CSCI572 - HW5 Report
Rahul Ethiraj – USC ID: 3765791028 – ethiraj@usc.edu

News Website: **LA Times**

Steps followed to complete the assignment:

Creation of big.txt:

1. Using Apache Tika, I extracted html content of all webpages and created big.txt, which will be a dictionary of words for spell check and auto suggestion.

Spelling Correction:

1. Peter Norvig's spelling corrector code in PHP was used to find suitable candidates, for the misspelled words in the query search box.
2. It requires bix.txt as input, which is a file containing all the terms in the inverted index of the search engine.
3. I used SpellCorrector.php to check the correctness and generated the serialized_dictionary.txt
4. When a user types in a word in the query search box, and submits the page, the program checks the validity of individual words.
5. If it is a valid word, then the search results are displayed, otherwise the user is given a suggestion based on the dictionary of words using Norvig's spell corrector algorithm.
6. Post which, if the user clicks the corrected suggested word, then the search results are displayed for the new term.

Auto Complete:

1. For auto-complete to work, I made changes to the solrconfig.xml by configuring it to use the SuggestComponent. Post which, I added a requestHandler to configure default parameters for suggestions.
2. I called the auto-complete function in main php file, using the below url to extract suggestions:
   "http://localhost:8983/solr/myexample/suggest?q="correct"&wt=json&indent=true";
3. The results are fetched in the AJAX call and gets displayed in a dropdown list.
4. On selection of the keyword suggestion, the word gets filled in the search box.

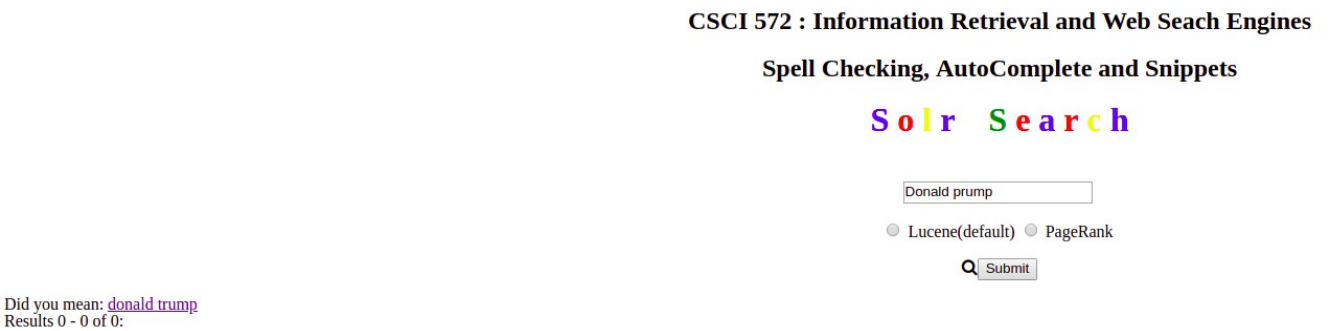
Snippet:

1. I used simple_html_dom.php to parse LA times html pages and extract the content required to show in the snippet section of SERP.
2. File_get_contents can be used to get the contents of the file.
3. I removed any special characters found, so that they do not affect while matching keywords.
4. The content was then split into an array & strings matched to find the index of the words in the query.
5. The string was exploded and found the index of the words in the sentences.
6. Using these indices, snippet string was found and printed accordingly with ellipses(…)
7. In case there are no snippet found, N/A would be displayed, otherwise snippet with boldfaced query words would printed in the snippet.

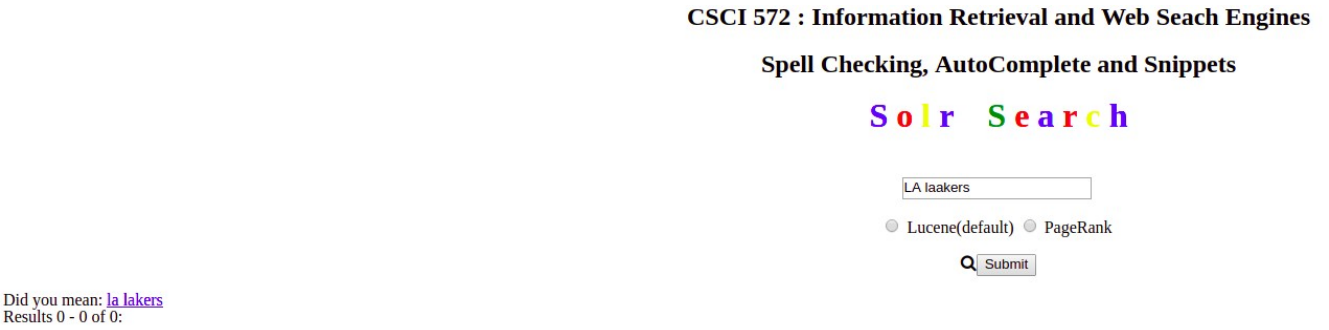
Screenshots of working enhanced Solr search engine:

Spell Correction:

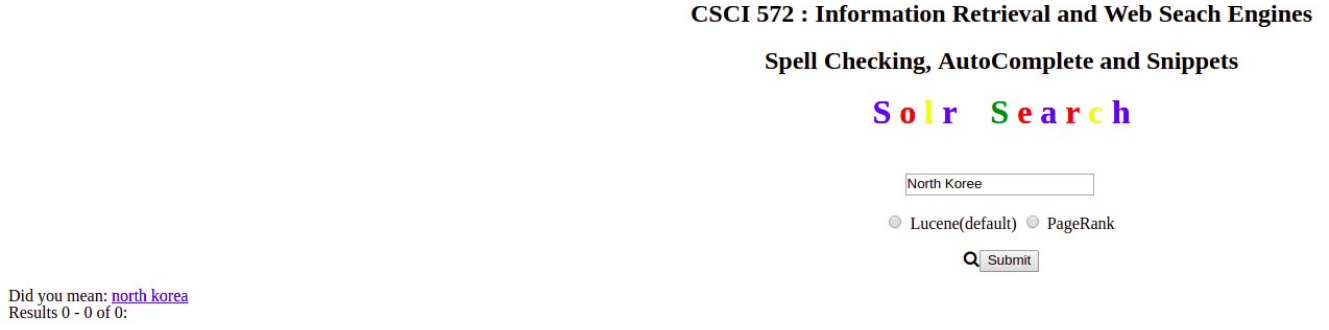
1. Donald trump



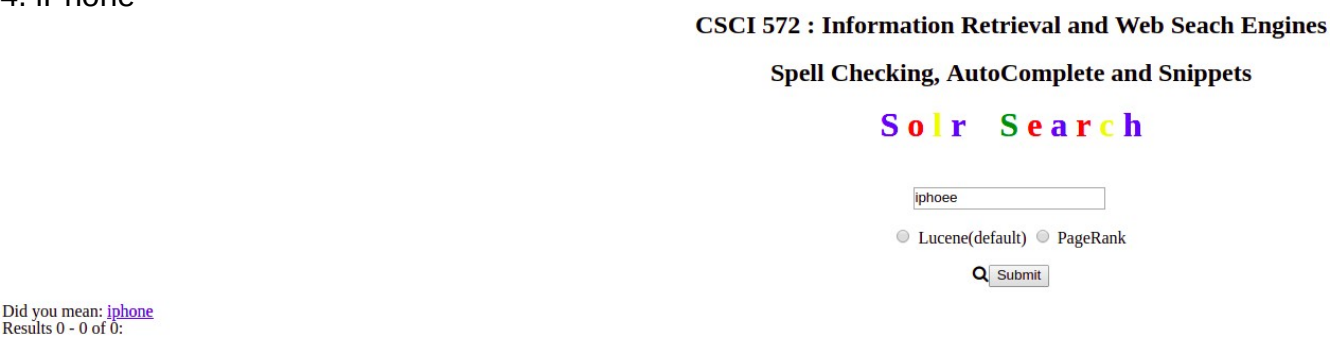
2. LA Lakers



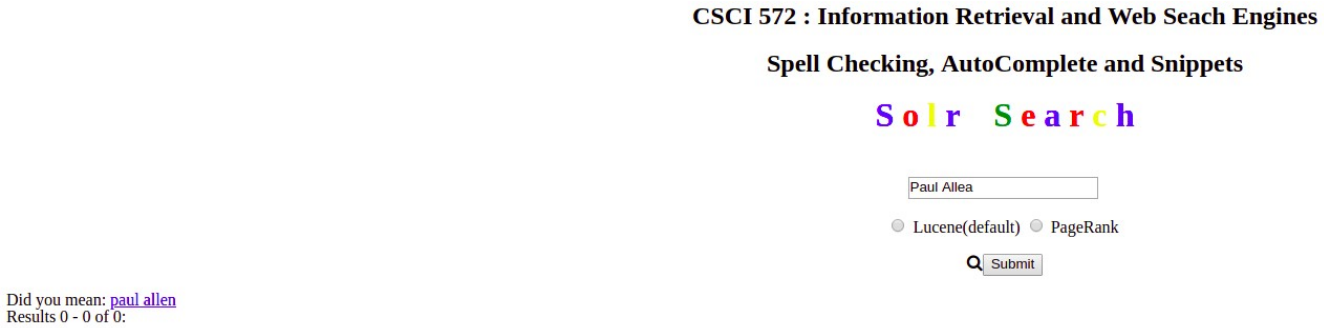
