Query Based Summarized Email Extraction

GROUP 07

[1]Keshaw Singh [2]Roshan Kumar [3]Sandipan Mandal

[1]13347, skeshaw@iitk.ac.in
[2]13590, roshan@iitk.ac.in
[3]13616, mandals@iitk.ac.in

CS671A: Final Report

Under the guidance of Prof. Harish Karnick
Indian Institute of Technology, Kanpur

Contents

1 Introduction

1.1 Problem Statement

Our Problem statement can be summarised in one line as follows

“**Given a query, extract ‘topic-wise’ summarized email content from various conversations**”

To explain further, our system takes a query as input and returns summarised content of relevant mails in the database. If our query can correspond to more than one meaning, it returns seperate summary corresponding to each.

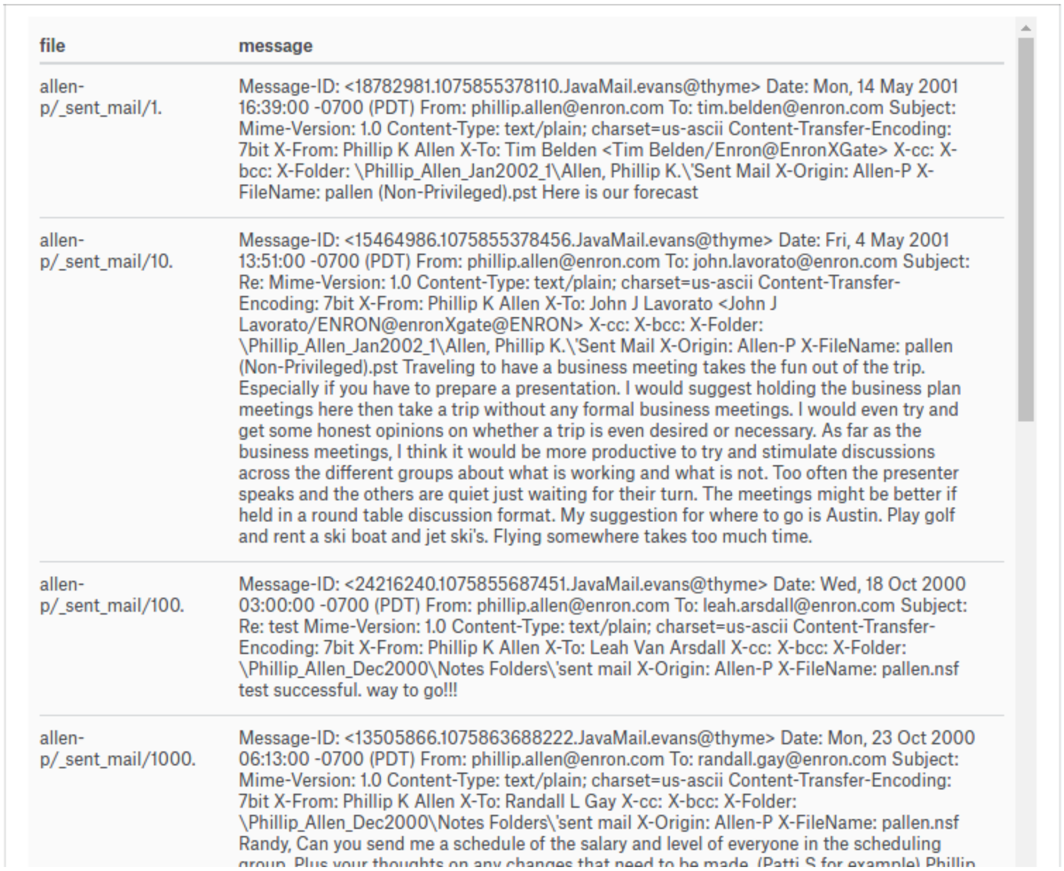
1.2 Dataset

The dataset we used is **Enron Email Dataset**
available at `https://www.kaggle.com/wcukierski/enron-email-dataset`.

Brief description of the dataset:-

- Consists of a single ‘emails.csv’ file.
- Data from 150 users, mostly senior management of enron.
- Has approximately 0.5M emails over course of 2 years.

