations/europe/uk/9171566/
97	The trail takes in the Titanic memorials dotted around the centre – to the post...	https://www.telegraph.co.uk/travel/destinations/europe/uk/9171566/
98	A century ago, more than 20 steamship companies were registered here and ...	https://www.telegraph.co.uk/travel/destinations/europe/uk/9171566/
99	Most of the Titanic's crew of stokers and stewards came from the neighbourh...	https://www.telegraph.co.uk/travel/destinations/europe/uk/9171566/

Each entry in the facts table has a sentence and its source URL.



Records: 24596

	sentence	tag
143	Staggering out, they had to wait for a docks train to pass and when they reached the ship, the gan...	ship
144	The Grapes makes the perfect finale to the Titanic trail.	titanic
145	"Across the North Coast the National Trust currently employs 85 permanent staff and 130 full time ...	sea
146	Titanic Belfast - 667,000 visitors.	titanic
147	This gallery will show 50 photos of the Titanic before, during, and after the sinking.	titanic
148	The Titanic sank in the early hours of April 15, 1912.	titanic
149	It was only four days into her maiden voyage when the ship struck an iceberg.	ship
150	It was only four days into her maiden voyage when the ship struck an iceberg.	iceberg
151	The ship took two hours and forty minutes to sink to the ocean floor.	ship
152	This gallery will show 50 photos of the Titanic before, during, and after the sinking.	titanic
153	The Titanic sank in the early hours of April 15, 1912.	titanic
154	It was only four days into her maiden voyage when the ship struck an iceberg.	ship
155	It was only four days into her maiden voyage when the ship struck an iceberg.	iceberg
156	The ship took two hours and forty minutes to sink to the ocean floor.	ship
157	The sinking of the Titanic, on April 15, 1912, was just the beginning of a protracted transatlantic le...	titanic
158	The sinking of the Titanic, on April 15, 1912, was just the beginning of a protracted transatlantic le...	ship
159	This document, prepared for Titanic passenger Anna Sofia Sjöblom, is one of many claims jointly fil...	titanic

Each entry in the tags table has a foreign key to a sentence in the facts in the table and a tag string.

This system allows each fact entry to be tagged by multiple keywords since the sentence may be relevant to multiple subjects. Sentences were put in the database if they contained at least one of the tags. The tags used were 'titanic', 'ship', 'maritime', 'salvage', 'wreck', 'bodies', 'dead', 'iceberg', 'sea', and 'steel' which were among the 25 most common words across all scraped websites or were relevant to the topic. The scraping was limited to the <p> tags to avoid website elements like menus.

Chatbot Sample Dialogue

Given below is an example dialogue for a chatbot using the information scraped from the Titanic webpage example.

RoseBot: Hello, my name is RoseBot, and I really like talking about the Titanic (the ship, not the movie). What's your name?

Jack (user): My name is Jack.

RoseBot: Hello, Jack. Is there anything you would like to know about the Titanic?

Jack (user): Yes, when did they spot the iceberg that hit the ship?

RoseBot: Good question! The iceberg was spotted on April 14th, 1912 at around 11:40 pm by the ship's local time. Would you have boarded the Titanic in 1912 if you had the chance?

Jack (user): That depends. How many people died?

RoseBot: 706 people survived and 1,517 people died.

Jack (user): I don't like those odds. No, I would not board the Titanic.

RoseBot: I would take my chances. I think I would survive by clinging to a piece of driftwood.