

Author: Yunbin Tu

System Design

The system is using flask and jinja, implementing a standard MVC framework.

Api

- "/" the home page for the service
- "/results" method=post using post request to retrieve matched queries from database, storing the result in the session.
- "/results/<int:page_id>" method=post using post and a path parameter to get the coordinate page of the result, the function next_page(page_id) is called, retrieve the result from the session.
- "/doc_data/<doc_id>" using get method and a path parameter doc_id to get the coordinate detail of doc

Frontend

- Doc.html
A template showing the author, date, content etc. of a specific doc.
- Home.html
Home page for the app, contains a search box for the user input.
- Result.html
Result page, showing the result, has a next button if there's more content in the next page.

Data Storage

Mongodb database was used to store index and wapo_doc.

collection called "wapo_docs"

- add a unique ascending index on the key "id"
- insert documents into the "wapo_docs" collection

One collection called "vs_index":

- add a unique ascending index on the key "term"
- insert posting lists (index_list) into the "inverted_index" collection\

Other collection called "doc_len_index"

- add a unique ascending index on the key "doc_id"

- insert list of document vector length (index_list) into the "doc_len_index" collection
Session was used to store the temporary query results.

Description and test queries

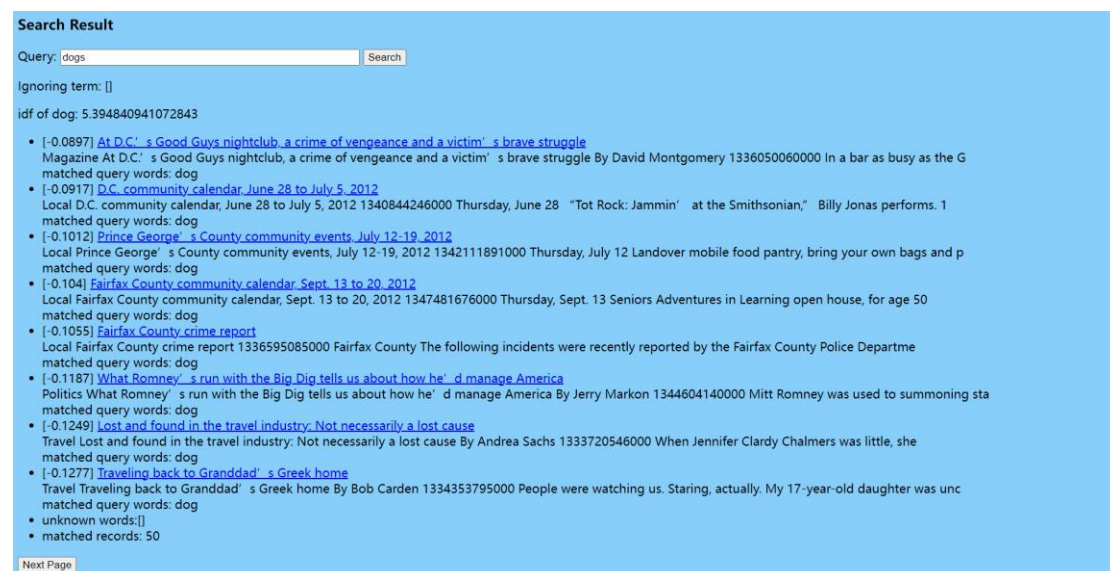
Small test file

The first 10 records was selected from the original file, building a new file call smalltest.jl.
Then using the first 10000 records to test the actual behavior when scaling up.

Test results

Note that the parameter k in topk is selected to 50.

Query: dogs



The screenshot shows a search interface with a blue header. Below the header, there is a search bar containing the text 'dogs' and a 'Search' button. Below the search bar, it says 'Ignoring term: []'. Then, it displays 'idf of dog: 5.394840941072843'. A list of search results follows, each starting with a score in brackets and a link to a document. The results are: 1. [-0.0897] [At D.C.' s Good Guys nightclub, a crime of vengeance and a victim' s brave struggle](#) Magazine At D.C.' s Good Guys nightclub, a crime of vengeance and a victim' s brave struggle By David Montgomery 1336050060000 In a bar as busy as the G matched query words: dog. 2. [-0.0917] [D.C. community calendar June 28 to July 5, 2012](#) Local D.C. community calendar, June 28 to July 5, 2012 1340844246000 Thursday, June 28 "Tot Rock: Jammin' at the Smithsonian," Billy Jonas performs. 1 matched query words: dog. 3. [-0.1012] [Prince George' s County community events July 12-19, 2012](#) Local Prince George' s County community events, July 12-19, 2012 1342111891000 Thursday, July 12 Landover mobile food pantry, bring your own bags and p matched query words: dog. 4. [-0.104] [Fairfax County community calendar Sept. 13 to 20, 2012](#) Local Fairfax County community calendar, Sept. 13 to 20, 2012 1347481676000 Thursday, Sept. 13 Seniors Adventures in Learning open house, for age 50 matched query words: dog. 5. [-0.1055] [Fairfax County crime report](#) Local Fairfax County crime report 1336595085000 Fairfax County The following incidents were recently reported by the Fairfax County Police Departme matched query words: dog. 6. [-0.1187] [What Romney' s run with the Big Dig tells us about how he' d manage America](#) Politics What Romney' s run with the Big Dig tells us about how he' d manage America By Jerry Markon 1344604140000 Mitt Romney was used to summoning sta matched query words: dog. 7. [-0.1249] [Lost and found in the travel industry: Not necessarily a lost cause](#) Travel Lost and found in the travel industry: Not necessarily a lost cause By Andrea Sachs 1333720546000 When Jennifer Clardy Chalmers was little, she matched query words: dog. 8. [-0.1277] [Traveling back to Granddad' s Greek home](#) Travel Traveling back to Granddad' s Greek home By Bob Carden 1334353795000 People were watching us. Staring, actually. My 17-year-old daughter was unc matched query words: dog. At the bottom, it says 'unknown words:[]' and 'matched records: 50'. There is a 'Next Page' button at the very bottom.