A small preview of the **emails.csv** file.



file	message
allen-p/_sent_mail/1.	Message-ID: <18782981.1075855378110.JavaMail.evans@thyme> Date: Mon, 14 May 2001 16:39:00 -0700 (PDT) From: phillip.allen@enron.com To: tim.belden@enron.com Subject: Mime-Version: 1.0 Content-Type: text/plain; charset=us-ascii Content-Transfer-Encoding: 7bit X-From: Phillip K Allen X-To: Tim Belden <Tim Belden/Enron@EnronXGate> X-cc: X-bcc: X-Folder: \Phillip_Allen_Jan2002_1\Allen, Phillip K.\'Sent Mail X-Origin: Allen-P X-FileName: pallen (Non-Privileged).pst Here is our forecast
allen-p/_sent_mail/10.	Message-ID: <15464986.1075855378456.JavaMail.evans@thyme> Date: Fri, 4 May 2001 13:51:00 -0700 (PDT) From: phillip.allen@enron.com To: john.lavorato@enron.com Subject: Re: Mime-Version: 1.0 Content-Type: text/plain; charset=us-ascii Content-Transfer-Encoding: 7bit X-From: Phillip K Allen X-To: John J Lavorato <John J Lavorato/ENRON@enronXgate@ENRON> X-cc: X-bcc: X-Folder: \Phillip_Allen_Jan2002_1\Allen, Phillip K.\'Sent Mail X-Origin: Allen-P X-FileName: pallen (Non-Privileged).pst Traveling to have a business meeting takes the fun out of the trip. Especially if you have to prepare a presentation. I would suggest holding the business plan meetings here then take a trip without any formal business meetings. I would even try and get some honest opinions on whether a trip is even desired or necessary. As far as the business meetings, I think it would be more productive to try and stimulate discussions across the different groups about what is working and what is not. Too often the presenter speaks and the others are quiet just waiting for their turn. The meetings might be better if held in a round table discussion format. My suggestion for where to go is Austin. Play golf and rent a ski boat and jet ski's. Flying somewhere takes too much time.
allen-p/_sent_mail/100.	Message-ID: <24216240.1075855687451.JavaMail.evans@thyme> Date: Wed, 18 Oct 2000 03:00:00 -0700 (PDT) From: phillip.allen@enron.com To: leah.arsdall@enron.com Subject: Re: test Mime-Version: 1.0 Content-Type: text/plain; charset=us-ascii Content-Transfer-Encoding: 7bit X-From: Phillip K Allen X-To: Leah Van Arsdall X-cc: X-bcc: X-Folder: \Phillip_Allen_Dec2000\Notes Folders\'sent mail X-Origin: Allen-P X-FileName: pallen.nsf test successful. way to go!!!
allen-p/_sent_mail/1000.	Message-ID: <13505866.1075863688222.JavaMail.evans@thyme> Date: Mon, 23 Oct 2000 06:13:00 -0700 (PDT) From: phillip.allen@enron.com To: randall.gay@enron.com Subject: Mime-Version: 1.0 Content-Type: text/plain; charset=us-ascii Content-Transfer-Encoding: 7bit X-From: Phillip K Allen X-To: Randall L Gay X-cc: X-bcc: X-Folder: \Phillip_Allen_Dec2000\Notes Folders\'sent mail X-Origin: Allen-P X-FileName: pallen.nsf Randy, Can you send me a schedule of the salary and level of everyone in the scheduling group. Plus your thoughts on any changes that need to be made. (Patti S for example) Phillip

2 Pre-processing

2.1 Reasons for Pre-processing

- Difficult to use in the format provided.
- Flooded with irrelevant information.
- Inclusion of forwarded mails result in repetitive dataset.
- Unexpected tokens, numeric need to be appropriately handled.

2.2 Method of Pre-processing

- Used **pandas** library to properly organize the mail for later use. After a few processing steps any field of the mail can be accessed easily using **pandas**. For example suppose in python, we want to access the content, subject, to and from field of i^{th} mail

```
content = pandas_frame[‘content’][i]
subject = pandas_frame[‘Subject’][i]
To = pandas_frame[‘To’][i]
From = pandas_frame[‘From’][i]
```

- Wrote Regular Expressions to remove appended ‘Forwarded’ message.

3 Initial Steps

3.1 Using LDA

Latent Dirichlet Allocation(LDA) [2] represents documents as mixtures of topics that spit out words with certain probabilities.
Suppose we are given a set of documents. We decide upon the number of topics, say, K. Then learning happens as follows:

- Go through each document, and randomly assign each word in the document to one of the K topics.
- This random assignment already gives both topics representations of all the documents and the word distributions of all topics, though not very good ones.
- To improve upon these distributions, for each document d:
    - Go through each word w in d and for each topic t, compute two things:
        1. p(topic t | document d) = the proportion of words in document d that are currently assigned to topic t
        2. p(word w | topic t) = the proportion of assignments to topic t over all documents that come from this word w.

