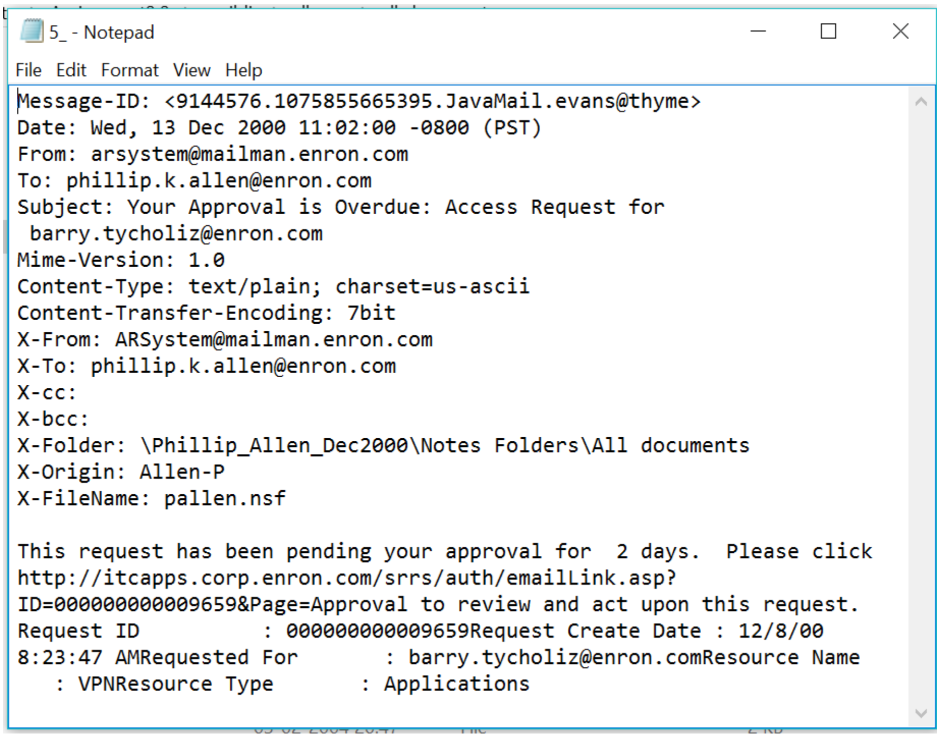
MACHINE LEARNING FOR BIG DATA ASSIGNMENT 2.2

Data Pre-Processing

- Header information such as “Mime-version” were removed from the mail
- Attachments are removed.
- Stemming is done, stopping words are removed using nltk.corpus- stopwords.
- Data is uploaded in AWS.

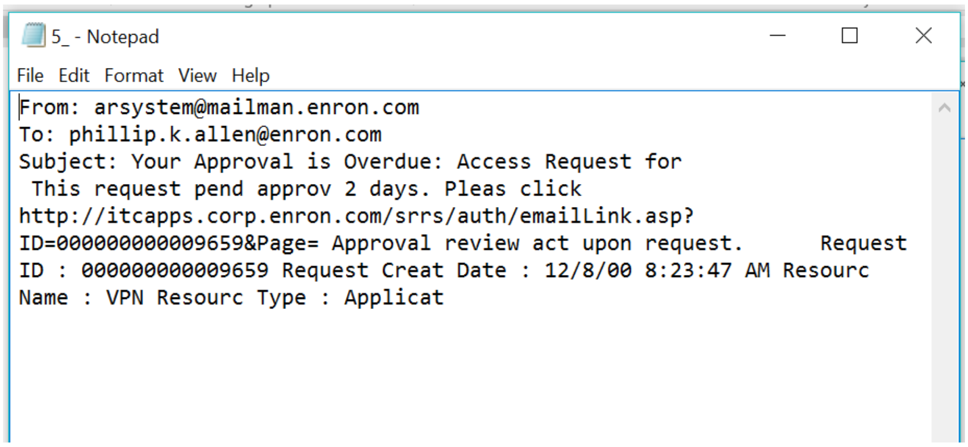
E-mail before Pre-Processing:



