Reducing Noise

CS5604: Final Presentation

Xiangwen Wang, Prashant Chandrasekar
Acknowledgements

- Integrated Digital Event Archiving and Library (IDEAL)
- Digital Libraries Research Laboratory (DLRL)
- Dr. Fox, IDEAL GRA’s (Sunshin & Mohamed)
- All of the teams especially the Hadoop & Solr team
Goal

- Providing “quality” data
  - Identify and remove “noisy” data
  - Process and clean “sound” data
  - Extract and organize data
Reducing Noise in IR System

NER  LDA  Social  Classification  Clustering

Raw Source Web pages (in Sequence File Format) → Webpage and Tweets Cleanup → Clean Web Pages (in AVRO File Format) → HBase
Cleanup Process

Raw Tweets Collections in Avro Format on HDFS

Lists of URLs

Nutch

Raw Files & Metadata in SequenceFile Format

Clean Tweets in Avro Format

Clean web pages in Avro Format
Libraries/Packages

- BeautifulSoup4:
  - Parse text out of HTML and XML files.
- Readability:
  - Pull out the title and main body text from a webpage.
- Langdetect:
  - Detect language of a text using naive Bayesian filter.
- Re:
  - Provide regular expression matching operations.
- NLTK:
  - Natural language processing (stemming, stopwords removal).
Cleaning Rules

- Remove Emoticons
- Remove Non-ASCII Characters
- Filter english text only
- Extract URLs, (Tweets) Hashtags, (Tweets) User Handles
- Stopwords removal, Stemming
- Remove invalid URLs
- Remove profane words

Implemented with regular expressions, e.g.

```
HashtagRegexp = r'(\<=^|(?<=[^a-zA-Z0-9\.\-\_]#)([A-Za-z_]+[A-Za-z0-9_]+)')
UserhandleRegexp = r'(\<=^|(?<=[^a-zA-Z0-9\.\-\_]@([A-Za-z_]+[A-Za-z0-9_]+)')
UrlRegexp = r'(\?P[url]https?://[a-zA-Z0-9\./-]+)'
```
Tweets cleaning process

Tweets collections in AVRO Format

ASCII characters only

English only Tweets

Matching regular expressions
Stop words and profane words removal

URLs, Hashtags, User Handles, etc.
Clean Tweets
Stemming
Clean Tweets 2

Clean Tweets in AVRO format with pre-setted schema
News: Ebola in Canada?: @CharlieHebdo #CharlieHebdo #Shooting Suspected patient returns from Nigeria to Ontario with symptoms http://t.co/KoOD8yweJd
Web pages cleaning process

- Raw Files & metadata in Sequence File format
  - Titles & “Useful” HTML
    - HTML only (raw)
      - Identified individual files with corresponding URLs
        - Raw consolidated files & metadata in Text format
          - Clean Text and Extract URLs
            - Clean web pages in AVRO format with pre-setted schema

- Readability
  - Beautifulsoup
  - Langdetect

- Content only
  - English only Text
    - ASCII only Text
      - Clean Text and Extract URLs
Eight in 10 say their families have seen either zero or not very much improvement in their living standards, according to pollsters Ipsos MORI.

Looking to the future, less than a quarter think they will be much better off in the next 12 months spanning the general election in May.

The grim findings come a day after the Bank of England Governor trumpeted the return of “real pay growth”, as official figures showed wages creeping up ahead of the cost of living.

Labour leader Ed Miliband, making a keynote comeback speech in London, said most families were simply missing out altogether on a recovery that favoured the better-off. For Tory MPs, the findings will fuel fears that the past five years of painful austerity are leading to a “voteless” recovery as households face up to repaying debts rather than enjoying spending sprees.

Mr Miliband told an audience in west London that the recovery was only working for “the privileged few”.

