Text Analytics and Entity Resolution Using Apache Spark

A Report submitted to the

Indian School of Business, Hyderabad

For Certificate Programme in Business Analytics

By

Hardik Gupta (71620027)
Praveen Bhandari (71620051)
Soumya Mallick (71620071)
Vaibhav Mhaske (71620083)

Under the guidance of
Prof. Peeyush Taori

2016-2018

Contents

Executive Summary

There is a vast and rapidly increasing quantity of online news content, each source publishing or covering almost the same news - content wise. Since these articles or entities do not share any common attribute but still have an underlying relationship, there is inherently the problem of identifying and linking or grouping different data content of the same real world entity. One solution to this problem can be achieved using Entity Resolution.

Entity Resolution, or "Record linkage" refers to the process of joining records from one data source with another that describe the same entity. Entity Resolution (ER) refers to the task of finding records in a dataset that refer to the same entity across different data sources – in this case news articles from two major Indian news channel, Indian Express and The Hindu. The task of resolving entities and detecting relationships becomes easy with ER, particularly when combining two datasets may or may not share a common identifier.

The objective of this report is to present the result of how Apache Spark can be used to apply powerful text analysis techniques and perform entity resolution across two datasets of news articles. The report highlights the process of data collection using web scraping RSS feeds of the news sources and cleaning, using Entity Resolution technique group similar news articles and also display similar news articles as per the search query input from the user, and finally divide the results as per the sentiment carried by the respective news article.

Keywords – Apache Spark, Text analytics, Entity Resolution, Sentiment Analysis

Business Problem

The report attempts to highlight the following business problem

- Given two dataset of news articles from different sources, identify and group similar news article content wise
- Given a large collection of all news articles, display/recommend articles to the user as per the input search provided and further identify the sentiment carried by the news articles and present the results as per their polarity – Positive, Negative or Neutral

Methodology

The following summarises the methodology used

a. Data Collection

We performed web scraping in R to procure data from two Indian news channel RSS feed – **The Hindu** and **The Indian Express** using 'rvest' package in R. The data was collected for around 40 days at an average interval of 10 hours from the following feeds

- http://indianexpress.com/section/india/feed/
- http://www.thehindu.com/?service=rss

The attributes that we collect from these feeds are

- Title – The main title of the news article
- Link – The web URL of the news article
- PubDate – The published date of the news article
- Desc – A short description of the news article to highlight what the article is about

R codes to collect data - **Webscraping_Feeds.R**

b. Data Cleaning

We performed data cleaning and corpus building in R using 'tm' package. In this process we collated the entire data collected over days into single CSV file individually for the two data sources, removed white space, converted to asci characters for attributes – Title and Description. We also add a unique ID to each news article. The final attributes in the data set are

- ID – The unique ID for each news article
- Title – The main title of the news article
- Link – The web URL of the news article
- PubDate – The published date of the news article
- Desc – A short description of the news article to highlight what the article is about

R code for cleaning data - **Cleaning_Corpus.R**

Final data files – **hindu.csv, ie.csv**

c. ER as Text Similarity - Weighted Bag-of-Words using TF-IDF in PySpark

In a given corpus, some tokens have higher importance than the others, which add value overall. We have used 'Term Frequency – Inverse Document Frequency' or TF-IDF to assign the correct weight to each token in the entire corpus. These weights correctly specify which tokens are to be favoured and give better results when comparing different documents

i. Term Frequency (TF)

Term Frequency computes the frequency of a token in the same document. E.g. if document d contains 100 tokens and token t appears in d 20 times, then the TF weight of t in d is 20/100 = 1/5. Usually the intuition is that if a token appears more often in a document, then it is more important or carries more meaning in the document

ii. Inverse Document Frequency (IDF)

IDF identifies all the tokens that are rare overall in the entire corpus. Rare tokens add more value than the common ones and it also helps in identifying the two documents sharing the rarer token as compared to the common token easily. IDF for a token t in a collection of documents D is calculated as follows

- D is the total number of documents in the entire corpus
- Compute n(t), the number of documents in D which contain the token t
- IDF(t) = D/n(t) (total documents in corpus) / (documents containing token t)

iii. Term Frequency- Inverse Document Frequency (TF-IDF)

The total TF-IDF weight for a token in a document is the product of its TF and IDF weights.

d. Compute Similarity between text using Cosine Similarity

We have used Cosine Similarity as the index to compute distance between the two documents (strings). Here each document is treated as a vector of their TF-IDF weights. Then to compute similarity, the cosine angle between the two documents is calculated.