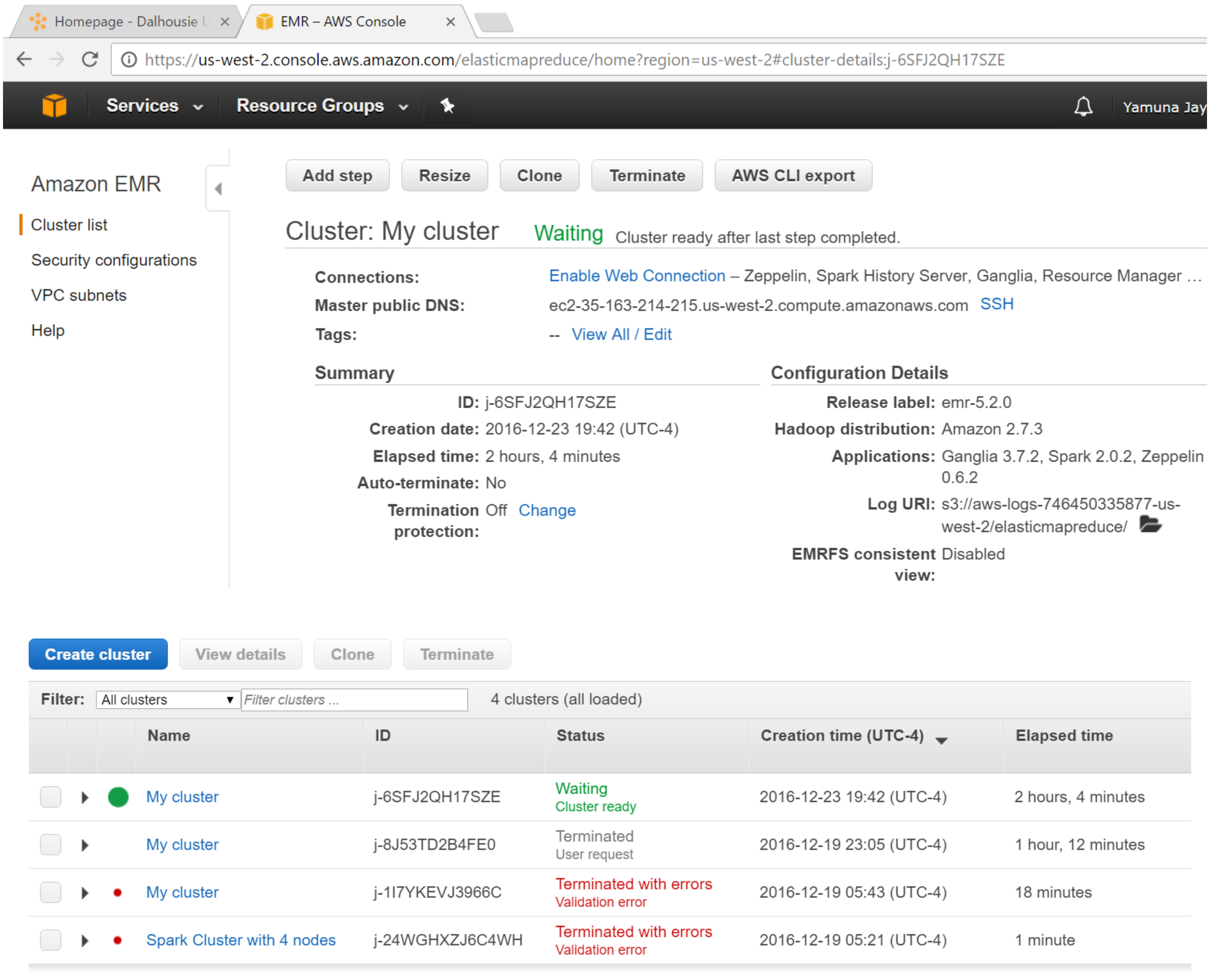
```
Message-ID: <9144576.1075855665395.JavaMail.evans@thyme>
Date: Wed, 13 Dec 2000 11:02:00 -0800 (PST)
From: arsystem@mailman.enron.com
To: phillip.k.allen@enron.com
Subject: Your Approval is Overdue: Access Request for
 barry.tycholiz@enron.com
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
X-From: ARSystem@mailman.enron.com
X-To: phillip.k.allen@enron.com
X-cc:
X-bcc:
X-Folder: \Phillip_Allen_Dec2000\Notes Folders\All documents
X-Origin: Allen-P
X-FileName: pallen.nsf

This request has been pending your approval for  2 days.  Please click
http://itcapps.corp.enron.com/srrs/auth/emailLink.asp?
ID=000000000009659&Page=Approval to review and act upon this request.
Request ID          : 000000000009659Request Create Date : 12/8/00
8:23:47 AMRequested For       : barry.tycholiz@enron.comResource Name
   : VPNResource Type       : Applications
```