    Reassign w a new topic, where we choose topic t with probability p(topic t | document d) * p(word w | topic t) (according to our generative model, this is essentially the probability that topic t generated word w, so it makes sense that we resample the current word's topic with this probability). In other words, in this step, we're assuming that all topic assignments except for the current word in question are correct, and then updating the assignment of the current word using our model of how documents are generated.

  - After repeating the previous step a large number of times, we'll eventually reach a roughly steady state where our assignments are pretty good. So use these assignments to estimate the topic mixtures of each document (by counting the proportion of words assigned to each topic within that document) and the words associated to each topic (by counting the proportion of words assigned to each topic overall).

3.2 Simple Mail Grouping

- Assume forwarded or replied mails to have same subject appended with 'FW:' or 'RE:'
- Find relevant emails this way and arrange them in order of time sent to get the whole string
- Follow this by topic analysis and summarization
- People often modify subject and/or don't continue the conversation by replying back
- As many details are lost this way summarization and query search won't give results.

3.3 Trying Different Mail Representation

Various representations for email were tried, which include:

- Bag of Words
- Weighted Bag of Words
- Average of Bag of Words
- Bag of Vectors
- Weighted Bag of Vectors
- Doc2Vec [4]

In the end, we settled for Doc2Vec representation for our model. Its working has been explained below.

3.3.1 Distributed Representations of Documents(Doc2Vec)

We use Doc2vec which is an unsupervised model that learns fixed-length feature representations from variable pieces of text, such as sentences, paragraphs or documents. This has the potential to overcome the weakness of the bag-of-words model. This has been known to perform excellently on text as well as sentiment classification tasks. The two models of training tried are:

3.3.1.1 Paragraph Vector: A distributed memory model The key points to note how the model is trained are:

1. Every paragraph is mapped to a unique vector, represented by a column in matrix D and every word is also mapped to a unique vector, represented by a column in matrix W.
2. The paragraph vector and word vectors are concatenated to predict the next word in a context.
3. The model is the trained in a similar manner to word2vec with an additional document token along with word tokens to predict the next word in a sequence

This technique is shown in Figure 1

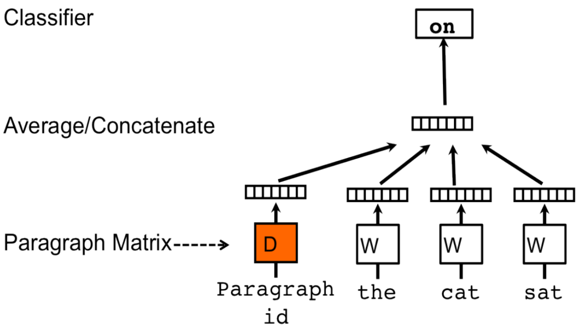


Figure 1: In this model, the concatenation or average of this vector with a context of three words is used to predict the fourth word. The paragraph vector represents the missing information from the current context and can act as a memory of the topic of the paragraph.

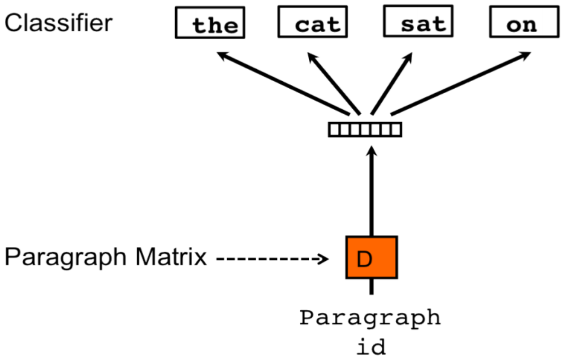


Figure 2: Distributed Bag of Words version of paragraph vectors. In this version, the paragraph vector is trained to predict the words in a small window.

3.3.1.2 Paragraph Vector without word ordering: Distributed bag of words The key points to note how the model is trained are:

1. We ignore the context words in the input, and force the model to predict words randomly sampled from the paragraph in the output.
2. We sample a text window, then sample a random word from the text window and form a classification task given the Paragraph Vector.

This technique is shown in Figure 2

3.4 Using K-Means

- Attempted clustering by K-Means using representations described before
- K is brought down from no. of groups obtained previously
- So obtained clusters are assigned “key-phrases” which serve to address our query

The nature of emails causes clustering to perform poorly.