Other people, he warned, were “asking, why are they being told there is a recovery when they aren’t feeling the benefits. People working so hard but not being rewarded, young people fearing that they are going to have a worse life than their parents, people making a decent living but still unable to afford to buy a house.”
eight 10 say families seen either zero much improvement living standards, according pollsters ipsos mori. looking future, less quarter think much better next 12 months spanning general election may. grim findings come day bank england governor trumpeted return real pay growth, official figures showed wages creeping ahead cost living. labour leader ed miliband, making keynote comeback speech london, said families simply missing altogether recovery favoured better-off. tory mps, findings fuel fears past five years painful austerity leading voteless recovery households face repaying debts rather enjoying spending sprees. mr miliband told audience west london recovery working privileged. people, warned, asking, told recovery aren feeling benefits. people working hard rewarded, young people fearing going worse life parents, people making decent living still unable afford buy house. ipsos mori found overall optimism economy lower summer. 42 per cent think things improve next year, 23 per cent think get worse. net figure plus 19 well plus 32 recorded august. asked personal finances, people pessimistic. 25 per cent expect get better, 22 per cent fear decline. half think situation change. comes family living standards, almost half say improved. three 10 seen bit improvement. 14 per cent feel fair amount better off, three per cent feel great deal better off. asked year ahead, people slightly less pessimistic. three quarters predict little better off. fifth think great fair amount better off. bobby duffy, head ipsos mori social research institute, said: results show britons economic optimism still converting feelgood factor personal level. conservatives vote share shown little change since end recession. whether convert recovery votes, labour take advantage ongoing concerns living standards, remain one key issues right election.
<table>
<thead>
<tr>
<th>Collection Name</th>
<th>Number of Tweets</th>
<th>Number of Non-English Tweets</th>
<th>Percent of Tweets cleaned</th>
<th>Execution Time (seconds)</th>
<th>Size of Output File (MB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>suicide_bomb_attack_S</td>
<td>39258</td>
<td>2083</td>
<td>95% (37175)</td>
<td>32.18</td>
<td>15.8</td>
</tr>
<tr>
<td>Jan_25</td>
<td>911684</td>
<td>415873</td>
<td>55% (495811)</td>
<td>475.18</td>
<td>210.2</td>
</tr>
<tr>
<td>charlie_hebdo_S</td>
<td>520211</td>
<td>346989</td>
<td>34% (173222)</td>
<td>85.46</td>
<td>70.1</td>
</tr>
<tr>
<td>ebola_S</td>
<td>621099</td>
<td>240655</td>
<td>62% (380444)</td>
<td>422.66</td>
<td>136.3</td>
</tr>
<tr>
<td>election_S</td>
<td>931436</td>
<td>101659</td>
<td>90% (829777)</td>
<td>823.77</td>
<td>323.5</td>
</tr>
<tr>
<td>plane_crash_S</td>
<td>273595</td>
<td>7561</td>
<td>98% (266034)</td>
<td>237.96</td>
<td>100.0</td>
</tr>
<tr>
<td>winter_storm_S</td>
<td>493812</td>
<td>7772</td>
<td>99% (486040)</td>
<td>444.90</td>
<td>192.9</td>
</tr>
<tr>
<td>egypt_B</td>
<td>11747983</td>
<td>3797211</td>
<td>68% (7950772)</td>
<td>7271.36</td>
<td>3313.2</td>
</tr>
<tr>
<td>Malaysia_Airlines_B</td>
<td>1239606</td>
<td>462651</td>
<td>63% (776955)</td>
<td>769.62</td>
<td>305.3</td>
</tr>
<tr>
<td>bomb_B</td>
<td>24676569</td>
<td>3955957</td>
<td>84% (20720612)</td>
<td>16506.19</td>
<td>6537.7</td>
</tr>
<tr>
<td>diabetes_B</td>
<td>8382585</td>
<td>2845395</td>
<td>67% (5537190)</td>
<td>5452.62</td>
<td>2073.2</td>
</tr>
<tr>
<td>shooting_B</td>
<td>26381867</td>
<td>3535187</td>
<td>87% (22846680)</td>
<td>19377.55</td>
<td>8403.2</td>
</tr>
<tr>
<td>storm_B</td>
<td>27286337</td>
<td>4949268</td>
<td>82% (22337069)</td>
<td>18649.99</td>
<td>7591.7</td>
</tr>
<tr>
<td>tunisia_B</td>
<td>6061926</td>
<td>3322288</td>
<td>46% (2739638)</td>
<td>3265.69</td>
<td>1004.4</td>
</tr>
</tbody>
</table>
## Output Statistics: Web pages

<table>
<thead>
<tr>
<th>Collection Name</th>
<th>Number of web pages</th>
<th>Percent of web pages cleaned</th>
<th>Execution Time (seconds)</th>
<th>Size of Output File (MB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ebola_S</td>
<td>1279</td>
<td>36%</td>
<td>64.51</td>
<td>28.58</td>
</tr>
<tr>
<td>election_S</td>
<td>266</td>
<td>72%</td>
<td>16.15</td>
<td>11.44</td>
</tr>
<tr>
<td>charlie_hebdo_S</td>
<td>339</td>
<td>87%</td>
<td>21.54</td>
<td>17.02</td>
</tr>
<tr>
<td>winter_storm_S</td>
<td>1726</td>
<td>66%</td>
<td>79.93</td>
<td>65.06</td>
</tr>
<tr>
<td>egypt_B</td>
<td>6,844</td>
<td>45%</td>
<td>330.12</td>
<td>143.42</td>
</tr>
<tr>
<td>shooting_B</td>
<td>10843</td>
<td>81%</td>
<td>524.49</td>
<td>462.11</td>
</tr>
</tbody>
</table>
Challenges

- Non-ASCII characters
- Language of document
- AVRO with Hadoop Streaming
Next Steps and Future Work

- **Next Steps**
  - Clean big collection
  - Reducing Noise using MR

- **Future Work**
  - Cleaning Document with multiple languages
  - Cleaning different document formats
References

- NLTK project, NLTK 3.0 documentation. [http://www.nltk.org](http://www.nltk.org), accessed on 02/05/2015.
- Michal Danilák, langdetect, [https://github.com/Mimino666/langdetect](https://github.com/Mimino666/langdetect), accessed on 04/30/2015.
- Python Software Foundation, Regular expression operations, [https://docs.python.org/2/library/re.html](https://docs.python.org/2/library/re.html), accessed on 04/30/2015.
Thank you

{wxw, peecee}@vt.edu