E-mail after Pre-Processing:



```
From: arsystem@mailman.enron.com
To: phillip.k.allen@enron.com
Subject: Your Approval is Overdue: Access Request for
 This request pend approv 2 days. Pleas click
http://itcapps.corp.enron.com/srrs/auth/emailLink.asp?
ID=000000000009659&Page= Approval review act upon request.      Request
ID : 000000000009659 Request Creat Date : 12/8/00 8:23:47 AM Resourc
Name : VPN Resourc Type : Applicat
```



Cluster: My cluster Waiting Cluster ready after last step completed.

Connections: Enable Web Connection – Zeppelin, Spark History Server, Ganglia, Resource Manager …
Master public DNS: ec2-35-163-214-215.us-west-2.compute.amazonaws.com SSH
Tags: -- View All / Edit

Summary

ID: j-6SFJ2QH17SZE
Creation date: 2016-12-23 19:42 (UTC-4)
Elapsed time: 2 hours, 4 minutes
Auto-terminate: No
Termination protection: Off Change

Configuration Details

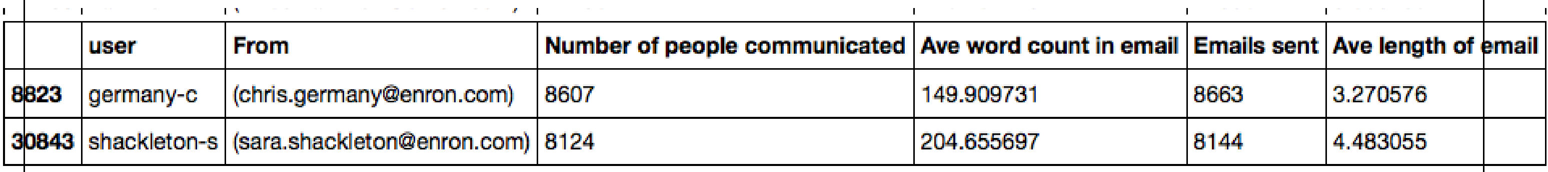
Release label: emr-5.2.0
Hadoop distribution: Amazon 2.7.3
Applications: Ganglia 3.7.2, Spark 2.0.2, Zeppelin 0.6.2
Log URI: s3://aws-logs-746450335877-us-west-2/elasticmapreduce/
EMRFS consistent view: Disabled

4 clusters (all loaded)

Name	ID	Status	Creation time (UTC-4)	Elapsed time
● My cluster	j-6SFJ2QH17SZE	Waiting Cluster ready	2016-12-23 19:42 (UTC-4)	2 hours, 4 minutes
My cluster	j-8J53TD2B4FE0	Terminated User request	2016-12-19 23:05 (UTC-4)	1 hour, 12 minutes
● My cluster	j-1I7YKEVJ3966C	Terminated with errors Validation error	2016-12-19 05:43 (UTC-4)	18 minutes
● Spark Cluster with 4 nodes	j-24WGHXZJ6C4WH	Terminated with errors Validation error	2016-12-19 05:21 (UTC-4)	1 minute

