CSCI-572 Information Retrieval and Web Search Engines

Spring 2019 Assignment 5

Adding Spell Checking, AutoComplete and Snippets to Your Search Engine

Approach

The Autocomplete Feature: To implement autocomplete the FuzzyLookupFactory from Solr / Lucene was leveraged. The following configurations were made in solrconfig.xml

Introducing Suggest Component in Solr	Introducing Suggest Handler in Solr	AND as default for Phrase search
<searchComponent class="solr.SuggestComponent" name="suggest"> <lst name="suggester"> <str name="name">suggest</str> <str name="lookupImpl">FuzzyLookupFactory</str> <str name="field">_text_</str> <str name="suggestAnalyzerFieldType">string</str> </lst> </searchComponent>	*<requestHandler class="solr.SearchHandler" name="/suggest"> <lst name="defaults"> <str name="suggest">true</str> <str name="suggest.count">5</str> <str name="suggest.dictionary">suggest</str> </lst> <arr name="components"> <str>suggest</str> </arr> </requestHandler>*	*<lst name="defaults"> <str name="q.op">AND</str> <str name="echoParams">explicit</str> <int name="rows">10</int> <str name="df">_text_</str> </lst>*

On the client side, the search engine text box was modified to listen to any “key-up” events. So, whenever a user types in a query keyword, an asynchronous HTTP call is made to Solr suggest handler to retrieve the list of all suggestions corresponding to that keyword or phrase. Once the suggestions are retrieved, the suggestions list under the search textbox is populated with the results. User has the ability to either use the up and down arrow key to select a suggestion or use the mouse. The background of an active suggestion in the list is highlighted when a user hovers over it or uses the arrow key to navigate.

Spelling Correction Feature: An existing JavaScript implementation of Peter Norvig’s spell corrector was used. The NodeJS package and details about the implementation can be found here: https://www.npmjs.com/package/spell. The library works off a corpus and a domain specific big.txt file needs to be generated. Tika Parser was leveraged for performing a batch parsing on all the crawled HTML files and text data from all these files were extracted. The textual data from these files were merged into a single file big.txt using a Windows batch utility. Finally the export feature of the NPM spell library was used to convert big.txt to big.json, that has the number of occurrences of each unique word in the corpus. The big.json has been submitted as a part of this assignment and resides in the location: solr-nodejs-client\dict\big.json.

Apache Tika CLI batch processing utility tool was used to parse textual content form all HTML files:

```
java -jar tika-app-1.20.jar --text-main -i "D:\usc_courses\Semester-4\Assignments\Assignment4\data\guardiannews" -o "D:\usc_courses\Semester-4\Assignments\Assignment4\data\big "
```

The files were merged using windows type command as below for the Apache Tika’s output location:

```
type *.txt >> big.txt
```

The following commands were used for converting big.txt to big.json using NPM Spell export feature:

```
let data = fs.readFileSync("dict/big.txt", "utf8");
dict.load(data);
fs.writeFileSync("dict/big.json",JSON.stringify(dict.export()), "utf8");
```

Finally, to get the corrected phrase the dictionary is loaded and the lucky method of NPM spell library is invoked:

```
let data = fs.readFileSync("dict/big.json", "utf8");
dict.load(JSON.parse(data));
```

Snippet Generation Feature:

For generating snippets, the data that was ingested into Solr, for the purpose of indexing and searching has been used. The batch processing in Tika extracted the textual content from each of the html files. When a search for a term is made and one of the documents is returned as the search result, the snippet corresponding to it is also sent to the client. The snippet generation algorithm starts by cleansing the textual content extracted by Tika. It first removes multiple line breaks and multiple spaces, then it uses a sentence tokenizer to iterate over every line and extracts sentences from those lines. Finally, it iterates over every line and tries to find if any of the lines matches the query terms. If the line matches it is appended to the result of the snippet. Once the size of the result exceeds 160 characters the algorithm returns the result.

The search is carried out in the following fashion for phrases or multiword queries

1. First all the phrases are searched in each sentence in the exact same order as in which they appear in the query and they should be present adjacent to one another.
2. Then all the sentences are scanned once again to check if all the query terms are present in the sentence in the same order in which they appear in the query but might not be present adjacent to one another. Meaning there could be other words between the query terms. This was achieved using a regular expression.
3. Finally, another round of scan was done to check if one or more query terms in any order is present in the sentences. This was also achieved using a regular expression.