3. North Korea



4. iPhone

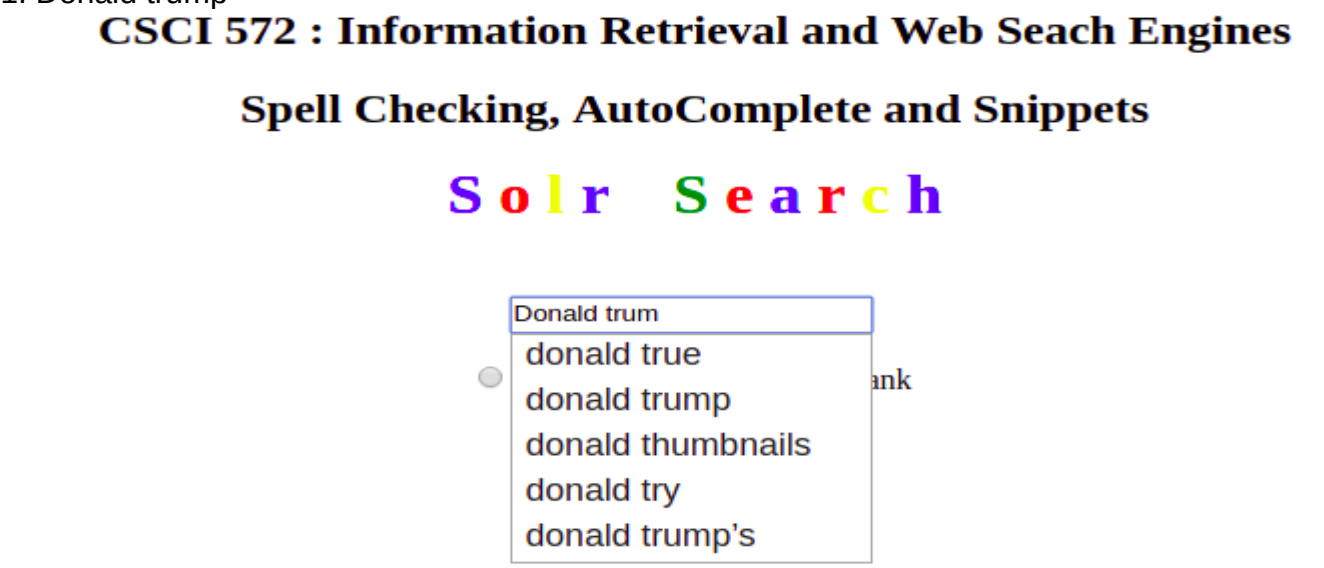


5. Paul Allen

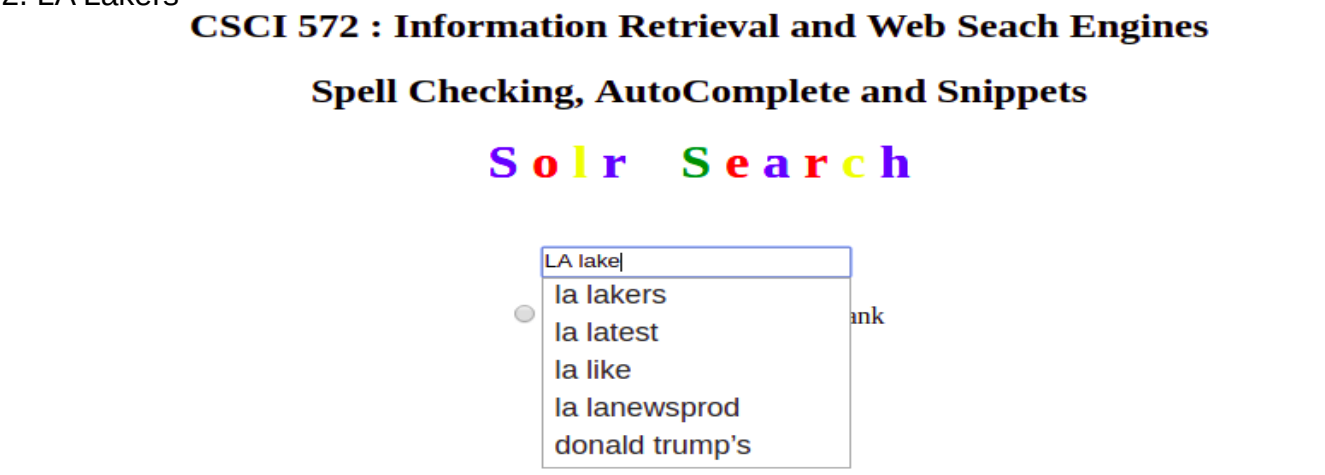


Auto Complete:

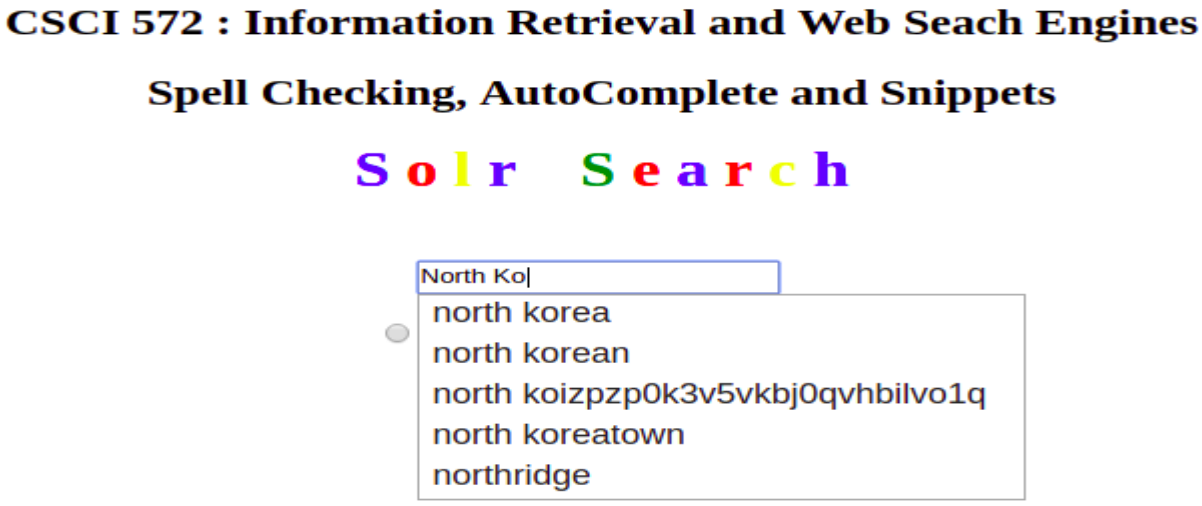
1. Donald trump



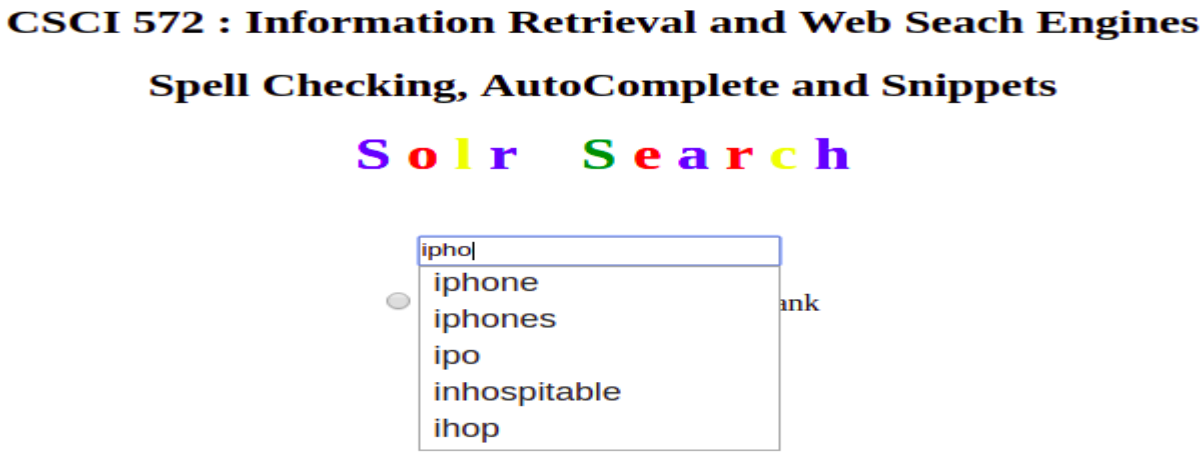
2. LA Lakers



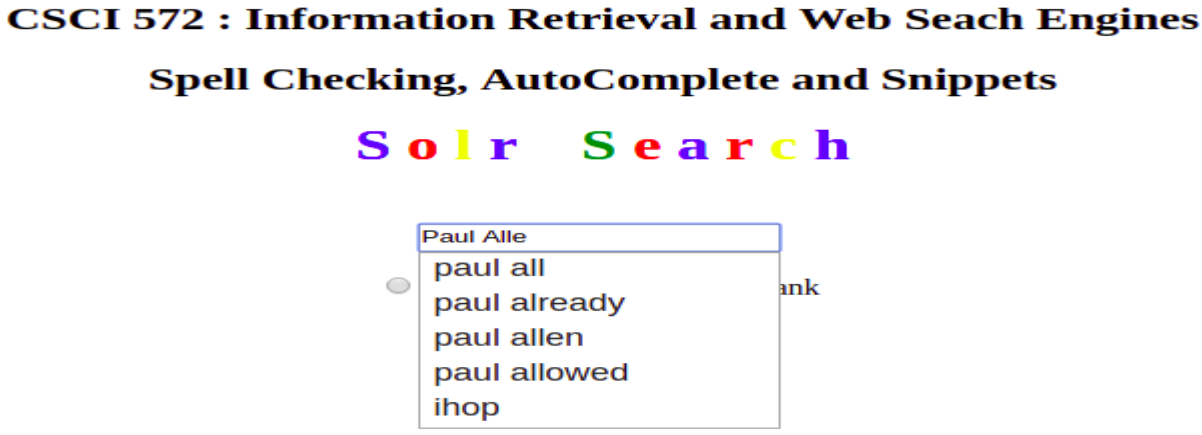
3. North Korea



4. iPhone



5. Paul Allen



Snippet:

CSCI 572 : Information Retrieval and Web Seach Engines

Spell Checking, AutoComplete and Snippets

Solr Search

Donald Trump

Lucene(default) PageRank

Submit

Results 1 - 10 of 2349:

1. **Donald Trump releases new doctor's letter declaring clean bill of health**
   Link http://www.latimes.com/nation/politics/trailguide/la-na-trailguide-updates-trump-release-new-doctor-s-letter-1473951652-htmlstory.html
   ID 292c1d87-3922-48e2-90ea-0b3db47fae4b.html
   Snippet ... The two have made it clear they will work hard to defeat **Donald Trump**, but will continue to hold Clinton accountable from the leftcom%2Fnation%2Fpolitics%2Ftrailguide%2Fla-na-trailguide-updates-now-we-know-how-much-**Donald-Trump**-1473948715-htmlstory...

2. **Trump offers people in the country illegally a way to stay: Join the military**
   Link http://www.latimes.com/nation/politics/trailguide/la-na-trailguide-updates-donald-trump-open-to-allowing-those-in-1473296491-htmlstory.html
   ID 73df42d1-bf4e-49d1-9e0e-140d48efb0ad.html
   Snippet ...com%2Fnation%2Fpolitics%2Ftrailguide%2Fla-na-trailguide-updates-**Donald-Trump**-open-to-allowing-those-in-1473296491-htmlstorycom%2Fnation%2Fpolitics%2Ftrailguide%2Fla-na-trailguide-updates-**Donald-Trump**-open-to-allowing-those-in-1473296491-htmlstory...