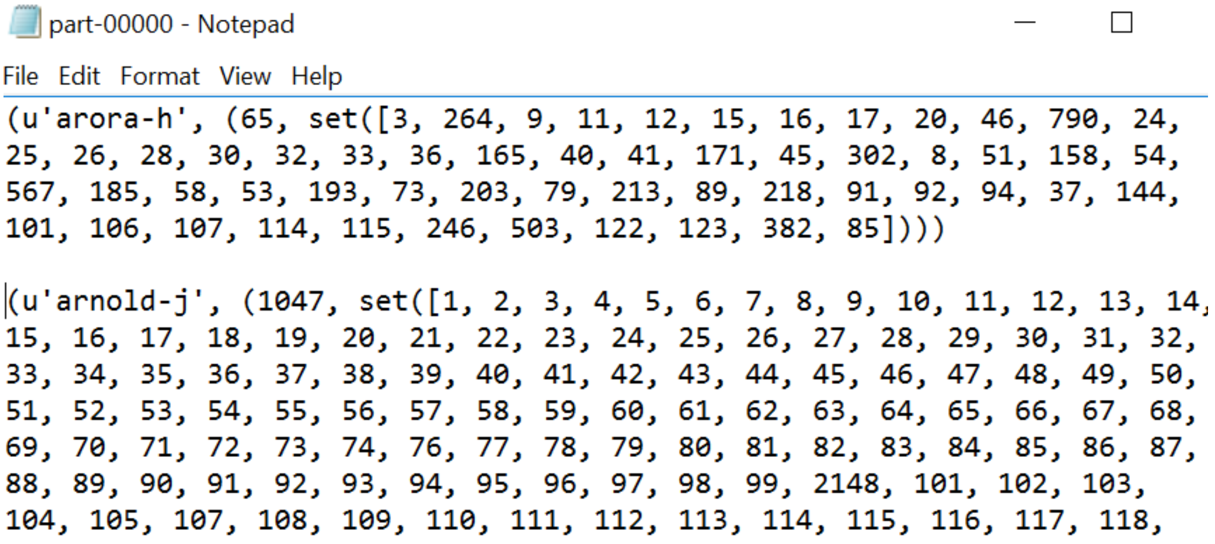
Data Discovery:

- The following are discovered in data discovery stage and are stored in file.



	user	From	Number of people communicated	Ave word count in email	Emails sent	Ave length of email
8823	germany-c	(chris.germany@enron.com)	8607	149.909731	8663	3.270576
30843	shackleton-s	(sara.shackleton@enron.com)	8124	204.655697	8144	4.483055