As any point in time if the result size of the snippet exceeded 160 characters, the remaining sentences or rules if any were ignored and the result was returned.

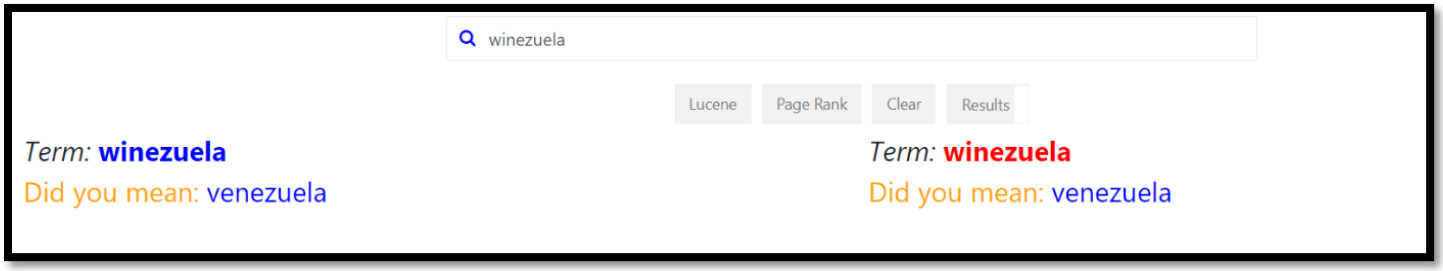
The solr-nodejs-client\lib\snippet.js file can be looked into for details.
A keyword highlighter NPM library has been used for highlighting the matching keywords in the snippet.