Some representations did provide decent results but not what we wanted

4 Algorithm Followed

4.1 Algorithm Details

Pseudocode of the algorithm is mentioned in **Algorithm 1**

Algorithm 1 Get Summarized Mails

```
procedure SummarizedMails(Query, MailCorpus, StopWords, Idf)    ▷ Mail corpus is pre-processed
    PreprocessedQuery $\leftarrow$ RemoveStopWordsAndStem(Query)
    ImportantTerms $\leftarrow$ k Terms in PreprocessedQuery having highest idf.
    RelevantMails $\leftarrow$ All Mails in MailCorpus having any word matching any terms from ImportantTerms
    Clusters $\leftarrow$ BirchClustering(RelevantMails)
    for each cluster $\in$ Clusters do
        summary $\leftarrow$ LSASummarization(cluster)
        topic $\leftarrow$ ExtractTopic(summary)
        print topic
        print summary
```

4.2 Details of Steps Involved

4.2.1 Stemming and Stop words Removal

The first step in our pipeline is to remove stopwords and stem the contents and queries. It is required because of the following reasons:

- Users may query for some stopword, which don't carry much information and occurs in many mails.
- Users may query for different form of the same word. Without stemming, we might miss out on some form.

4.2.2 Email Extraction

- Use idf value of words to get the 3 most important words from the query. It is an important step because some words appear in many mails and hence don't carry any specific information. High idf implies appearance in lesser number of mails and hence carrying more information.
- Now use these 3(max) words to extract the mails where they occur.

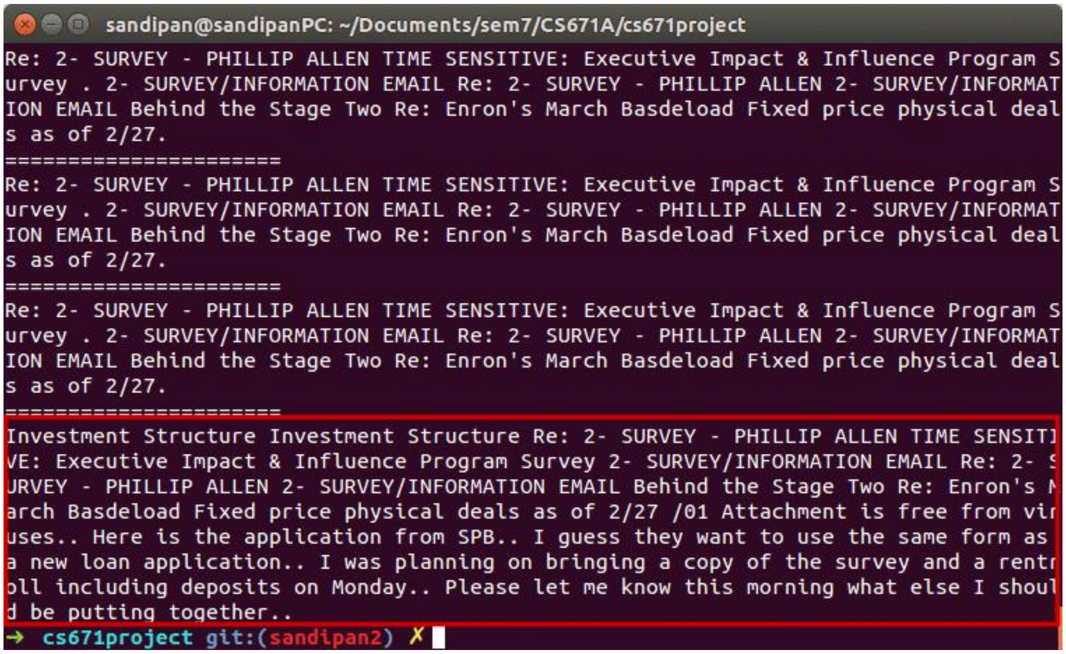
4.2.3 Clustering

BIRCH (balanced iterative reducing and clustering using hierarchies) [1] is an unsupervised data mining algorithm used to perform hierarchical clustering over particularly large data-sets. An advantage of BIRCH is its ability to incrementally and dynamically cluster incoming, multi-dimensional metric data points in an attempt to produce the best quality clustering for a given set of resources.