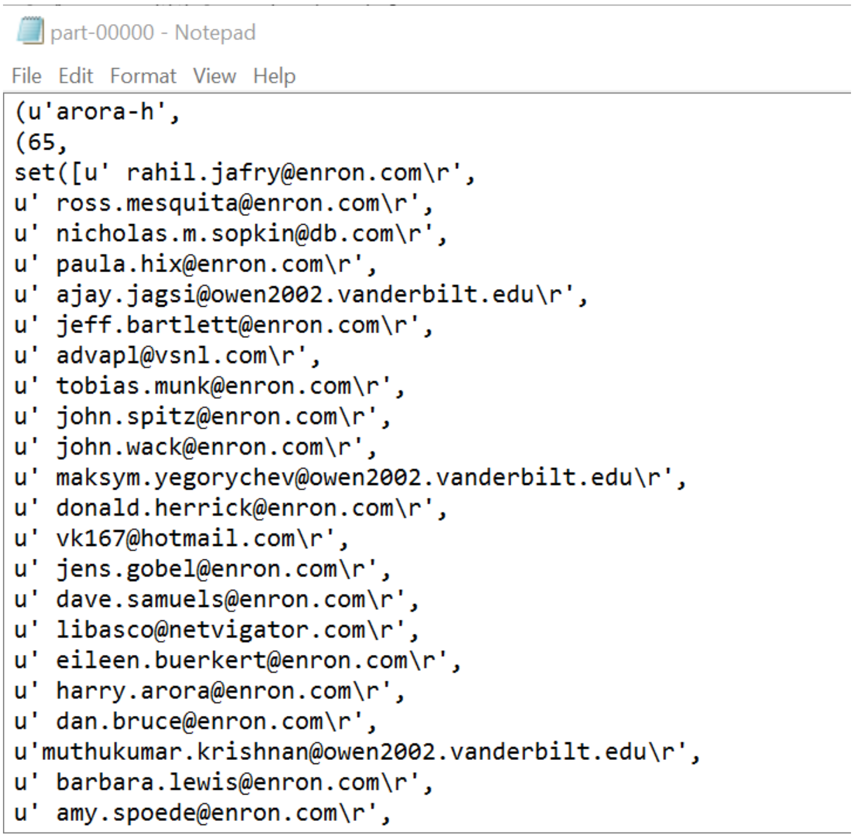
- Number of mails received – mails_received file
- Number of mails sent – mails_sent file
- User name Word count in each email – wordcount file



```
(u'arora-h', (65, set([3, 264, 9, 11, 12, 15, 16, 17, 20, 46, 790, 24,
25, 26, 28, 30, 32, 33, 36, 165, 40, 41, 171, 45, 302, 8, 51, 158, 54,
567, 185, 58, 53, 193, 73, 203, 79, 213, 89, 218, 91, 92, 94, 37, 144,
101, 106, 107, 114, 115, 246, 503, 122, 123, 382, 85])))

(u'arnold-j', (1047, set([1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14,
15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32,
33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50,
51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68,
69, 70, 71, 72, 73, 74, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87,
88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 2148, 101, 102, 103,
104, 105, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118,
```

- Number of mails sent and received by each user – numberofmails



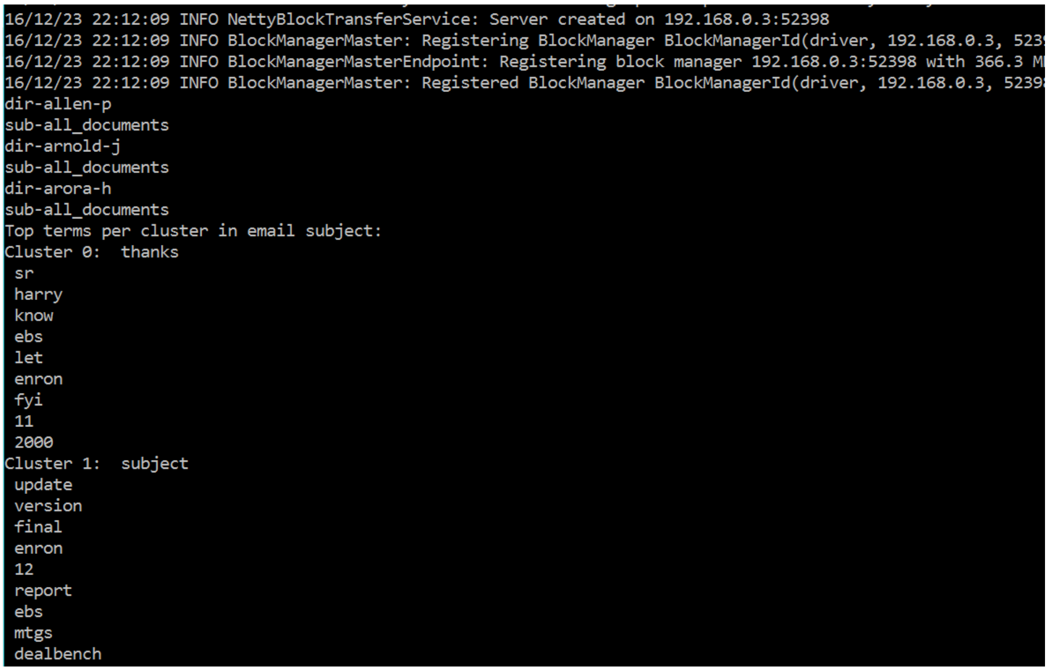
```
(u'arora-h',
(65,
set([u' rahil.jafry@enron.com\r',
u' ross.mesquita@enron.com\r',
u' nicholas.m.sopkin@db.com\r',
u' paula.hix@enron.com\r',
u' ajay.jagsi@owen2002.vanderbilt.edu\r',
u' jeff.bartlett@enron.com\r',
u' advapl@vsnl.com\r',
u' tobias.munk@enron.com\r',
u' john.spitz@enron.com\r',
u' john.wack@enron.com\r',
u' maksym.yegorychev@owen2002.vanderbilt.edu\r',
u' donald.herrick@enron.com\r',
u' vk167@hotmail.com\r',
u' jens.gobel@enron.com\r',
u' dave.samuels@enron.com\r',
u' libasco@netvigator.com\r',
u' eileen.buerkert@enron.com\r',
u' harry.arora@enron.com\r',
u' dan.bruce@enron.com\r',
u'muthukumar.krishnan@owen2002.vanderbilt.edu\r',
u' barbara.lewis@enron.com\r',
u' amy.spoede@enron.com\r',
```