One word query working. The matching is listed from the highest similarity to the lowest. Only the top 50 mathcing records was selected.

Query: dogs human

Search Result

Query:

Ignoring term: []

idf of dog: 5.394840941072843

idf of human: 3.7577899292032333

- [-0.0917] [D.C. community calendar June 28 to July 5, 2012](#)
Local D.C. community calendar, June 28 to July 5, 2012 1340844246000 Thursday, June 28 "Tot Rock: Jammin' at the Smithsonian," Billy Jonas performs. 1
matched query words: dog
- [-0.0919] [Magnetic stimulation helps some people with treatment-resistant depression](#)
Health & Science Magnetic stimulation helps some people with treatment-resistant depression By by Julie Appleby Kaiser Health News and Julie Appleby 1
matched query words: human
- [-0.092] [Congress finally gets its business done. But first, lots of something else.](#)
Politics Congress finally gets its business done. But first, lots of something else. By David A. Fahrenthold 1341181800000 This is what the most detes
matched query words: human
- [-0.0921] [Antipsychotic drugs grow more popular for patients without mental illness](#)
Health & Science Antipsychotic drugs grow more popular for patients without mental illness By Sandra G. Boodman and Kaiser Health News 1331586444000 A
matched query words: human
- [-0.0925] [Prince William County community calendar March 25 to April 1, 2012](#)
Local Prince William County community calendar, March 25 to April 1, 2012 1332774161000 Sunday, March 25 "From Oquossoc to Occoquan: An Artist's Journ
matched query words: human
- [-0.0926] [CES 2012: Wrap-up of the show's best and most bizarre gadgets](#)
Business CES 2012: Wrap-up of the show's best and most bizarre gadgets 1326845979000 Scores of new products were unveiled at this year's Consumer Elec
matched query words: human
- [-0.0935] [Prince William County community calendar March 22 to 29, 2012](#)
Local Prince William County community calendar, March 22 to 29, 2012 1332338776000 Thursday, March 22 Injured Marines exhibit, "Focus on Ability: Cele
matched query words: human
- [-0.0938] [Tough economic times bring a more subdued approach to kitchen renovations](#)
Real Estate Tough economic times bring a more subdued approach to kitchen renovations By Deborah K. Dietsch 1327502460000 Kitchen remodeling used to b
matched query words: human
- unknown words:[]
- matched records: 50

Notice that dog has higher idf than human, which means 'dog' is more informational than 'human', which is more or less expected.

Query: on which

Search Result

Query:

Ignoring term: ['on', 'which']

Unknown search term: []

- unknown words:[]
- matched records: 0

They are both stop words, so they are just neglected. And showing no results could be acceptable though we could present an error message to tell the user why.

Query: common monster

Search Result

Query:

Ignoring term: []

idf of common: 4.2875469279009355

idf of monster: 8.16037568743582

- [-0.1084] [Battle over athlete Jim Thorpe' s burial site continues](#)
Magazine Battle over athlete Jim Thorpe' s burial site continues By Neely Tucker 1331827492000 Corra
matched query words: common monster
- [-0.1087] [Man' s persistent rash seemed ordinary, but it was actually ominous](#)
Health & Science Man' s persistent rash seemed ordinary, but it was actually ominous By Sandra G. Bo
matched query words: common
- [-0.1091] [Angry about inequality? Don' t blame the rich.](#)
Opinions Angry about inequality? Don' t blame the rich. By James Q. Wilson 1327613760000 There is
matched query words: common
- [-0.1099] [At Matchbox restaurants, lunch is big business](#)
Capital Business At Matchbox restaurants, lunch is big business By Thomas Heath 1347832297000 It is
matched query words: common
- [-0.1133] [Aging power grid on overload as U.S. demands more electricity](#)
Transportation Aging power grid on overload as U.S. demands more electricity By Ashley Halsey III 134
matched query words: common
- [-0.1133] [In Md., fear for the turtles](#)
Local In Md., fear for the turtles By Katherine Shaver 1329097080000 Maryland biologists studying box
matched query words: common
- [-0.1135] [House hunters find it' s a jungle](#)
Real Estate House hunters find it' s a jungle By Olga Khazan 1336150500000 On a recent rainy Sunday
matched query words: common
- [-0.1143] [Review: Phillips Collection show provides perspective to photos' role in paintings](#)
Style Review: Phillips Collection show provides perspective to photos' role in paintings By Philip Kenn
matched query words: common
- unknown words:[]
- matched records: 50

Monster is much more informational than common, so the idf is much higher.