Results

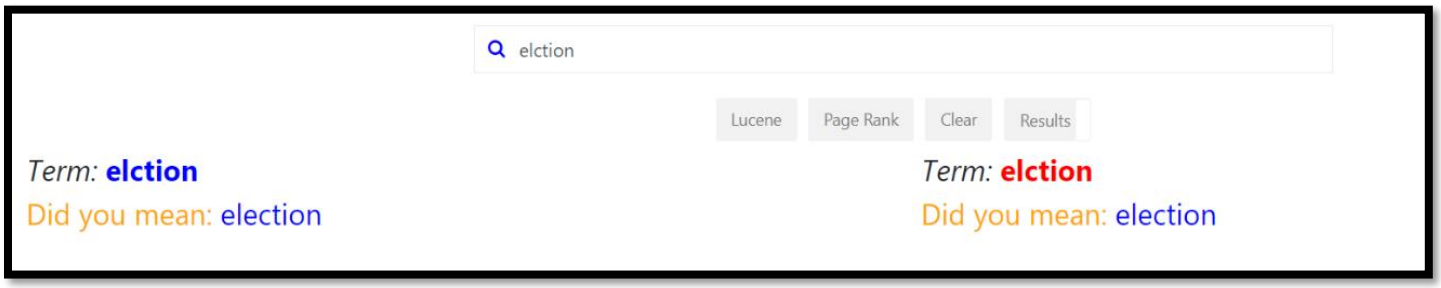
Spell Correction

1. Seenate was corrected to Senate

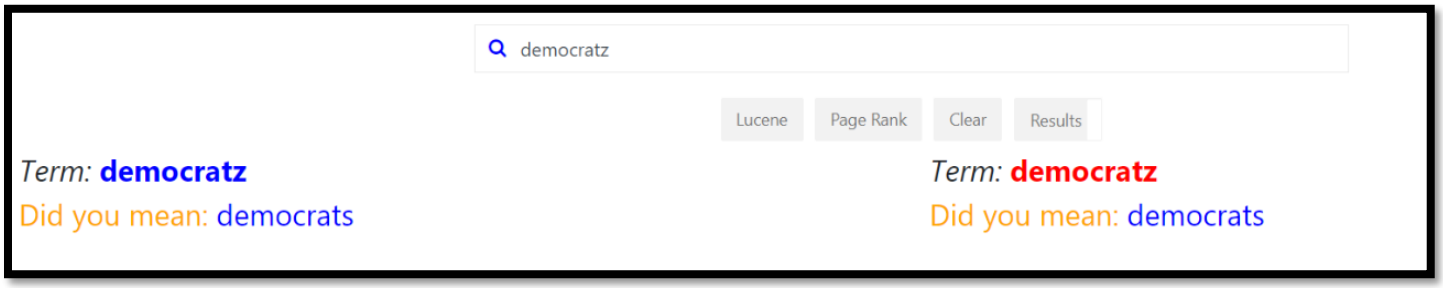
2. Winezuela was corrected to Venezuela



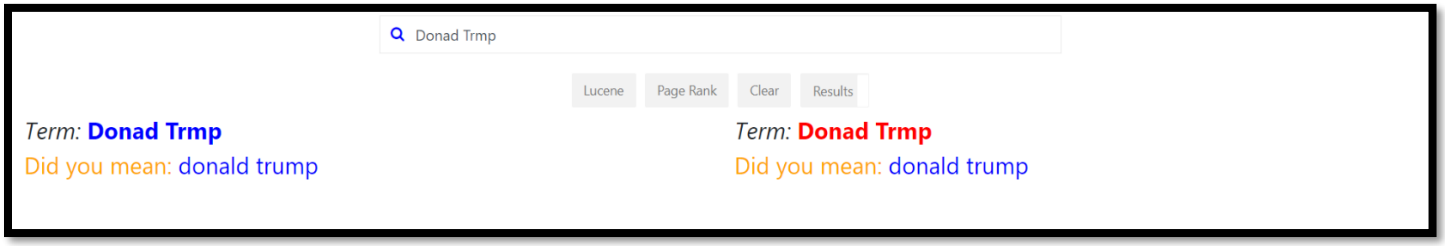