Data Analysis – Task 1 :

K-Means is applied to email subject, body and following results were obtained.

For K-Means, allocating 2 clusters returned results with words having close relation with each other inside the cluster, than increasing the number of clusters.

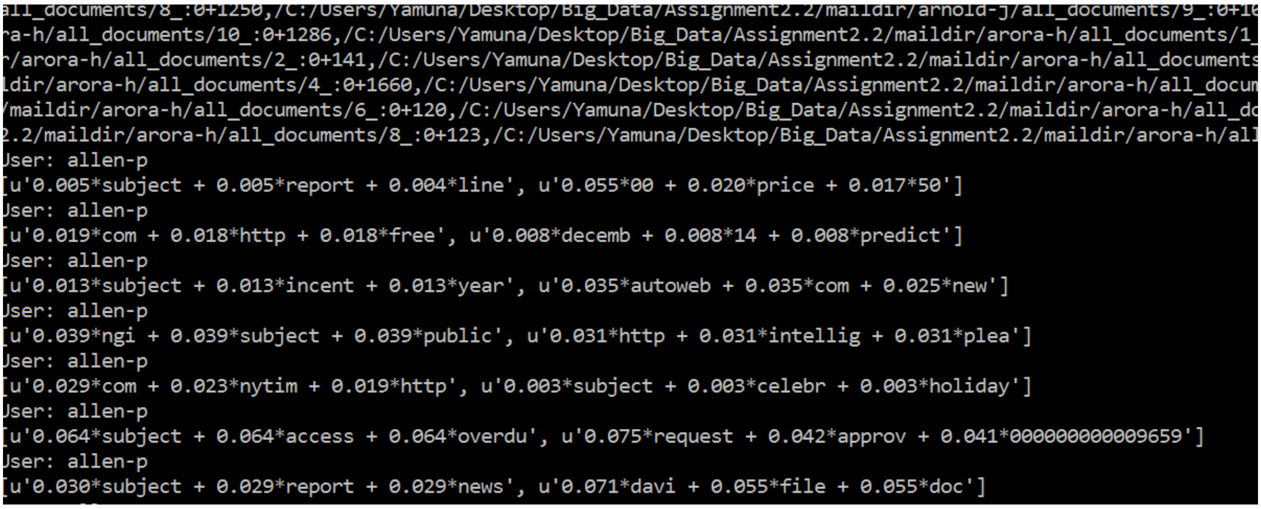
K-Means - Sample Results



```
16/12/23 22:12:09 INFO NettyBlockTransferService: Server created on 192.168.0.3:52398
16/12/23 22:12:09 INFO BlockManagerMaster: Registering BlockManager BlockManagerId(driver, 192.168.0.3, 523
16/12/23 22:12:09 INFO BlockManagerMasterEndpoint: Registering block manager 192.168.0.3:52398 with 366.3 M
16/12/23 22:12:09 INFO BlockManagerMaster: Registered BlockManager BlockManagerId(driver, 192.168.0.3, 5239
dir-allen-p
sub-all_documents
dir-arnold-j
sub-all_documents
dir-arora-h
sub-all_documents
Top terms per cluster in email subject:
Cluster 0:  thanks
 sr
 harry
 know
 ebs
 let
 enron
 fyi
 11
 2000
Cluster 1:  subject
 update
 version
 final
 enron
 12
 report
 ebs
 mtgs
 dealbench
```