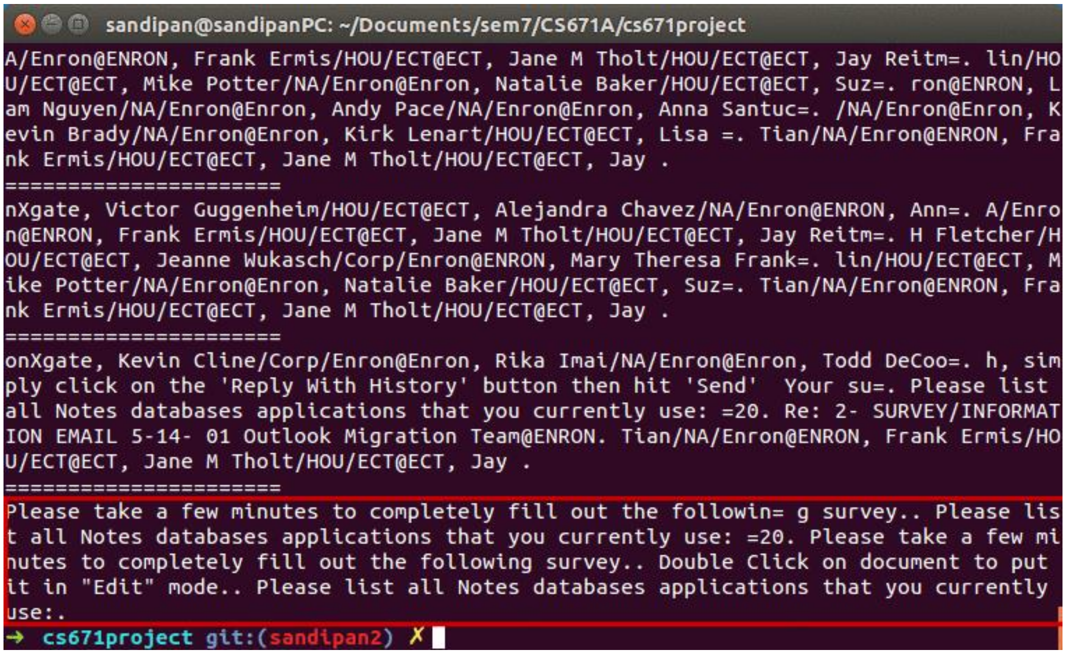
- Used BIRCH, with email vectors trained with Doc2Vec. Clustering was better at this stage because number of clusters was limited. No need to specify the number of clusters.

- Another reason for clustering was that there may be different instances of the same event. We need to differentiate between those instances.

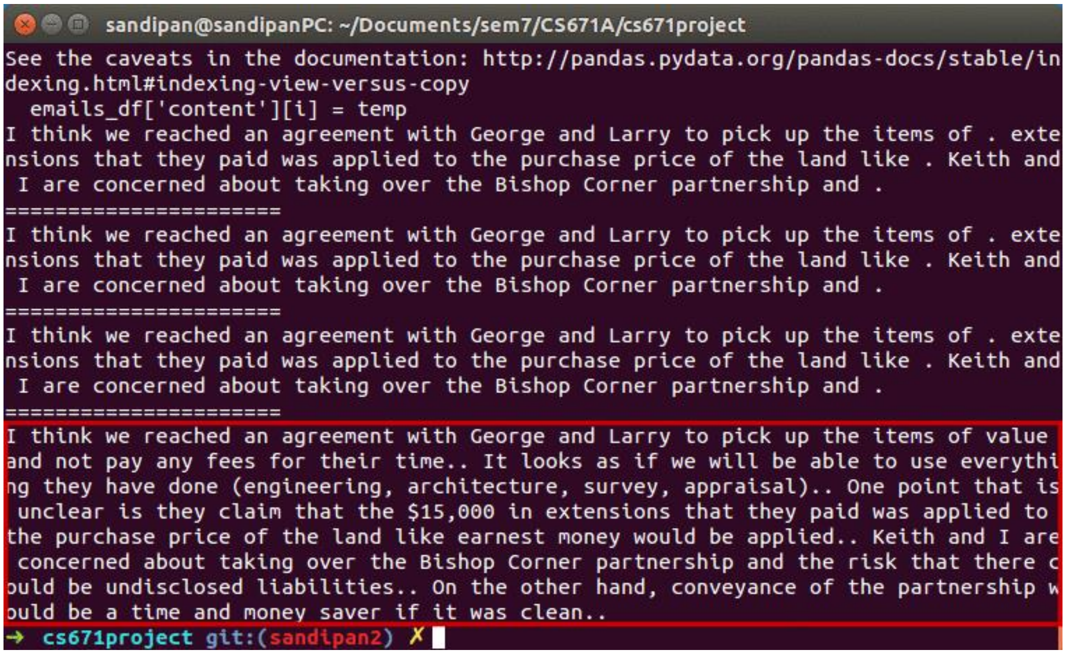
Consider the search term **survey**. As we can see from the results shown below, the clustering algorithm is able to properly differentiate between all the senses.