If *A* and *B* are two documents whose similarity is to be calculated, then we calculate token vectors as

A – Vector of its tokens represented by their TF-IDF weights
B – Vector of its tokens represented by their TF-IDF weights

Then determine their similarity using the formula

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Where,

A.B – Dot product between the two vectors

||A|| - Magnitude of vector A

||B|| - Magnitude of vector B

The angle between the two documents is small if they share many common tokens and hence their similarity score is high (ideally = 1). This is because small angle means that they are pointing in the same direction. If the angle between them increases, it means they are pointing in different direction and hence their similarity score tends to 0.

e. Grouping similar news articles using Python-Igraph package

Python igraph is the graph package used to create and perform graph analysis. Here we have used this package to group all similar news articles and cluster them together

f. Sentiment Analysis using TextBlob package

TextBlob is a python library used to process textual data. As part of its API, it provides many common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more.

For section c, d, e, f code - **computeSimilarity.py**

Analysis and Results

Analysis

We read in the two CSV files to create two data frames. We also read the stopwords file which contains all the commonly used words such as articles, conjunctions – words that don't add value.

The data frames are used to create two dictionaries, one for each source. The two dictionaries are then merged together to form a single dictionary which has key as the news article id and value as title, link and publication date of the news article. The output of merged dictionary looks like this

```
File  Edit  View  Search  Terminal  Help
          'Fri, 09 Dec 2016 14:05:18 +0000'],
 'IE990': ['alienation of any segment of countrys citizens not healthy says hamid ansari',
           'http://indianexpress.com/article/india/alienation-of-any-segment-of-countrys-citizens-not-healthy-says-hamid-ansari-4449663/',
           'Wed, 28 Dec 2016 23:06:10 +0000'],
 'IE991': ['why its barter or cash in this cashless ambala village',
           'http://indianexpress.com/article/india/why-its-barter-or-cash-in-this-cashless-ambala-village-4449656/',
           'Wed, 28 Dec 2016 23:02:02 +0000'],
 'IE992': ['bank manager kills self in bengaluru',
           'http://indianexpress.com/article/india/bank-manager-kills-self-in-bengaluru-4449662/',
           'Wed, 28 Dec 2016 23:00:12 +0000'],
 'IE993': ['pm modi intervenes gets railways and defence to put aside land tussle',
           'http://indianexpress.com/article/india/pm-modi-intervenes-gets-railways-and-defence-to-put-aside-land-tussle-4449631/',
           'Wed, 28 Dec 2016 22:56:20 +0000'],
 'IE994': ['bastar rights advocate writes to nhrc says threat tactics being used',
           'http://indianexpress.com/article/india/bastar-rights-advocate-writes-to-nhrc-says-threat-tactics-being-used-4449638/',
           'Wed, 28 Dec 2016 22:51:42 +0000'],
 'IE995': ['pathankot repeat terrorists used tree along nagrota base wall to gain entry',
           'http://indianexpress.com/article/india/pathankot-attack-repeat-terrorists-used-tree-along-nagrota-base-wall-to-gain-entry-4449629/',
           'Wed, 28 Dec 2016 22:48:18 +0000'],
 'IE996': ['wb its economy crippled by clashes dhulagarh picks up the pieces',
           'http://indianexpress.com/article/india/west-bengal-dhulagarh-violence-economy-crippled-4449637/',
           'Wed, 28 Dec 2016 22:47:58 +0000'],
 'IE997': ['ratan tata calls on rss chief in nagpur bjp says courtesy call',
           'http://indianexpress.com/article/india/ratan-tata-calls-on-rss-chief-in-nagpur-bjp-says-courtesy-call-4449632/',
           'Wed, 28 Dec 2016 22:38:13 +0000'],
 'IE998': ['up elections akhilesh yadav hits back after mulayam puts out poll list rules out pact',
           'http://indianexpress.com/article/india/up-elections-snubbing-akhilesh-mulayam-puts-out-assembly-poll-list-rules-out-pact/',
           'Wed, 28 Dec 2016 22:30:27 +0000'],
 'IE999': ['up sealdah ajmer train derails at least 63 injured',
           'http://indianexpress.com/article/india/kanpur-train-derailed-sealdah-ajmer-express-indore-patna-4449570/',
           'Wed, 28 Dec 2016 22:04:04 +0000']}
```

Our results are based on comparing the title of each article. We create two RDDs for each news source containing the ID and title of the news article.