3. Elction was corrected to Election



4. Democratz was corrected to Democrats

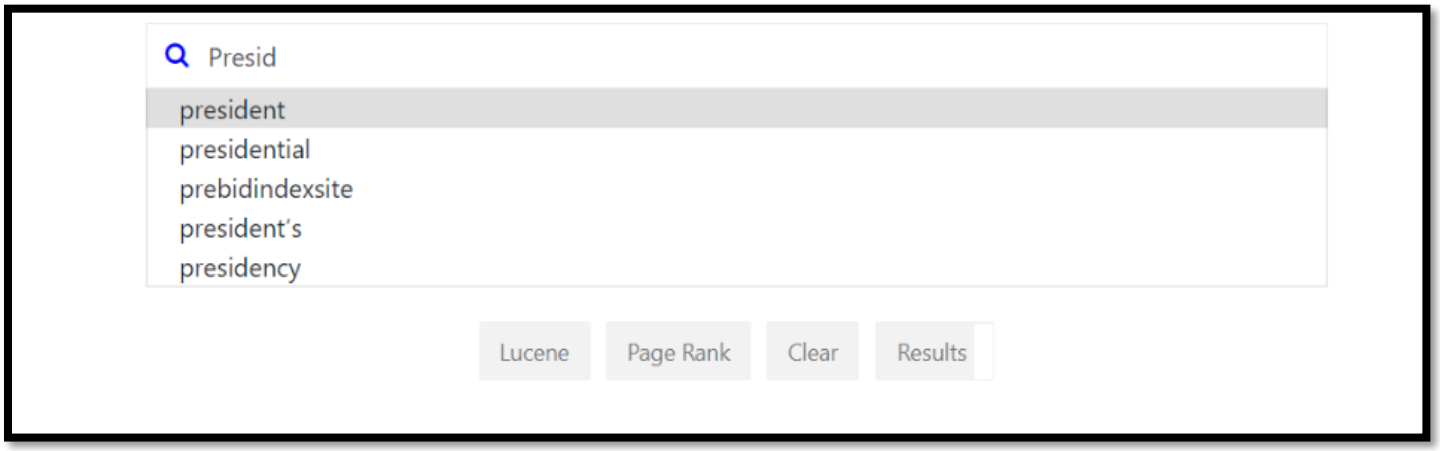


5. Donad Trmp was corrected to Donald Trump

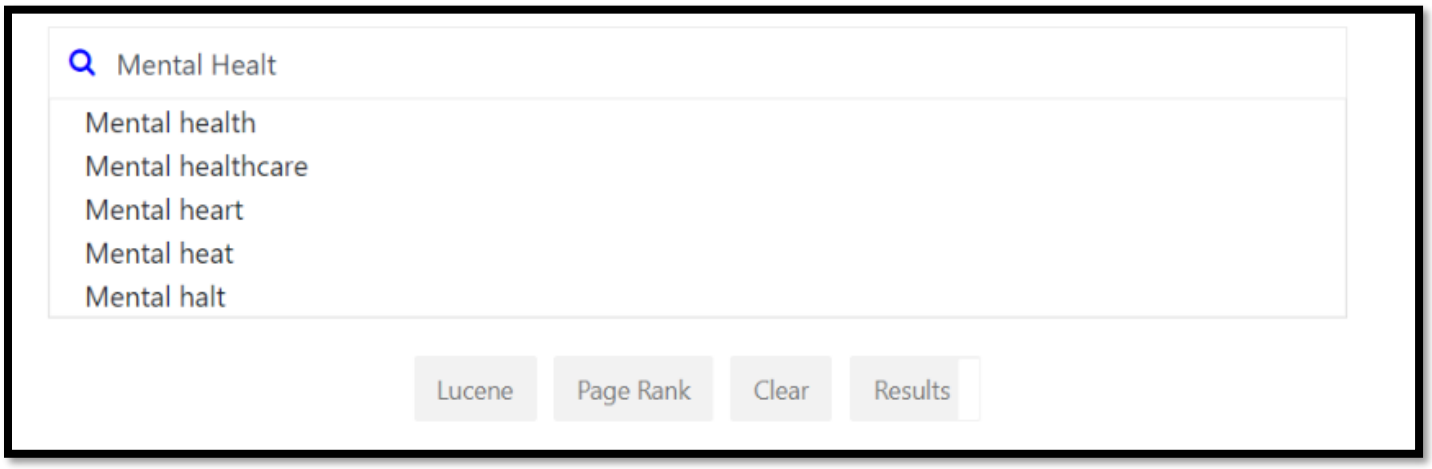


Autocomplete:

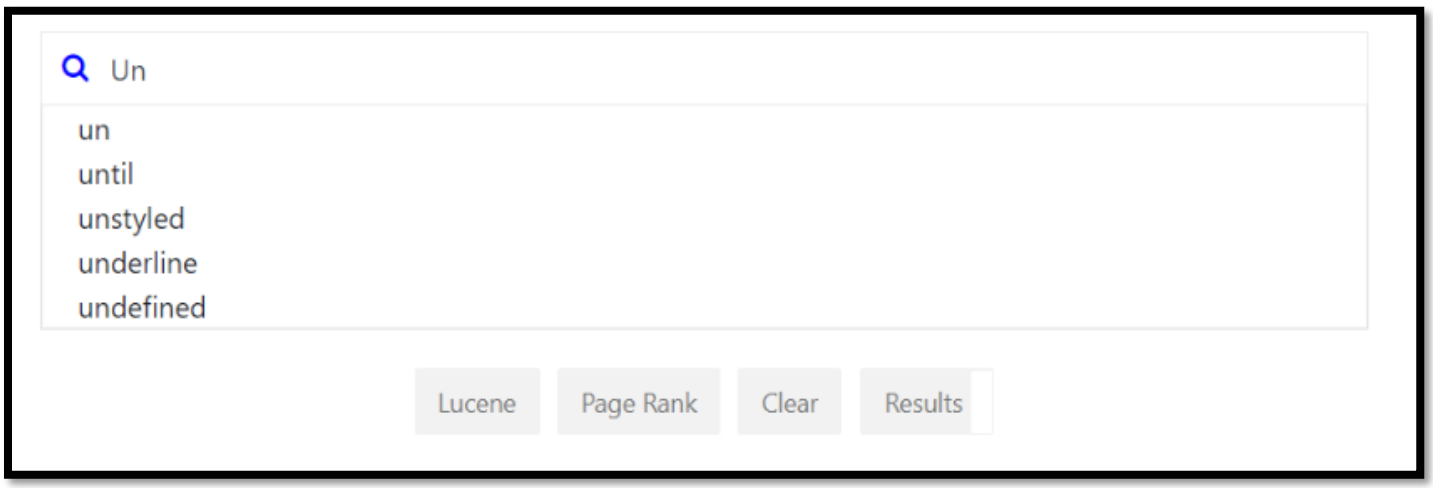
1. Autosuggestions for Presid



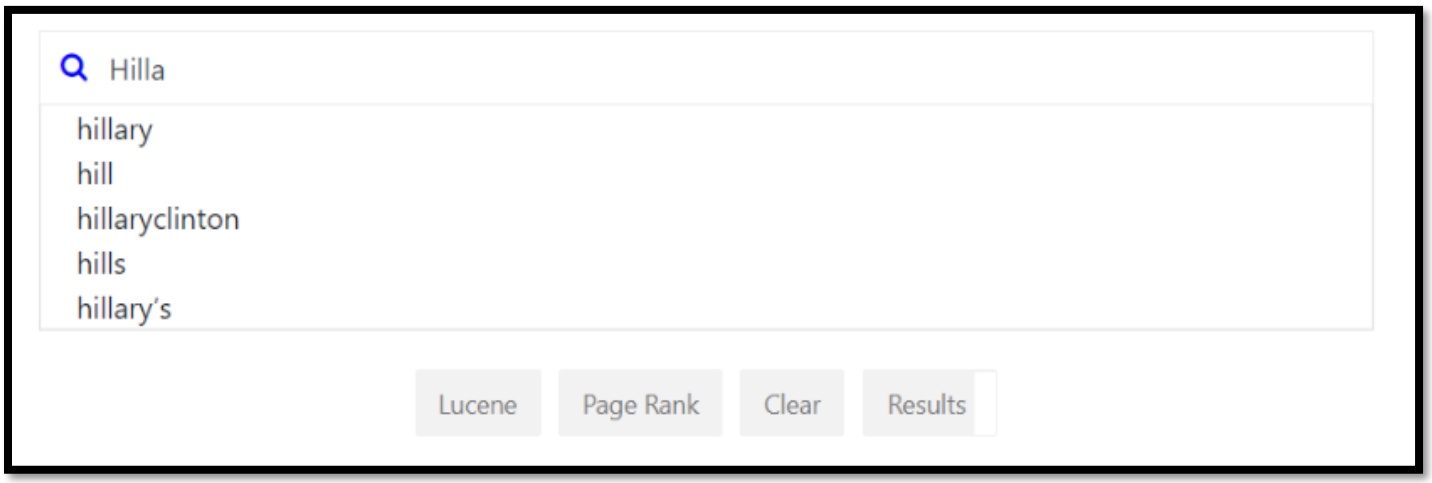
2. Autosuggestions for Mental Healt



3. Autosuggestions for Un



4. Autosuggestion for Hilla



5. Autosuggestion for Oscar 20

P.S.: Next Page contains a complete output with snippets and results fetched.

Overlaps: **3**

Oscar 2019

Lucene | Page Rank | Clear | Results

Term: **Oscar 2019** *Result:* **1 - 10 of 106**

Did you mean: oscar

LUCENE BASED SEARCH RESULTS

1 **Obama commutes sentence for political prisoner Oscar López Rivera | World news | The Guardian**
http://www.theguardian.com/us-news/2017/jan/17/barack-obama-commutes-sentence-oscar-lopez-rivera-puerto-rico-activist
/home/koustav/shared/guardiannews/5e45285d-7078-473d-b3e2-b01ae8b7b2d1.html
Obama commutes sentence for political prisoner **Oscar** López Rivera. A mural dedicated to **Oscar** López Rivera in Puerto Rico.

2 **Richard Trentlage, man behind beloved Oscar Mayer Wiener song, dies at 87 | US news | The Guardian**
http://www.theguardian.com/us-news/2016/oct/01/richard-trentlage-oscar-mayer-wiener-song-dies-hot-dogs
/home/koustav/shared/guardiannews/23f70c63-a005-42f2-8f1b-58294cca2004.html
Richard Trentlage, man behind beloved **Oscar** Mayer Wiener song, dies at 87. © **2019** Guardian News & Media Limited or its affiliated companies.