LDA Clustering – Sample Result

Based on K-means clustering results, same number of clusters were assigned for LDA.



```
all_documents/8_:0+1250,/C:/Users/Yamuna/Desktop/Big_Data/Assignment2.2/maildir/arnold-j/all_documents/9_:0+1
ra-h/all_documents/10_:0+1286,/C:/Users/Yamuna/Desktop/Big_Data/Assignment2.2/maildir/arora-h/all_documents/1_
r/arora-h/all_documents/2_:0+141,/C:/Users/Yamuna/Desktop/Big_Data/Assignment2.2/maildir/arora-h/all_documents
ldir/arora-h/all_documents/4_:0+1660,/C:/Users/Yamuna/Desktop/Big_Data/Assignment2.2/maildir/arora-h/all_docum
/maildir/arora-h/all_documents/6_:0+120,/C:/Users/Yamuna/Desktop/Big_Data/Assignment2.2/maildir/arora-h/all_do
2.2/maildir/arora-h/all_documents/8_:0+123,/C:/Users/Yamuna/Desktop/Big_Data/Assignment2.2/maildir/arora-h/all
User: allen-p
[u'0.005*subject + 0.005*report + 0.004*line', u'0.055*00 + 0.020*price + 0.017*50']
User: allen-p
[u'0.019*com + 0.018*http + 0.018*free', u'0.008*decemb + 0.008*14 + 0.008*predict']
User: allen-p
[u'0.013*subject + 0.013*incent + 0.013*year', u'0.035*autoweb + 0.035*com + 0.025*new']
User: allen-p
[u'0.039*ngi + 0.039*subject + 0.039*public', u'0.031*http + 0.031*intellig + 0.031*plea']
User: allen-p
[u'0.029*com + 0.023*nytim + 0.019*http', u'0.003*subject + 0.003*celebr + 0.003*holiday']
User: allen-p
[u'0.064*subject + 0.064*access + 0.064*overdu', u'0.075*request + 0.042*approv + 0.041*000000000009659']
User: allen-p
[u'0.030*subject + 0.029*report + 0.029*news', u'0.071*davi + 0.055*file + 0.055*doc']
```

Data Analysis – Task 2 :