Each RDD is then mapped to tokenise the title into words or tokens

```
>>>
>>> hinduRecToToken.take(5)
[('HIN1', ['wild', 'elephant', 'sidda', 'breathes']), ('HIN2', ['farmers', 'tiruchi', 'district', 'sc', 'decisio
ning', 'suspended', 'tiruchi']), ('HIN4', ['obama', 'orders', 'full', 'review', 'russian', 'hacking', 'election'
reddy', 'traced', 'back', 'days', 'ballari', 'tahsildar'])]
>>> ieRecToToken.take(5)
[('IE1', ['rajagopalacharis', 'vision', 'political', 'space', 'honouring', 'individual', 'liberty', 'free', 'fai
collapse', '11', 'bodies', 'recovered', 'debris', 'owner', 'arrested']), ('IE3', ['delhi', '550', 'students', 'd
 'record']), ('IE4', ['badals', 'public', 'money', 'organise', 'moga', 'rally']), ('IE5', ['kaithal', 'embezzlem
, 'sit', 'complete', 'probe', 'weeks'])]
>>>
```

We then merge the two token data set to form a corpus. This step is done to calculate the IDF weights of each token across the corpus

```
>>> nitems = take(20, idfsWeights.iteritems())
>>> pprint(nitems)
[('mohini', 24992.0),
 ('paperless', 8330.666666666666),
 ('unscientific', 12496.0),
 ('radiologists', 24992.0),
 ('kadekar', 24992.0),
 ('institutionalised', 24992.0),
 ('yellow', 12496.0),
 ('kalyana', 8330.666666666666),
 ('katha', 12496.0),
 ('prefix', 24992.0),
 ('jihad', 6248.0),
 ('spiders', 24992.0),
 ('hanging', 2499.2),
 ('enlighten', 12496.0),
 ('woody', 12496.0),
 ('indranis', 24992.0),
 ('cyprus', 24992.0),
 ('towns', 24992.0),
 ('kandhamal', 24992.0),
 ('functioningfrom', 24992.0)]
>>>
```

Next, we compute Cartesian product of each news article from source 1 with each news article for source 2. This is done by using the 'cartesian' function on the two RDDs. Cartesian Product will result into tuples of news articles which is then used to calculate the similarity.

```
>>> crossProduct.take(10)
[(('HIN1', 'wild elephant sidda breathes its last'), ('IE1', 'c rajagopalacharis vision a political space honouring individual liberty
HIN1', 'wild elephant sidda breathes its last'), ('IE2', 'hyderabad building collapse 11 bodies recovered from debris owner arrested')
breathes its last'), ('IE3', 'delhi 550 students dress up as albert einstein eye guinness world record')), (('HIN1', 'wild elephant si
badals used public money to organise moga rally')), (('HIN1', 'wild elephant sidda breathes its last'), ('IE5', 'kaithal embezzlement
rs sit to complete probe in six weeks')), (('HIN1', 'wild elephant sidda breathes its last'), ('IE6', 'scrapping of import duty on whe
te threaten to launch agitation soon')), (('HIN1', 'wild elephant sidda breathes its last'), ('IE7', '40 lakh drug addicts in punjab s
'wild elephant sidda breathes its last'), ('IE8', 'india must spend 6 of gdp on education says manmohan singh')), (('HIN1', 'wild elep
IE9', 'punjab polls congressmen should get preference over turncoats says sonia gandhi')), (('HIN1', 'wild elephant sidda breathes its
t now when he was settled destiny did this to family there is wave of grief in the village'))]
>>>
```

The steps to find similar news articles can be summarised as below

a) Given two news article

   Article 1: “goa stares at a dull tourist season this year end”
   Article 2: “note ban goa stares at a dull tourist season this year end”

b) Tokenise each news article title and compute a list of tokens without stopwords. Fetch TF-IDF weight for each token

Tokens Article 1	Tokens Article 2
goa	note
stares	ban
dull	goa
tourist	stares
season	dull
year	tourist
end	season
	year
	end

c) The output from step b is a dictionary of tokens and their TF-IDF weights. This is then used to compute similarity using the cosine formula. The dot product of each vector is calculated, which is divided by the product of the magnitude of each individual vector
d) The final output is a tuple of (Article 1 ID, Article 2 ID, Cosine Similarity score)
e) The score is in the range of 0 to 1. Higher the score, better is the similarity between the two articles

