Reducing Noise

CS5604: Final Presentation

Xiangwen Wang, Prashant Chandrasekar

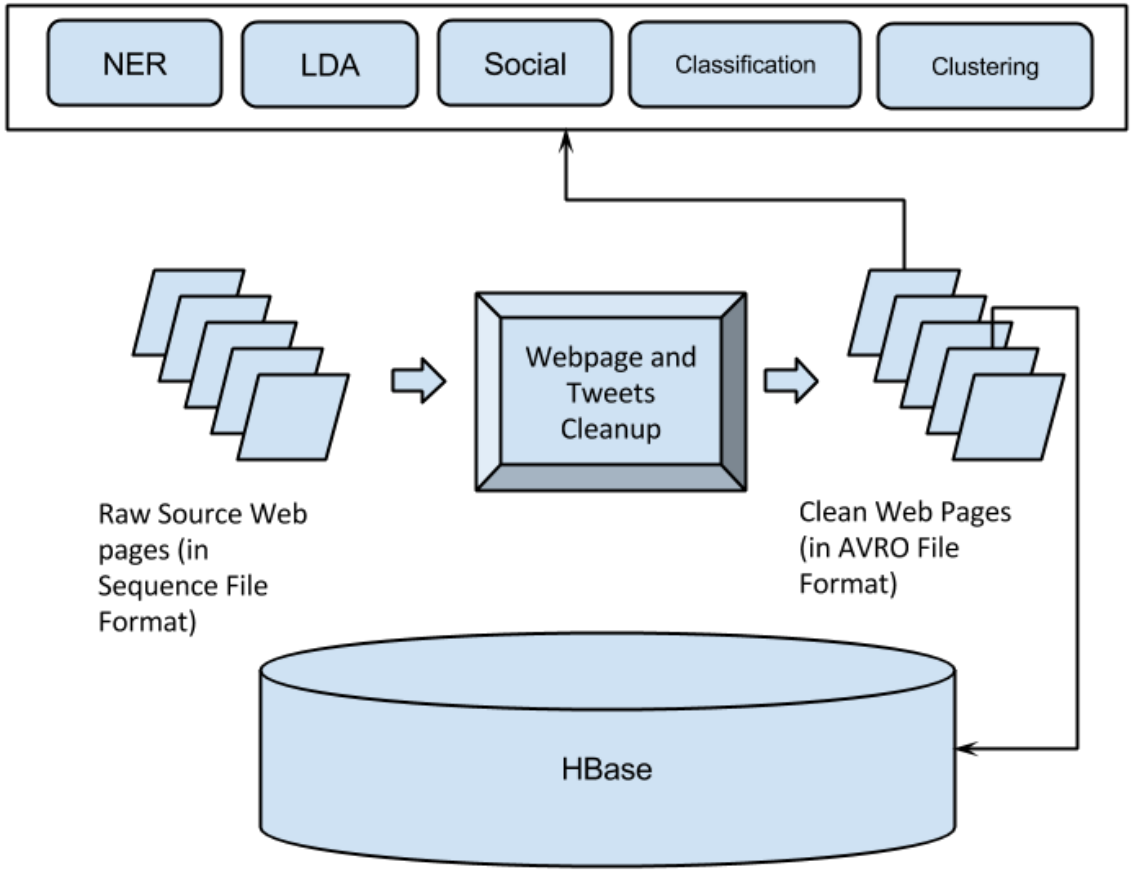
Acknowledgements

- 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

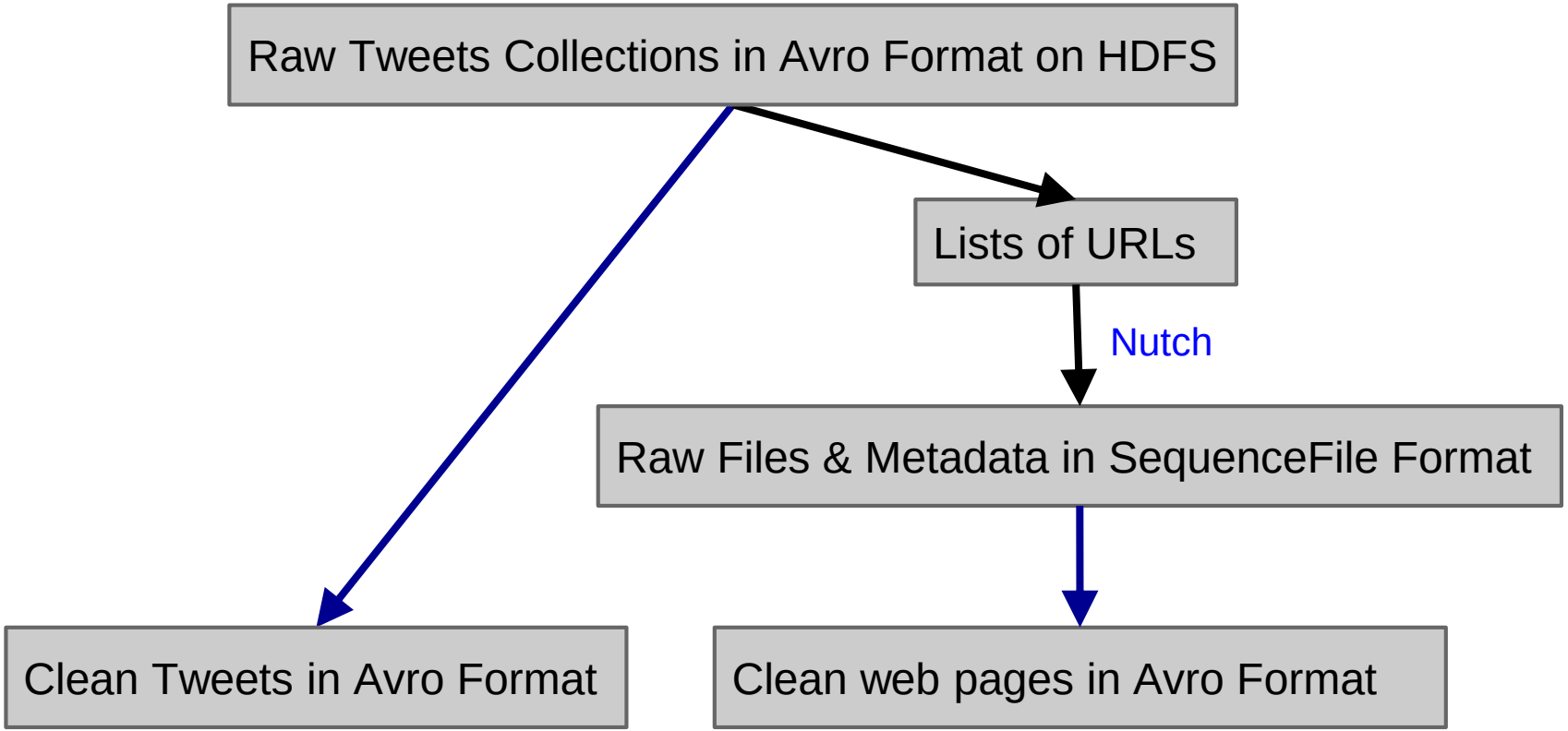
Goal

- Providing “quality” data
  - Identify and remove “noisy” data
  - Process and clean “sound” data
  - Extract and organize data

Reducing Noise in IR System



Cleanup Process



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).