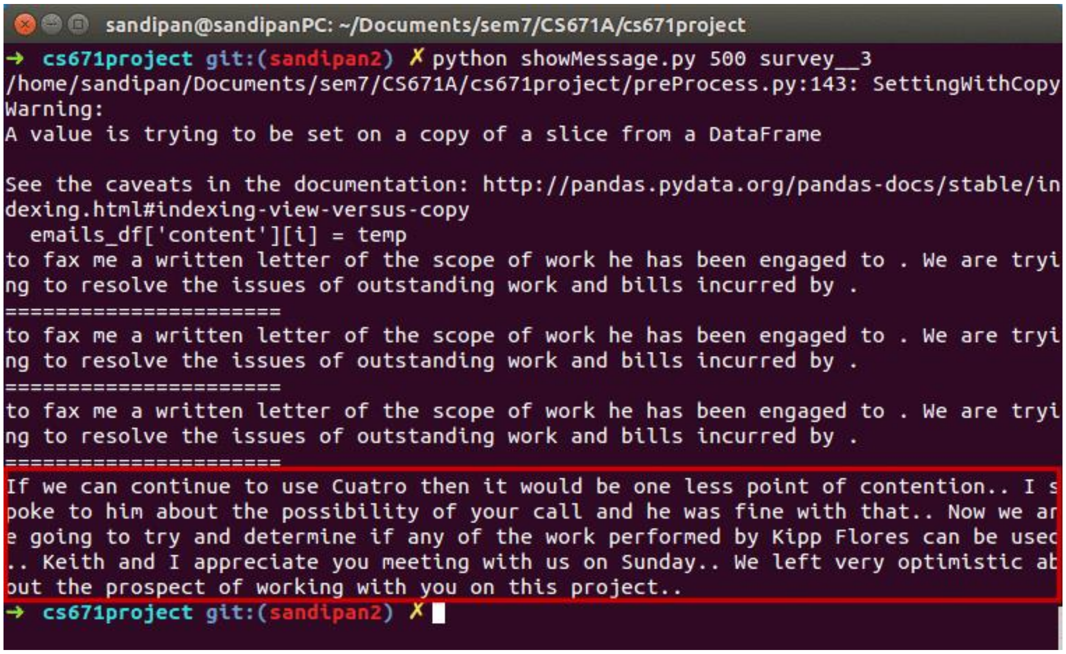
```
Re: 2- SURVEY - PHILLIP ALLEN TIME SENSITIVE: Executive Impact & Influence Program S
urvey . 2- SURVEY/INFORMATION EMAIL Re: 2- SURVEY - PHILLIP ALLEN 2- SURVEY/INFORMAT
ION EMAIL Behind the Stage Two Re: Enron's March Basdeload Fixed price physical deal
s as of 2/27.
======================
Re: 2- SURVEY - PHILLIP ALLEN TIME SENSITIVE: Executive Impact & Influence Program S
urvey . 2- SURVEY/INFORMATION EMAIL Re: 2- SURVEY - PHILLIP ALLEN 2- SURVEY/INFORMAT
ION EMAIL Behind the Stage Two Re: Enron's March Basdeload Fixed price physical deal
s as of 2/27.
======================
Re: 2- SURVEY - PHILLIP ALLEN TIME SENSITIVE: Executive Impact & Influence Program S
urvey . 2- SURVEY/INFORMATION EMAIL Re: 2- SURVEY - PHILLIP ALLEN 2- SURVEY/INFORMAT
ION EMAIL Behind the Stage Two Re: Enron's March Basdeload Fixed price physical deal
s as of 2/27.
======================
Investment Structure Investment Structure Re: 2- SURVEY - PHILLIP ALLEN TIME SENSITI
VE: Executive Impact & Influence Program Survey 2- SURVEY/INFORMATION EMAIL Re: 2- S
URVEY - PHILLIP ALLEN 2- SURVEY/INFORMATION EMAIL Behind the Stage Two Re: Enron's M
arch Basdeload Fixed price physical deals as of 2/27 /01 Attachment is free from vir
uses.. Here is the application from SPB.. I guess they want to use the same form as
a new loan application.. I was planning on bringing a copy of the survey and a rentr
oll including deposits on Monday.. Please let me know this morning what else I shoul
d be putting together..
➜ cs671project git:(sandipan2) ✗
```



```
A/Enron@ENRON, Frank Ermis/HOU/ECT@ECT, Jane M Tholt/HOU/ECT@ECT, Jay Reitm=. lin/HO
U/ECT@ECT, Mike Potter/NA/Enron@Enron, Natalie Baker/HOU/ECT@ECT, Suz=. ron@ENRON, L
am Nguyen/NA/Enron@Enron, Andy Pace/NA/Enron@Enron, Anna Santuc=. /NA/Enron@Enron, K
evin Brady/NA/Enron@Enron, Kirk Lenart/HOU/ECT@ECT, Lisa =. Tian/NA/Enron@ENRON, Fra
nk Ermis/HOU/ECT@ECT, Jane M Tholt/HOU/ECT@ECT, Jay .
======================
nXgate, Victor Guggenheim/HOU/ECT@ECT, Alejandra Chavez/NA/Enron@ENRON, Ann=. A/Enro
n@ENRON, Frank Ermis/HOU/ECT@ECT, Jane M Tholt/HOU/ECT@ECT, Jay Reitm=. H Fletcher/H
OU/ECT@ECT, Jeanne Wukasch/Corp/Enron@ENRON, Mary Theresa Frank=. lin/HOU/ECT@ECT, M
ike Potter/NA/Enron@Enron, Natalie Baker/HOU/ECT@ECT, Suz=. Tian/NA/Enron@ENRON, Fra
nk Ermis/HOU/ECT@ECT, Jane M Tholt/HOU/ECT@ECT, Jay .
======================
onXgate, Kevin Cline/Corp/Enron@Enron, Rika Imai/NA/Enron@Enron, Todd DeCoo=. h, sim
ply click on the 'Reply With History' button then hit 'Send'  Your su=. Please list
all Notes databases applications that you currently use: =20. Re: 2- SURVEY/INFORMAT
ION EMAIL 5-14- 01 Outlook Migration Team@ENRON. Tian/NA/Enron@ENRON, Frank Ermis/HO
U/ECT@ECT, Jane M Tholt/HOU/ECT@ECT, Jay .
======================
Please take a few minutes to completely fill out the followin= g survey.. Please lis
t all Notes databases applications that you currently use: =20. Please take a few mi
nutes to completely fill out the following survey.. Double Click on document to put