Results

Part 1: Similar News Articles having similarity score greater than 0.7

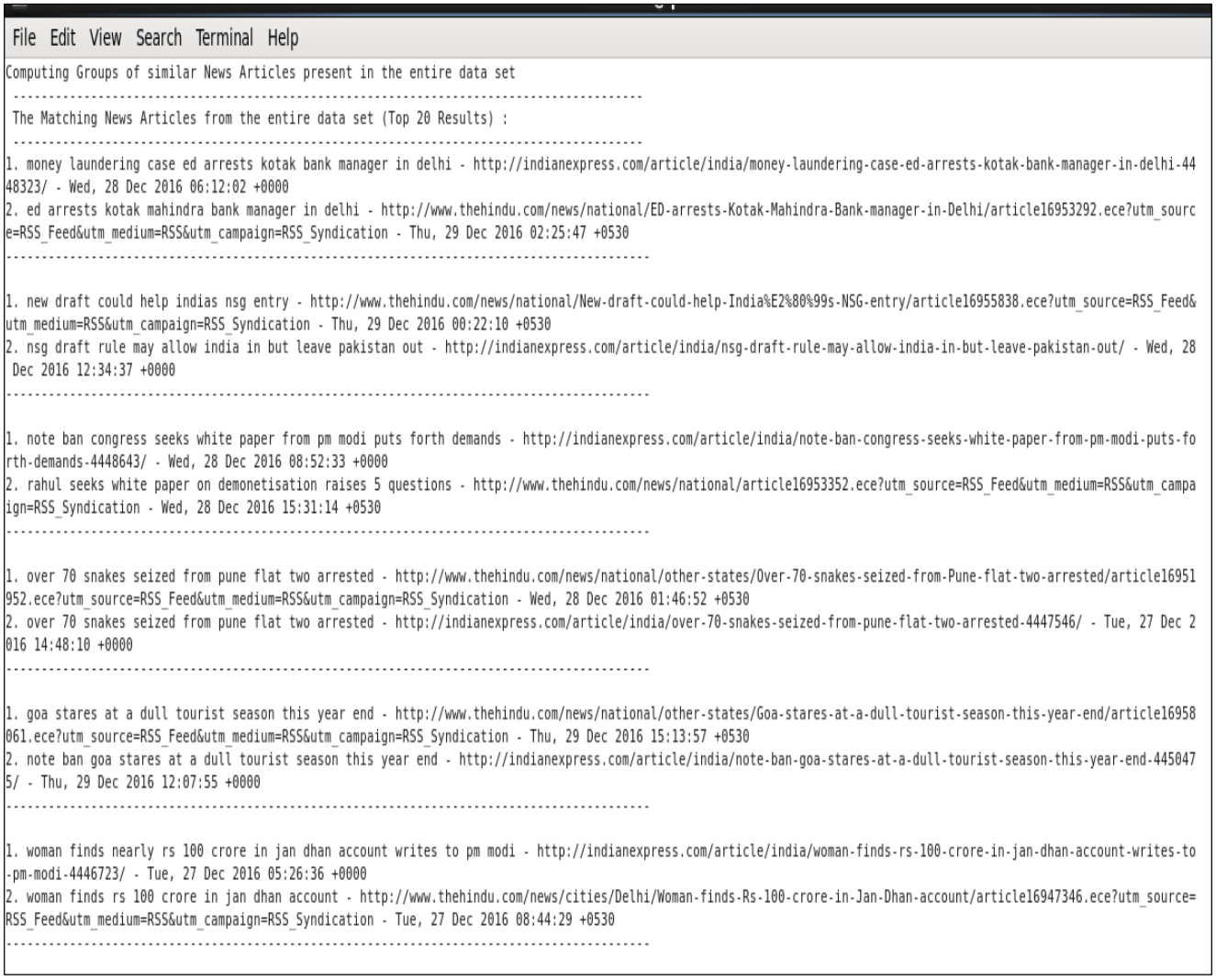
After calculating the similarity scores, we group all the news articles which are similar in content. We first filter our results for similarity score greater than 0.7, and then using ‘python-igraph’ package we compute groups of similar news articles. Below screenshot shows few groups of articles

e.g.

