

Gebresilassie Takele (gtt210000)

Biruk Mamo (bgm210001)

Date: 3/11/2023

Course: CS 4395.001 - Portfolio Assignmet 6

Web Crawler Assignment 6

The Starting URL 1 started with the Wikipedia page about the Titanic. Then we found 15 URL using the same contain of the original ideas. After that we extract each of the 15 URL data, we cleaned them up by removing newlines and tabs in the clean file text. We used the.split() method the text and then joined them using spaces. The output is a collection of very long line of text. After cleaning up we took all the text, tokenized it, lowercased it, and removed stop words and punctuation. Then we outputted the top 40 terms from the text. From these 40 terms we picked

```
my_10_terms = ["film", "Ship", "Titanic", "Atlantic", "Sink", "James", "Cameron", "story", "Ocean", "Dicaprio"].
```

Then I used these 10 terms to create a knowledge base.

```
film: ['Titanic Studios are a group of film and TV studios in the former shipyard.', 'After the 1997 film, tourists flocked to the grave because they thought this was Jack Dawson, the lead character played by Leonardo Dicaprio.', 'Legacy I edit: Dorothy Gibson survived and released her film just 31 days later "Ark wonderful".', 'But for most people nowadays, film is the medium in which they first encounter Titanic.', 'and each of these has its own take on the legend.', 'To date there have been 18 feature-length films, here's just a selection.', 'The film actress Dorothy Gibson survived through playing cards late at night, so she was among the first to go to the lifeboats.', 'Her film premiered just 31 days later, but she was traumatised by events, never worked again, and all prints of this film are lost.', 'But by the time the film was complete the tide of war had chan director was disgraced and hanged, and disaster faced Germany.', 'As for the film's subtext of overconfident authority - hell no!', 'The Six (2020): documentary film about the six Chinese survivors of the Titanic and t struggles with racism during a time when the Chinese Exclusion Act was still in force in the United States.', 'Robert Balford's film about the ship's sinking, which was the first of a series of films about the ship's sinking, was the first to be made in the United States.', 'Titanic's impact on the ship's sinking is still felt today, as the ship's sinking is still a major event in the world of film and television.', 'Titanic's impact on the ship's sinking is still felt today, as the ship's sinking is still a major event in the world of film and television.', '1997-ci ildə çəkilmiş Titanic filmi 13 il ərzində ən çox gəlir gətirən film ünvanını qoruyubdur.', '($1 845 034 188, ondan $600 788 188 - ABS -da) 2010-cu ildə çəkilmiş Titanic 2 filmi ən çox uduşuzluğa uğramış filmə bəridir.', 'Bela ki: S. O. S. Titanic filmində çəkilişləri üçün gəminin 16 metrlik modeli hazırlanmışdır.', 'Bu rekord yalnız Çeyms Kameronun Titanic filmi üçün 2 addad 19 metrlik model hazırlanmışdır keçilmiş.', 'Foxyqatla təqərrürləndirə (6-cı növbədən 17-ci Növbə), "My Heart Will Go On" mahnısı Baltazar Kconssis sənəbatı edir və filmin edilməsi səbəblərindən qonşarı məsullarının rəhbərliyinə əldə edilməsi məqsədilə qonşarı xilə edir, lakin o, Vinchestər nəvələri üçün əvvəlki edərəkən Titanic filmindən və Səlin Dionun əvvəlki bəzi səhifələrindən zəhləsi pərdəliyinə görə bu addımı atdığını deyir.', 'Titanic (1979) Titanic (1996) Nəmə vəke (Liberty) (1996) "A FI Remember" (Futuram) (1999) "Voyage of the Damned" (Doctor Who) (2007) Titanic (2012) Titanic: Krv i čelak (2012) Sənəvarənə Titanica (2012) Muzika "The Titanic (It Was Sad When That Great Ship Went Down)" (folk: The Unsinkable Molly Brown (muzika) The Sinking of the Titanic (kompozisiya) Titanic (muzika) Titanic (muzika iz filma) Back to Titanic (muzika iz filma) "My Heart Will Go On" (qışma) "Nearer, My God, to Thee" (Videoqıraq Titanic: Adventure Out of Time (1996) Dive to the Titanic (2010) Titanic VR (2018) Titanic: Honor and Glory (işıq məşhurları) Muzeyi, izləyib SeaCity Museum (Southampton) Titanic Museum (Branson, Missouri) Tita Museum (Pigeon Forge, Tennessee) Pomoraki muzeyi Atlantika (Halifax) Titanic Belfast Miestä Titanic (Kanada) Četyrti Titanic (Belfast) RT Race (Newfoundland) Groblje Fairview (Halifax) Groblje Mount Olivet (Halifax) Arzo Titanic (Oklahoma) Povestina Historičko društvo Titanic Encyclopedia Titanica Halmonas Titanicae Cyro Zene i dičica SOS ČSD RDS Olympic HMS Britannic Robert Ballard Replika Titanica Titanic II esi «esi:include src=7e fahed9598347898/content"/> Preuzeto iz: "https://bn.wikipedia.org/w/index.php?title=wp5.Titanic&oldid=3461566" Kategorije: RMS Titanic Paroli brodolovi Brodolovi iz 1911., "Gövellet izve (Kremel) | Reman ar yamam | Ti (film, 1997) NewPP limit report: Parsed by mw2388 Cached time: 20230311214923 Cache expiry: 1814400 Reduced expiry: false Complications: [no-toc:conversion] CPU time usage: 0.010 seconds Real time usage: 0.018 seconds Preprocessor visited node count: 23/1000000 Post-expand include size: 0/2097152 bytes Template argument size: 15/2097152 bytes Highest expansion depth: 4/100 Expensive parser function count: 0/500 Unstrip recursion dep Unstrip post-expand size: 0/5000000 bytes Number of Wikibase entities loaded: 0/400 Transclusion expansion time report (w,ns,call,time) 100.00% 1.999 1.999 1.999 1.999 -total Saved in parser c with key brwikik:pcache:19339-0:canonical and timestamp 20230311214923 and revision ID 1835495.", "": milionär John Jacob Astor IV a jeho manželka Madeleine Force Astorová, průmyslník Benjamin Guggenheim, majite obchodního řetězce Macy's: Isidor Straus a jeho manželka Ida, milionářka Margaret "Molly" Brownová z Denveru, později přezdívaná "Nepotopitelná Molly" pro pomoc ostatním při ztroskotání, sir Cosmo Duff-Gordon a jeho manželka, ední návrhářka Lucy Duffová-Gordonová, George Dunton Widener a jeho manželka Eleanor se synem Horjez, hráčem kriketu, obchodník John Borland Thayer s manželkou Marian a sedmiletým synem Jackem, rovin William Thomas Stead, který v roce 1892 napsal knihu o zámožném parníku potopeném ledovou krou, hraběna Rothes, asistent prezidenta Spojených států amerických Archibald Butt, spisovatelka a socioložka Helen Chur Candeeová, spisovatel Jacques Futrelle s manželkou May a jejich přátelé, producenti z Broadwaye, Henry a Rene Harrisovi, herečka německého filmu Dorothy Gibsonová a další.", "Poslední přeživší z Titanicu, která měla na vzpomínky, byla tehdy sedmiletá holčička Eva Hartová, která zemřela ve věku 91 let v Londýně dne 16. února 1996 (rok před tím, než byl uveden do kin oscarový film založený na vzpomínkách poslední přeživší, které tehdy m v 1. let). "Epizodová vzore s roztomilými epizodami, která někdy líná ve vzpomínkách neuvěřitelné, odpovídá smyšlené scéně z německého filmu o Titaniku, který se v Československu promítal v době německé okupace.", "Třic
```

Sample dialogue:

- Chatbot: Hi My name is Chatbot, what's your name
- User: My name is Gersie.
- Chatbot: do you know the film Titanic?
- User: Yeah, I know it
- Chatbot: That's good! Who's your favorite character?
- User: Rose, do you know the director of the movie?
- Chat bot: James Cameron