it in "Edit" mode.. Please list all Notes databases applications that you currently
use:.
➜ cs671project git:(sandipan2) ✗
```



```
See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/in
dexing.html#indexing-view-versus-copy
  emails_df['content'][i] = temp
I think we reached an agreement with George and Larry to pick up the items of . exte
nsions that they paid was applied to the purchase price of the land like . Keith and
 I are concerned about taking over the Bishop Corner partnership and .
======================
I think we reached an agreement with George and Larry to pick up the items of . exte
nsions that they paid was applied to the purchase price of the land like . Keith and
 I are concerned about taking over the Bishop Corner partnership and .
======================
I think we reached an agreement with George and Larry to pick up the items of . exte
nsions that they paid was applied to the purchase price of the land like . Keith and
 I are concerned about taking over the Bishop Corner partnership and .
======================
I think we reached an agreement with George and Larry to pick up the items of value
and not pay any fees for their time.. It looks as if we will be able to use everythi
ng they have done (engineering, architecture, survey, appraisal).. One point that is
 unclear is they claim that the $15,000 in extensions that they paid was applied to
the purchase price of the land like earnest money would be applied.. Keith and I are
 concerned about taking over the Bishop Corner partnership and the risk that there c
ould be undisclosed liabilities.. On the other hand, conveyance of the partnership w
ould be a time and money saver if it was clean..
→ cs671project git:(sandipan2) ✗ █
```



```
→ cs671project git:(sandipan2) ✗ python showMessage.py 500 survey__3
/home/sandipan/Documents/sem7/CS671A/cs671project/preProcess.py:143: SettingWithCopy
Warning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/in
dexing.html#indexing-view-versus-copy
  emails_df['content'][i] = temp
to fax me a written letter of the scope of work he has been engaged to . We are tryi
ng to resolve the issues of outstanding work and bills incurred by .
======================
to fax me a written letter of the scope of work he has been engaged to . We are tryi
ng to resolve the issues of outstanding work and bills incurred by .
======================
to fax me a written letter of the scope of work he has been engaged to . We are tryi
ng to resolve the issues of outstanding work and bills incurred by .
======================
If we can continue to use Cuatro then it would be one less point of contention.. I s
poke to him about the possibility of your call and he was fine with that.. Now we ar
e going to try and determine if any of the work performed by Kipp Flores can be usec
.. Keith and I appreciate you meeting with us on Sunday.. We left very optimistic ab
out the prospect of working with you on this project..
→ cs671project git:(sandipan2) ✗ █
```