"HIN1968","goa stares at a dull tourist season this year end”, "http://www.thehindu.com/news/national/other-states/Goa-stares-at-a-dull-tourist-season-this-year-end/article16958061.ece?utm_source=RSS_Feed&utm_medium=RSS&utm_campaign=RSS_Syndication","Thu, 29 Dec 2016 15:13:57 +0530","goa seems to be reeling under the after effects of demonetisation as the regular week long beach parties ahead of the new year celebrations in the coastal state are yet to begin this time a much sou"

"IE946","note ban goa stares at a dull tourist season this year end”, "http://indianexpress.com/article/india/note-ban-goa-stares-at-a-dull-tourist-season-this-year-end-4450475/","Thu, 29 Dec 2016 12:07:55 +0000","state tourism minister dilip parulekar however ruled out the worries stating that this new year season will have tourists flooding the beaches"

Few Other Matching Groups:



```
1. bjp warns congress of legal action for baseless allegations against modi - http://www.thehindu.com/news/national/BJP-warns-Congress-of-legal-action-for-%E2%80%98base
less-allegations%E2%80%99-against-Modi/article16958976.ece?utm_source=RSS_Feed&utm_medium=RSS&utm_campaign=RSS_Syndication - Thu, 29 Dec 2016 18:34:15 +0530
2. bjp warns congress of legal action for its baseless allegations against pm modi amit shah - http://indianexpress.com/article/india/bjp-warns-congress-of-legal-actior
-for-its-baseless-allegations-against-pm-modi-amit-shah-4450533/ - Thu, 29 Dec 2016 12:31:07 +0000
---------------------------------------------------------------------------------------

1. us gang rape victim identifies three accused - http://www.thehindu.com/news/cities/Delhi/US-gang-rape-victim-identifies-three-accused/article16951840.ece?utm_source=
RSS_Feed&utm_medium=RSS&utm_campaign=RSS_Syndication - Wed, 28 Dec 2016 01:22:12 +0530
2. us woman gangrape victim identifies 3 of 4 accused - http://indianexpress.com/article/india/us-woman-gangrape-victim-identifies-3-of-4-accused-4447680/ - Tue, 27 Dec
 2016 16:29:30 +0000
---------------------------------------------------------------------------------------

1. rohit deo to be next advocate general - http://www.thehindu.com/news/cities/mumbai/Rohit-Deo-to-be-next-Advocate-General/article16951002.ece?utm_source=RSS_Feed&utm_
medium=RSS&utm_campaign=RSS_Syndication - Tue, 27 Dec 2016 23:47:06 +0530
2. maharashtra cabinet nod for rohit deo as new advocate general - http://indianexpress.com/article/india/maharashtra-cabinet-nod-for-rohit-deo-as-new-advocate-general-
4447229/ - Tue, 27 Dec 2016 11:00:47 +0000
---------------------------------------------------------------------------------------

1. pil in delhi hc on applicability of constitutional amendments to jammu kashmir - http://indianexpress.com/article/india/pil-in-delhi-hc-on-applicability-of-constitut
ional-amendments-to-jammu-kashmir-4447492/ - Tue, 27 Dec 2016 14:36:36 +0000
2. pil in delhi hc on applicability of constitutional amendments to j k - http://www.thehindu.com/news/national/other-states/PIL-in-Delhi-HC-on-applicability-of-Constit
utional-amendments-to-JK/article16950026.ece?utm_source=RSS_Feed&utm_medium=RSS&utm_campaign=RSS_Syndication - Tue, 27 Dec 2016 18:38:18 +0530
---------------------------------------------------------------------------------------

1. manmohan singh to release congress poll manifesto for punjab - http://www.thehindu.com/news/national/Manmohan-Singh-to-release-Congress-poll-manifesto-for-Punjab/art
icle16960404.ece?utm_source=RSS_Feed&utm_medium=RSS&utm_campaign=RSS_Syndication - Thu, 29 Dec 2016 23:25:34 +0530
2. manmohan singh to release punjab congress election manifesto - http://indianexpress.com/article/india/manmohan-singh-to-release-punjab-congress-election-manifesto-44
50528/ - Thu, 29 Dec 2016 11:51:31 +0000
---------------------------------------------------------------------------------------

1. indrani mukerjea performs post death rituals for father - http://indianexpress.com/article/india/indrani-mukerjea-performs-post-death-rituals-for-father-4447651/ - T
ue, 27 Dec 2016 16:10:53 +0000
2. indrani mukerjea performs fathers post death rituals - http://www.thehindu.com/news/cities/mumbai/Indrani-Mukerjea-performs-father%E2%80%99s-post-death-rituals/artic
le16951194.ece?utm_source=RSS_Feed&utm_medium=RSS&utm_campaign=RSS_Syndication - Wed, 28 Dec 2016 01:44:04 +0530
---------------------------------------------------------------------------------------
```