3. **Why Donald Trump keeps popping up in local races he has nothing to do with - Los Angeles Times**
   Link http://www.latimes.com/politics/la-pol-ca-donald-trump-california-races-downticket-20160527-snap-htmlstory.html
   ID a3d9c7e6-d00d-4d23-b449-23c8271be423.html
   Snippet ...com%2Fpolitics%2Fla-pol-ca-**Donald-Trump**-california-races-downticket-20160527-snap-htmlstorycom%2Fpolitics%2Fla-pol-ca-**Donald-Trump**-california-races-downticket-20160527-snap-htmlstory...

4. **Following Trump's money exposes the awful truth: Our president is a 'financial vampire'**
   Link http://www.latimes.com/opinion/op-ed/la-oe-johnston-trump-cons-and-cheats-20181004-story.html
   ID 78ff4abb-877a-4c13-a5e2-33fa09d5ef8e.html
   Snippet ...As that paper’s former tax reporter and a journalist who has covered **Donald Trump** for more than 30 years, this was no surprise **Donald Trump** is already a proven tax cheat...

5. **The moment when the Donald Trump and Kim Jong Un impersonators were escorted back to their seats**
   Link http://www.latimes.com/sports/olympics/la-sp-olympics-live-updates-the-moment-when-the-donald-trump-and-kim-1518187562-htmlstory.html
   ID 1d0a2e04-2478-44d9-9234-b959a0a25746.html
   Snippet ... President **Donald Trump** and North Korean leader Kim Jong Un came down the steps leading to the area where the media was sitting and once at the first row President **Donald Trump** and North Korean leader Kim Jong Un came down the steps leading to the area where the media was sitting and once at the first row...

6. **Whom do you believe, Michael Cohen or Donald Trump? Yes, that's a rhetorical question**
   Link http://www.latimes.com/opinion/la-ol-enter-the-fray-whom-to-believe-michael-cohen-or-donald-1532702692-htmlstory.html
   ID e78af783-64ca-4d2d-8004-3aa5c6f5debd.html
   Snippet ... They definitely don’t want **Trump**!— **Donald** J...

7. **Dos funerales y una boda: el rechazo a Donald Trump**
   Link http://www.latimes.com/espanol/politica/la-es-dos-funerales-y-una-boda-el-rechazo-a-donald-trump-20180828-story.html
   ID ed6c17fa-2399-4f5d-8d69-37315a2fde3d.html
   Snippet ...com%2Fespanol%2Fpolitica%2Fla-es-dos-funerales-y-una-boda-el-rechazo-a-**Donald-Trump**-20180828-storycom%2Fespanol%2Fpolitica%2Fla-es-dos-funerales-y-una-boda-el-rechazo-a-**Donald-Trump**-20180828-story...

8. **Audio reveals Donald Trump making lewd comments about women**
   Link http://www.latimes.com/nation/politics/trailguide/la-na-live-updates-trailguide-1475872277-htmlstory.html
   ID de5298bb-0d80-43d5-be92-dfb62c18d84a.html
   Snippet ...com%2Fnation%2Fpolitics%2Ftrailguide%2Fla-na-live-updates-trailguide-paul-ryan-wants-**Donald-Trump**-to-emulate-1475866340-htmlstorycom%2Fnation%2Fpolitics%2Ftrailguide%2Fla-na-live-updates-trailguide-paul-ryan-wants-**Donald-Trump**-to-emulate-1475866340-htmlstory...

9. **Donald Trump might actually have to shoot someone on Fifth Avenue before GOP leaders say, 'Enough'**
   Link http://www.latimes.com/opinion/la-ol-enter-the-fray-donald-trump-might-actually-have-to-1535040294-htmlstory.html
   ID e1308a8f-ab4b-4f98-b0c0-8a2e3087303f.html
   Snippet ...com%2Fopinion%2Fla-ol-enter-the-fray-**Donald-Trump**-might-actually-have-to-1535040294-htmlstorycom%2Fopinion%2Fla-ol-enter-the-fray-**Donald-Trump**-might-actually-have-to-1535040294-htmlstory...

10. **On Sunday's '60 Minutes,' meet President-elect Donald Trump - Los Angeles Times**
    Link http://www.latimes.com/entertainment/tv/la-et-st-donald-trump-60-minutes-20161110-story.html
    ID 26eea782-433e-43f1-8866-4a79a4961a41.html
    Snippet ...Additionally, Stahl will be the first to interview **Trump**'s closest family members, including First Lady-elect Melania, and his children Ivanka, Tiffany, Eric and **Donald** Jr...