Query: big bad cow animal

Search Result

Query:

Ignoring term: []

idf of big: 3.260716661058378

idf of bad: 4.031092670490855

idf of cow: 7.43790966296473

idf of anim: 4.831252091144255

- [-0.047] [D.C. community calendar Sept. 20 to 27, 2012](#)
Local D.C. community calendar, Sept. 20 to 27, 2012 1348119875000 Thursday, Sept. 20 Cathedral bird walk and woods exploration, birder Sheila Cochran I
matched query words: big
- [-0.0563] [After the death of Jack Kevorkian, a new public face of American assisted suicide](#)
Magazine After the death of Jack Kevorkian, a new public face of American assisted suicide By Manuel Roig-Franzia 1326986690000 Dust speckles shelves
matched query words: bad
- [-0.0595] [Charitable donations benefit telemarketers](#)
Business Charitable donations benefit telemarketers By David Evans 1347745260000 Carol Patterson was waiting for a call from her doctor. When the phon
matched query words: big
- [-0.0639] [State of the Union 2013: President Obama' s address to Congress \(Transcript\)](#)
None State of the Union 2013: President Obama' s address to Congress (Transcript) 1327525740000 Here is a full transcript of President Obama' s 2013 Sta
matched query words: big bad
- [-0.0695] [Marco Rubio' s grandfather had difficult transition to U.S.](#)
Style Marco Rubio' s grandfather had difficult transition to U.S. By Manuel Roig-Franzia 1339980720000 An excerpt from "The Rise of Marco Rubio" by Man
matched query words: big
- [-0.0697] [Mitt Romney' s road to Tampa](#)
Politics Mitt Romney' s road to Tampa By Karen Tumulty 1345931700000 Just under a year ago, Mitt Romney was looking at what promised to be a rough even
matched query words: big
- [-0.071] [High-tech vs. no-tech: D.C. area schools take opposite approaches to education](#)
Technology High-tech vs. no-tech: D.C. area schools take opposite approaches to education By Cecilia Kang 1336825320000 The sixth-graders are lighting
matched query words: big

Working for 4 words queries, taking about 3-4 seconds to finish.

Query: fairfax county

Search Result

Query:

Ignoring term: []

idf of fairfax: 5.384271699362657

idf of counti: 3.665320159067802

- [-0.1103] [A Local Life: Malcolm Davis, 74, pastor-turned-potter ministered through clay](#)
Obituaries A Local Life: Malcolm Davis, 74, pastor-turned-potter ministered through clay By Matt Schudel 1326588672000 As early as 1960, Malcolm Davis matched query words: counti
- [-0.1107] [Federal study estimates 1 in 88 children has symptoms of autism](#)
Health & Science Federal study estimates 1 in 88 children has symptoms of autism By David Brown 1333052940000 About 1 in 88 children in the United Sta matched query words: counti
- [-0.1112] [Maryland set to ban arsenic-containing drug in chicken feed](#)
Health & Science Maryland set to ban arsenic-containing drug in chicken feed By Darryl Fears 1337550777000 At his family farm on Maryland' s Eastern Sh matched query words: counti
- [-0.1114] [New North Korean leader Kim Jong Eun speaks publicly for first time](#)
Asia & Pacific New North Korean leader Kim Jong Eun speaks publicly for first time By Chico Harlan 1334507400000 SEOUL — Newly in charge of a country matched query words: counti
- [-0.1118] [Escapes: A tour of western Pennsylvania' s Amish country](#)
Travel Escapes: A tour of western Pennsylvania' s Amish country By Lindsay J. Westley 1351195181000 Correction: An earlier version of this article inco matched query words: counti
- [-0.1132] [D.C. drivers hurt by tough interpretation of Va. offenses](#)
D.C. Politics D.C. drivers hurt by tough interpretation of Va. offenses By Mike DeBonis 1343685195000 Yes, Tom Selden admits, he was speeding. Maybe h matched query words: counti
- [-0.114] [Justice Department trains prosecutors to combat cyber-espionage](#)
National Security Justice Department trains prosecutors to combat cyber-espionage By Sari Horwitz 1343250780000 Confronting a growing threat to nation matched query words: counti
- [-0.1151] [Alexandria and Arlington events, May 3 to 10, 2012](#)
Local Alexandria and Arlington events, May 3 to 10, 2012 1335982957000 Thursday, May 3 "Conversations With My Mother" exhibit, sculpture by Elissa Fa matched query words: fairfax counti

• unknown words:[]
• matched records: 50

Notice that 2 matched words is working as expected as shown in the last record of the page.

Query: fjdkfdkj

Search Result

Query:

Ignoring term: []

Unknown search term: ['fjdkfdkj']

- unknown words:['fjdkfdkj']
- matched records: 0

Shows unknown as expected