Note: The groups are not restricted to just two in number, it will form a group of all the matching articles greater than 0.7 if there any. Also this can be tuned further as per the similarity score desired

Part 2: Similar News Articles as per the User Input, Classify the output as per its Sentiment – Positive, Negative, Neutral

We gave the input string as 'what is demonetisation'. We obtained the matching news articles with similarity score 0.01 and above as shown in the screenshot below.

We used 'textblob' package in python to compute the Sentiment of each article. The sentiment was calculated on the attribute 'title' for each article. We obtained the results as shown below in the screenshot

Matching News Articles - Positive Sentiments

1. demonetisation rahul gandhi demands answers from pm modi mamata banerjee calls it super emergency - http://indianexpress.com/article/india/demonetisation-oppositon-joint-press-conference-rahul-gandhi-mamata-banerjee-tmc-congress-4447120/

Matching News Articles - Negative Sentiments

1. demonetisation a serious administrative flaw - http://www.thehindu.com/news/cities/Madurai/%E2%80%9CDemonetisation-a-serious-administrative-flaw%E2%80%9D/article16954347.ece?utm_source=RSS_Feed&utm_medium=RSS&utm_campaign=RSS_Syndication

2. post demonetisation auto sector to see sharp decline in dec - http://www.thehindu.com/business/Industry/Post-demonetisation-auto-sector-to-see-sharp-decline-in-Dec/article16958866.ece?utm_source=RSS_Feed&utm_medium=RSS&utm_campaign=RSS_Syndication

3. rbi refuses to give reasons behind demonetisation - http://www.thehindu.com/business/Economy/RBI-refuses-to-give-reasons-behind-demonetisation/article16958525.ece?utm_source=RSS_Feed&utm_medium=RSS&utm_campaign=RSS_Syndication

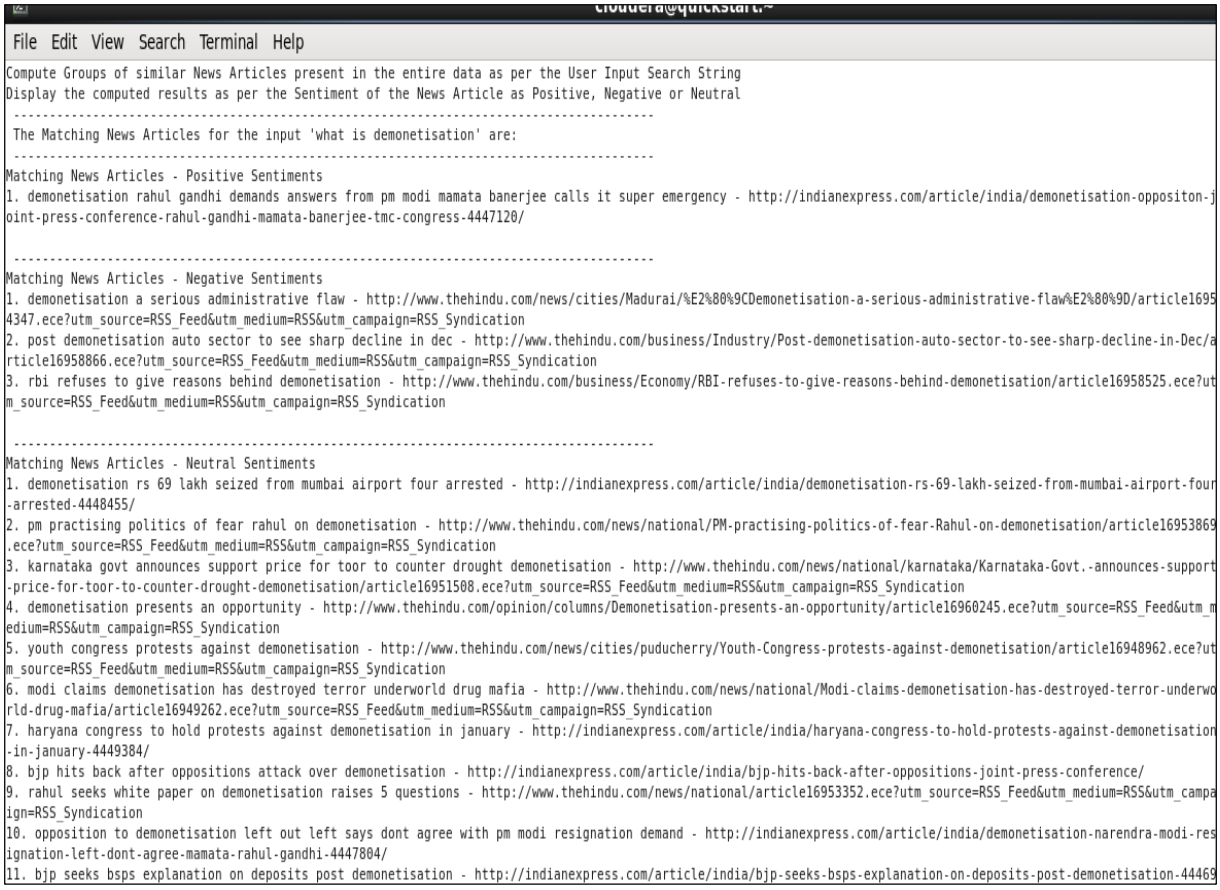
Matching News Articles - Neutral Sentiments