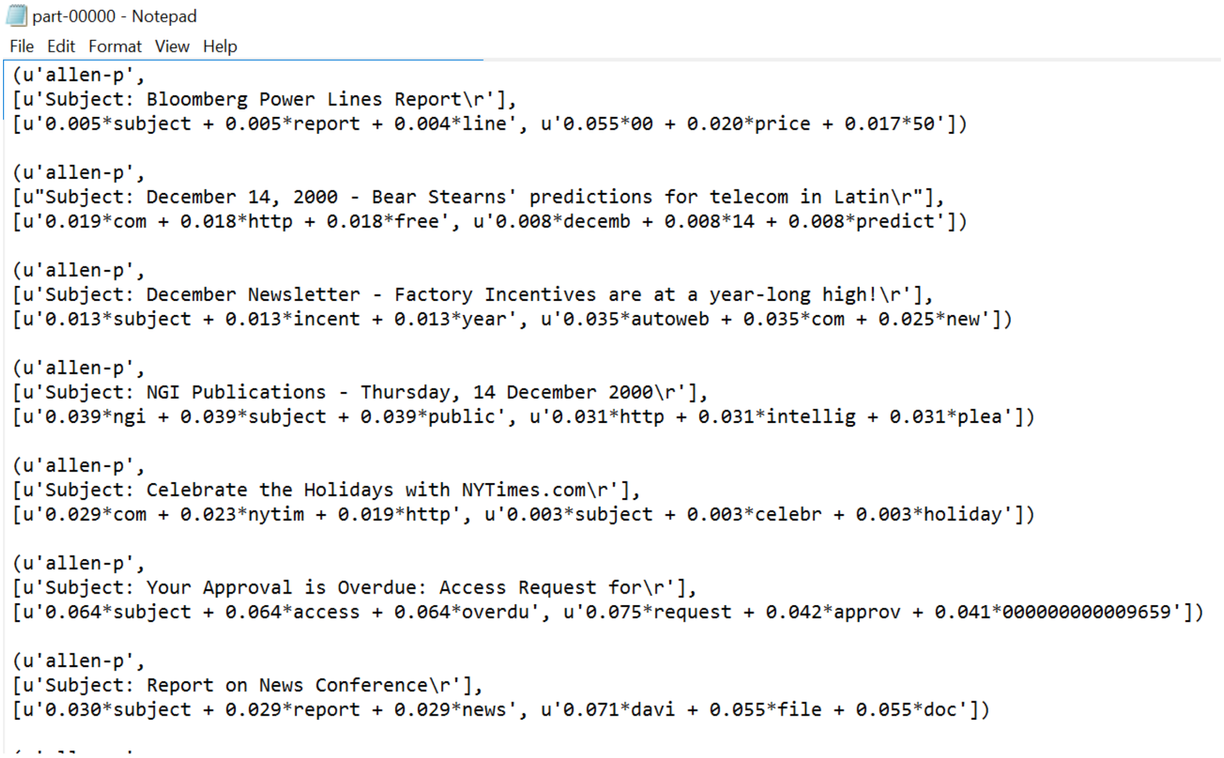
K-Means a classical clustering algorithm returned a single word per email

LDA(Latent Dirichlet Allocation) which is a Topic modeling algorithm retuned a probabilistic composition of the weighted labels in each email subject and body.

The difference is that K-means is to partition the mails in K disjoint clusters. On the other hand, LDA gives more relative comparison based on the word frequency.

Therefore each mail is clustered based on multiple key words.

Hence, LDA gives more realistic results than k-means for analysis of e-mails.



```
(u'allen-p',
[u'Subject: Bloomberg Power Lines Report\r'],
[u'0.005*subject + 0.005*report + 0.004*line', u'0.055*00 + 0.020*price + 0.017*50'])

(u'allen-p',
[u"Subject: December 14, 2000 - Bear Stearns' predictions for telecom in Latin\r"],
[u'0.019*com + 0.018*http + 0.018*free', u'0.008*decemb + 0.008*14 + 0.008*predict'])

(u'allen-p',
[u'Subject: December Newsletter - Factory Incentives are at a year-long high!\r'],
[u'0.013*subject + 0.013*incent + 0.013*year', u'0.035*autoweb + 0.035*com + 0.025*new'])

(u'allen-p',
[u'Subject: NGI Publications - Thursday, 14 December 2000\r'],
[u'0.039*ngi + 0.039*subject + 0.039*public', u'0.031*http + 0.031*intellig + 0.031*plea'])

(u'allen-p',
[u'Subject: Celebrate the Holidays with NYTimes.com\r'],
[u'0.029*com + 0.023*nytim + 0.019*http', u'0.003*subject + 0.003*celebr + 0.003*holiday'])

(u'allen-p',
[u'Subject: Your Approval is Overdue: Access Request for\r'],
[u'0.064*subject + 0.064*access + 0.064*overdu', u'0.075*request + 0.042*approv + 0.041*000000000009659'])

(u'allen-p',
[u'Subject: Report on News Conference\r'],
[u'0.030*subject + 0.029*report + 0.029*news', u'0.071*davi + 0.055*file + 0.055*doc'])
```