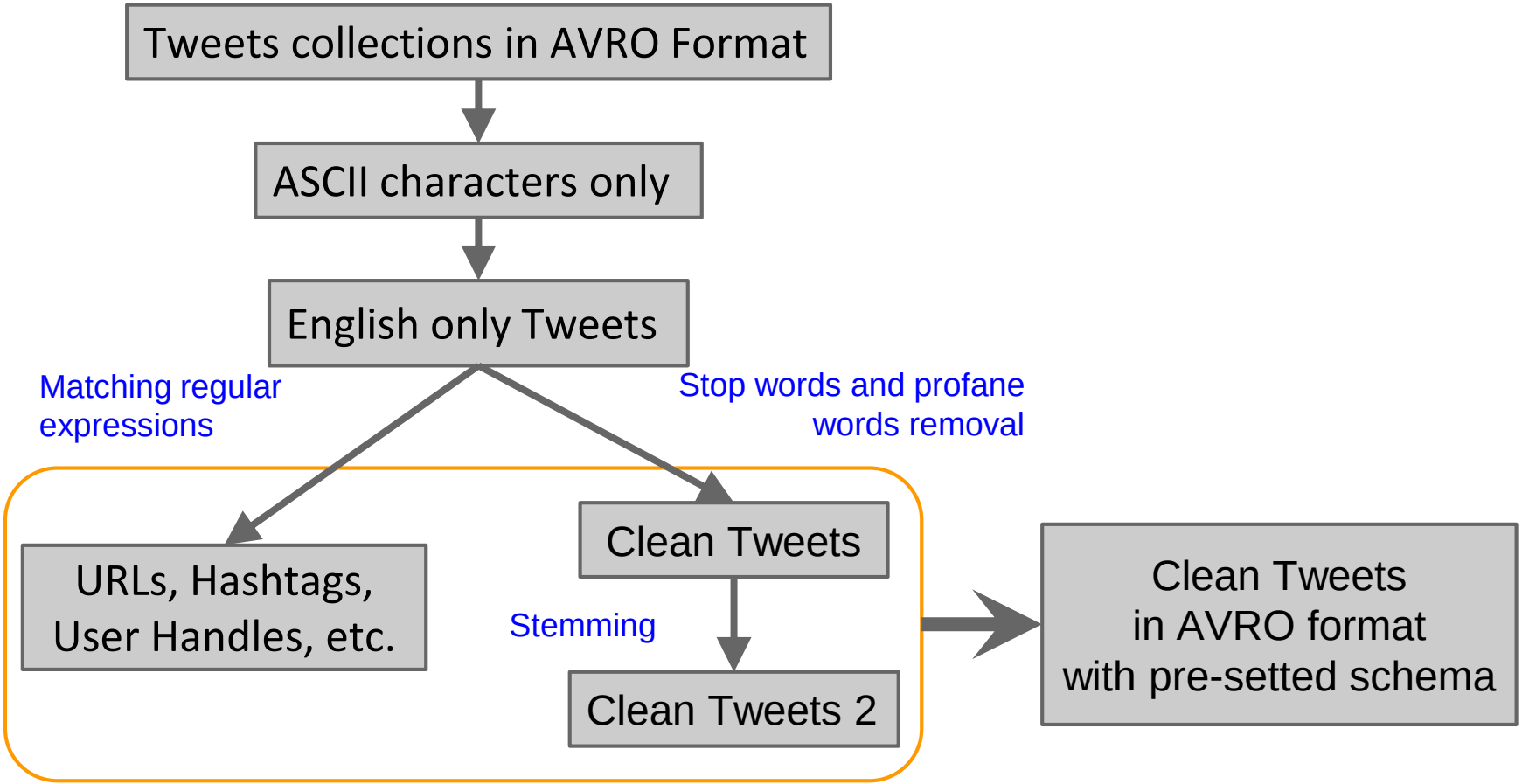
Cleaning Rules

- 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



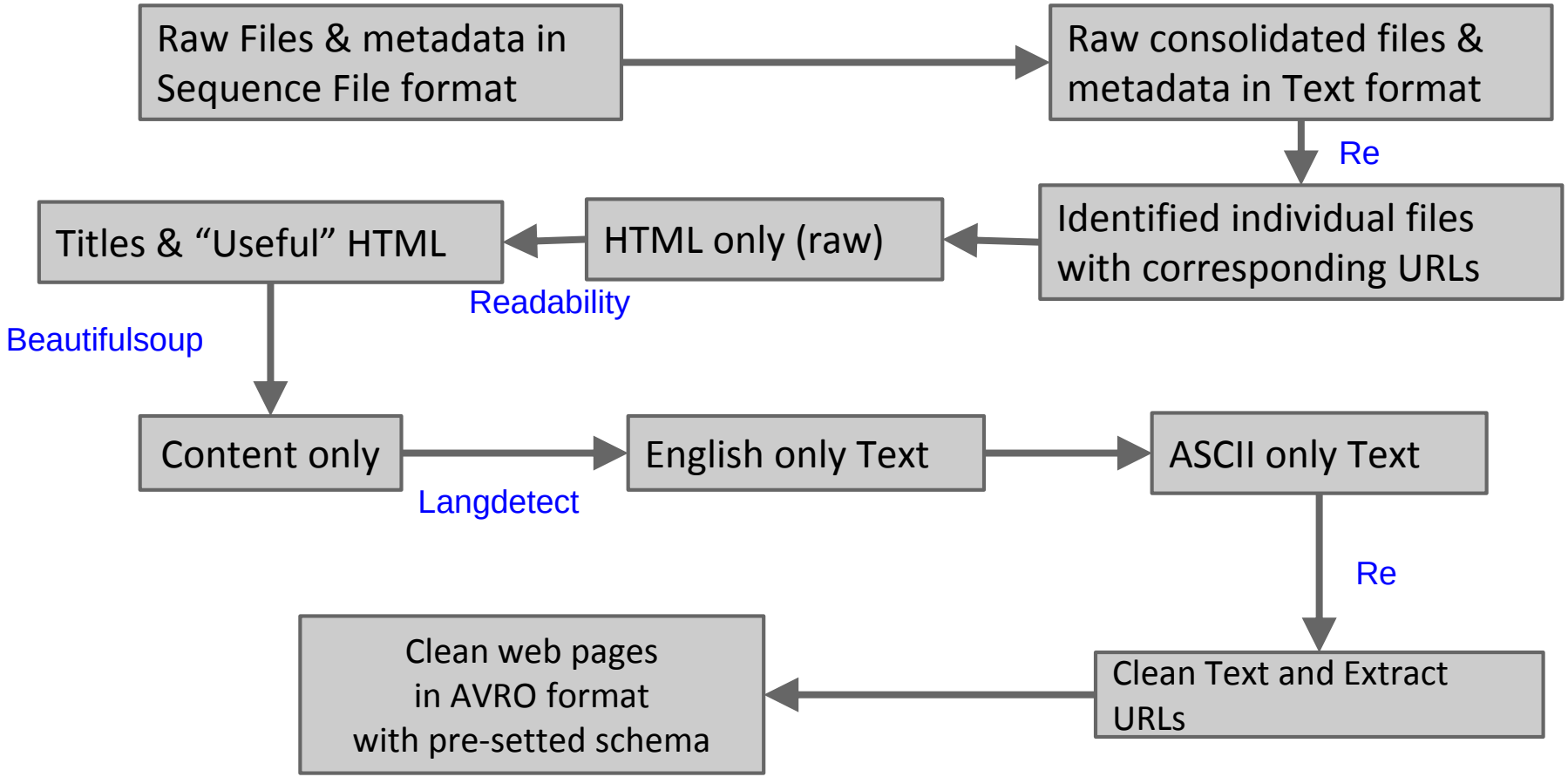
```
{
    'iso_language_code': 'en',
    'text': u"News: Ebola in Canada?: @CharlieHebdo
#CharlieHebdo #Shooting Suspected patient returns
from Nigeria to Ontario with symptoms
http://t.co/KoOD8yweJd",
    'time': 1407706831,
    'from_user': 'toyeenb',
    'from_user_id': '81703236',
    'to_user_id': '',
    'id': '498584703352320001'
    . . .
}
```


Original Tweet Example


```
{
    'lang': 'English',
    'user_id': '81703236',
    'text_clean': 'News Ebola in Canada Suspected
patient returns from Nigeria to Ontario with symptoms',
    'text_clean2': 'new ebol canad suspect paty return
niger ontario symptom ',
    'created_at': '1407706831',
    'hashtags': 'CharlieHebdo|Shooting',
    'user_mentions_id': 'CharlieHebdo',
    'tweet_id': '498584703352320001',
    'urls': 'http://t.co/KoOD8yweJd',
    'collection': 'charlie_hebdo_S',
    'doc_id': 'charlie_hebdo_S--8515a3c7-1d97-3bfa-
a264-93ddb159e58e',
    'user_screen_name': 'toyeenb',
    . . .
}
```


Clean Tweet Example

Web pages cleaning process



Web Page Cleaning: Original Page

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."