3 **Movies | The Guardian**
http://www.theguardian.com/us/film
/home/koustav/shared/guardiannews/6b7af0c2-07bc-4245-ae8f-e628f42b2ea1.html
Spike Lee unhappy with Green Book **Oscar** win: 'The ref made a bad call' – video. Spike Lee unhappy with Green Book **Oscar** win: 'The ref made a bad call' – video.

4 **Dick Cheney | Us-news | The Guardian**
http://www.theguardian.com/us-news/dick-cheney
/home/koustav/shared/guardiannews/1b4eb923-8bea-4b07-8897-d8173f51f697.html
Oscar nominations **2019**: Roma and The Favourite deserve acclaim, but no female directors is woeful. February **2019**. Published: 18 Feb **2019**. January **2019**.

5 **'We are going to die from sadness': the fathers and sons reunited behind bars | US news | The Guardian**
http://www.theguardian.com/us-news/2018/oct/14/zero-tolerance-family-separation-indefinite-detention-karnes-texas
/home/koustav/shared/guardiannews/0770c277-105a-47a8-95e3-89829dbd09e9.html
I've been here with **Oscar** [Jr] for 53 days. But he says he got on a plane, so I think they sent him to California," **Oscar** said. **Oscar**, Honduran immigrant.

6 **'Vamos a morir de tristeza': los padres y sus hijos se reunieron tras las rejas | US news | The Guardian**
http://www.theguardian.com/us-news/2018/oct/15/tolerancia-cero-familias-separadas-karnes-texas-detencion-infinita
/home/koustav/shared/guardiannews/dc706bb2-5396-4e20-a5ef-18fed5cf75e7.html
He estado aquí con **Oscar** [Jr] durante 53 días. Pero él dice que se subió a un avión, así que creo que lo enviaron a California," dice **Oscar**. **Oscar** Jr.

7 **California | Us-news | The Guardian**
http://www.theguardian.com/us-news/california
/home/koustav/shared/guardiannews/d7595d11-f749-4896-b90d-fe83b326e910.html
2019. Spike Lee unhappy with Green Book **Oscar** win: 'The ref made a bad call' – video. **Oscars 2019**.

8 **Nate Silver | Us-news | The Guardian**
http://www.theguardian.com/us-news/nate-silver
/home/koustav/shared/guardiannews/a1ee7afd-381e-499e-9cae-cdac761262b3.html
Film blog **Oscar** prognosticators: how did they do at picking the Academy Award winners?. © **2019** Guardian News & Media Limited or its affiliated companies.

9 **Black Lives Matter movement | Us-news | The Guardian**
http://www.theguardian.com/us-news/black-lives-matter-movement
/home/koustav/shared/guardiannews/addfab1f-125f-4cc3-b070-6a8d404f8c43.html
February **2019**. Published: 24 Feb **2019**. Published: 9 Feb **2019**. Published: 6 Feb **2019**. Published: 5 Feb **2019**. January **2019**. Published: 27 Jan **2019**.

10 **US briefing: Oscars shocks, pro-voter bill and Venezuela meeting | US news | The Guardian**
http://www.theguardian.com/us-news/2019/feb/25/us-briefing-oscars-shocks-pro-voter-bill-and-venezuela-meeting
/home/koustav/shared/guardiannews/c37b6a35-a317-48ac-a9df-36258265a41e.html
Mon 25 Feb **2019** 07.08 EST. Olivia Colman celebrates her **Oscar** win. **Oscars 2019**: surprise wins for Green Book and Olivia Colman. The full list of **Oscar** winners.

Term: **Oscar 2019** *Result:* **1 - 10 of 106**

Did you mean: oscar

PAGE RANK BASED SEARCH RESULTS

1 **Movies | The Guardian**
http://www.theguardian.com/us/film
/home/koustav/shared/guardiannews/6b7af0c2-07bc-4245-ae8f-e628f42b2ea1.html
Spike Lee unhappy with Green Book **Oscar** win: 'The ref made a bad call' – video. Spike Lee unhappy with Green Book **Oscar** win: 'The ref made a bad call' – video.