1. demonetisation rs 69 lakh seized from mumbai airport four arrested - http://indianexpress.com/article/india/demonetisation-rs-69-lakh-seized-from-mumbai-airport-four-arrested-4448455/

2. pm practising politics of fear rahul on demonetisation - http://www.thehindu.com/news/national/PM-practising-politics-of-fear-Rahul-on-demonetisation/article16953869.ece?utm_source=RSS_Feed&utm_medium=RSS&utm_campaign=RSS_Syndication

3. karnataka govt announces support price for toor to counter drought demonetisation - http://www.thehindu.com/news/national/karnataka/Karnataka-Govt.-announces-support-price-for-toor-to-

counter-drought-demonetisation/article16951508.ece?utm_source=RSS_Feed&utm_medium=RSS&utm_campaign=RSS_Syndication



We observe that textblob has done a decent classification of articles

Data and Code

The entire code and data sets can be procured from the following GitHub repository. The code is built with Spark 1.6.0

- https://github.com/HardikLGupta/BigDataSparkProject

File to refer to run the code - **Steps**

Conclusion

Entity Resolution is a good solution to apply when joining data sets from different sources and which don't share a common identifier. We observed how efficiently it created similar news article groups from a collection of over 24,000 articles. It also neatly fetched the news articles as per the user input string. Further the use of Apache Spark can boost the computation time with the aid of parallel computing.

Future Scope

This project can be further extended to include more news articles from different sources and perform computation. Further using the MLLib libraries of Apache Spark, we can perform Clustering analysis using algorithms such as Latent Dirichlet Allocation to perform topic modelling and compute different topic emerging from the underlying news articles.

References

Databricks - Text Analysis and Entity Resolution. (n.d.). Retrieved from https://docs.cloud.databricks.com/docs/latest/courses/Introduction%20to%20Big%20Data%20with%20Apache%20Spark%20(CS100-1x)/Solutions/Module%204:%20Text%20Analysis%20and%20Entity%20Resolution%20Lab%20Solutions.html

Entity Resolution for Big Data. (n.d.). Retrieved from Data Community DC: http://www.datacommunitydc.org/blog/2013/08/entity-resolution-for-big-data

Entity Resolution in the Web of Data. (n.d.). Retrieved from http://www.csd.uoc.gr/~vefthym/er/

IBM Knowledge Center - Entity resolution. (n.d.). Retrieved from IBM: http://www.ibm.com/support/knowledgecenter/SS2HSB_8.1.0/com.ibm.iis.ii.overview.doc/topics/eas_con_entityresolution.html

Merge tuples having atleast one common element to form a common tuple. (n.d.). Retrieved from Stack Overflow: http://stackoverflow.com/questions/41332988/merge-tuples-having-atleast-one-common-element-to-form-a-common-tuple

Stanford Entity Resolution Framework . (n.d.). Retrieved from http://infolab.stanford.edu/serf/