David Cameron talks to the Standard with eight days to go until the General Election

Will battle after north London man leaves £500k to builder who cleaned his gutters for free

Web Page Cleaning: Clean Text

http://www.standard.co.uk/news/politics/david-camerons-hopes-of-feelgood-election-boost-are-dashed-as-80-per-cent-say-recovery-is-not-helping-9858273.html

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.

Output Statistics: Tweets

Collection Name	Number of Tweets	Number of Non-English Tweets	Percent of Tweets cleaned	Execution Time (seconds)	Size of Output File (MB)
suicide_bomb_attack_S	39258	2083	95% (37175)	32.18	15.8
Jan_25	911684	415873	55% (495811)	475.18	210.2
charlie_hebdo_S	520211	346989	34% (173222)	85.46	70.1
ebola_S	621099	240655	62% (380444)	422.66	136.3
election_S	931436	101659	90% (829777)	823.77	323.5
plane_crash_S	273595	7561	98% (266034)	237.96	100.0
winter_storm_S	493812	7772	99% (486040)	444.90	192.9
egypt_B	11747983	3797211	68% (7950772)	7271.36	3313.2
Malaysia_Airlines_B	1239606	462651	63% (776955)	769.62	305.3
bomb_B	24676569	3955957	84% (20720612)	16506.19	6537.7
diabetes_B	8382585	2845395	67% (5537190)	5452.62	2073.2
shooting_B	26381867	3535187	87% (22846680)	19377.55	8403.2
storm_B	27286337	4949268	82% (22337069)	18649.99	7591.7
tunisia_B	6061926	3322288	46% (2739638)	3265.69	1004.4

Output Statistics: Web pages

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

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

- Steven Bird, Ewan Klein, and Edward Loper, *Natural language processing with Python.* O'Reilly Media, 2009.
- NLTK project, NLTK 3.0 documentation. http://www.nltk.org, accessed on 02/05/2015.
- The Apache Software Foundation, Solr Download. http://lucene.apache.org/solr/mirrors-solr-latest-redir.html, accessed on 02/05/2015.
- Leonard Richardson, Beautiful Soup Documentation. http://www.crummy.com/software/BeautifulSoup/bs4/doc/, accessed on 02/05/2015.
- Christopher Manning, Prabhakar Raghavan, and Hinrich Schütze, *Introduction to information retrieval*. Vol. 1. Cambridge: Cambridge University Press, 2008.
- Alex J. Champandard, The Easy Way to Extract Useful Text from Arbitrary HTML. http://ai-depot.com/articles/the-easy-way-to-extract-useful-text-from-arbitrary-html/, accessed on 02/05/2015
- Joseph Acanfora, Stanislaw Antol, Souleiman Ayoub, et al. Vtechworks: CS4984 Computational Linguistics. https://vtechworks.lib.vt.edu/handle/10919/50956 accessed on 02/05/2015.
- Ari Pollak, Include OutputFormat for a specified Avro schema that works with Streaming. https://issues.apache.org/jira/browse/AVRO-1067, accessed on 03/29/2015.
- Michael G. Noll, Using Avro in MapReduce Jobs With Hadoop, Pig, Hive. http://www.michael-noll.com/blog/2013/07/04/using-avro-in-mapreduce-jobs-with-hadoop-pig-hive/, accessed on 03/29/2015.
- Leonard Richardson, BeautifulSoup, http://www.crummy.com/software/BeautifulSoup/, accessed on 04/30/2015.
- Yuri Baburov, python-readability, https://github.com/buriy/python-readability, accessed on 04/30/2015.
- Michal Danilák, 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, accessed on 04/30/2015.

Thank you

{wxw, peecee}@vt.edu