2 **Cory Booker: Ivy League elite or hunger-striking hero? | US news | The Guardian**
http://www.theguardian.com/us-news/2019/mar/13/who-is-cory-booker-democrat-2020
/home/koustav/shared/guardiannews/57218bff-0dd7-433c-b928-9a264d89827f.html
Wed 13 Mar **2019** 06.00 EDT. © **2019** Guardian News & Media Limited or its affiliated companies.

3 **US morning briefing | Us-news | The Guardian**
http://www.theguardian.com/us-news/series/guardian-us-briefing
/home/koustav/shared/guardiannews/3ece6966-6137-477e-baaa-679474a5decf.html
14 March **2019**. 13 March **2019**. 12 March **2019**. 11 March **2019**. 8 March **2019**. 7 March **2019**. 6 March **2019**. 5 March **2019**. 4 March **2019**. 1 March **2019**. 28 February **2019**.

4 **Bernie Sanders' Chicago 2020 speech focuses on fight against racism | US news | The Guardian**
http://www.theguardian.com/us-news/2019/mar/03/bernie-sanders-chicago-speech-brooklyn-college-2020
/home/koustav/shared/guardiannews/25f20fac-2aae-4edb-8617-37798e68f89d.html
Sun 3 Mar **2019** 22.56 EST. © **2019** Guardian News & Media Limited or its affiliated companies.

5 **Black Lives Matter movement | Us-news | The Guardian**
http://www.theguardian.com/us-news/black-lives-matter-movement
/home/koustav/shared/guardiannews/addfab1f-125f-4cc3-b070-6a8d404f8c43.html
February **2019**. Published: 24 Feb **2019**. Published: 9 Feb **2019**. Published: 6 Feb **2019**. Published: 5 Feb **2019**. January **2019**. Published: 27 Jan **2019**.

6 **California | Page 2 of 234 | Us-news | The Guardian**
http://www.theguardian.com/us-news/california
/home/koustav/shared/guardiannews/3a3a4abd-70ef-4dc5-aee2-7847f15e7cca.html
3 March **2019**. 2 March **2019**. 1 March **2019**. 28 February **2019**. 27 February **2019**. 26 February **2019**. 25 February **2019**. 22 February **2019**. 21 February **2019**.

7 **Obama talks empowerment at My Brother's Keeper event: 'You matter' | US news | The Guardian**
http://www.theguardian.com/us-news/2019/feb/20/obama-my-brothers-keeper-oakland-masculinity-empowerment
/home/koustav/shared/guardiannews/2595246d-eb32-44c7-806d-16a8d6e8de2a.html
Wed 20 Feb **2019** 16.14 EST. © **2019** Guardian News & Media Limited or its affiliated companies.

8 **US foreign policy | Page 4 of 551 | Us-news | The Guardian**
http://www.theguardian.com/us-news/us-foreign-policy
/home/koustav/shared/guardiannews/e6002b98-550c-4b53-8336-d965c9513932.html
February **2019**. Published: 6 Feb **2019**. Published: 5 Feb **2019**. Published: 5 Feb **2019**. Published: 4 Feb **2019**. Published: 4 Feb **2019**. Published: 4 Feb **2019**.

9 **Trump faces rebuke from former top officials over 'national emergency' – as it happened | US news | The Guardian**
http://www.theguardian.com/us-news/live/2019/feb/25/trump-washington-live-politics-latest-news-updates
/home/koustav/shared/guardiannews/ec11fb19-40c3-47e6-a55a-08fbd78da305.html
Mon 25 Feb **2019** 20.20 EST. February 25, **2019**. February 25, **2019**. Lee received an honorary **Oscar** in 2015.

10 **Protests in California after police kill black man carrying only his phone | US news | The Guardian**
http://www.theguardian.com/us-news/2018/mar/23/stephon-clark-police-shooting-sacramento-protests-california
/home/koustav/shared/guardiannews/85af151a-412e-4c27-9ac1-38558af3c843.html
© **2019** Guardian News & Media Limited or its affiliated companies.