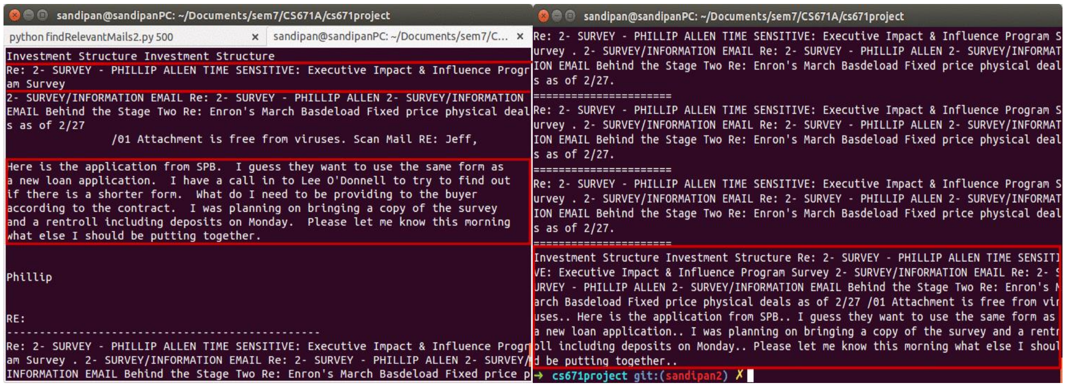
4.2.4 Summarization

We then executed summarizer on the cluster. We cleaned the emails for forwarded mails and other text irrelevant at this stage to come up with proper summary.

We used four summarizers for this purpose. The first three come as a part of the **pyTLDR** [8] package in python. All three of these implementations are extractive - that is, they simply extract and display the most relevant sentences from the input text. PyTLDR comes with a built-in sentence tokenizer that is used for summarization. The tokenizer performs stemming in several languages as well as stop-word removal. The last one comes as a part of the **sumy** [7] package.

- TextRankSummarizer [pyTLDR]

  This approach uses PageRank [5] algorithm, where "votes" or "in-links" are represented by words shared between sentences. The length of the summary can be specified either as a number of sentences, or a percentage of the original text.

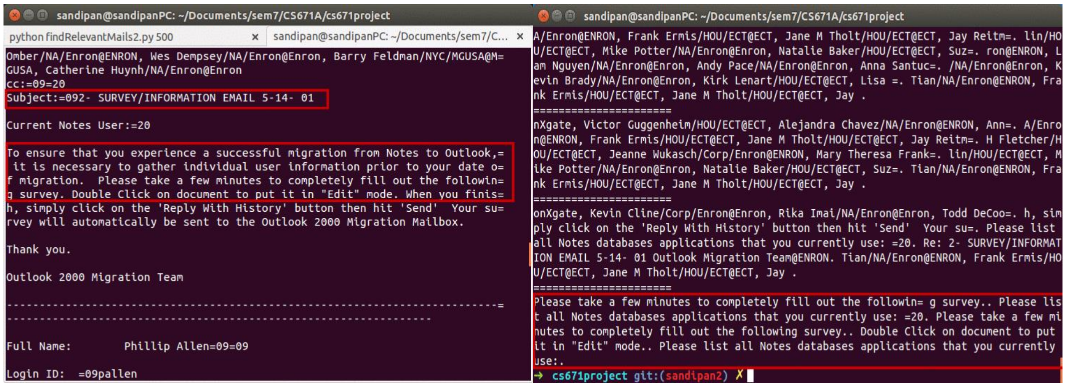

sandipan@sandipanPC: ~/Documents/sem7/CS671A/cs671project
python findRelevantMails2.py 500
sandipan@sandipanPC: ~/Documents/sem7/C...
sandipan@sandipanPC: ~/Documents/sem7/CS671A/cs671project

```
Investment Structure Investment Structure
Re: 2- SURVEY - PHILLIP ALLEN TIME SENSITIVE: Executive Impact & Influence Progr
am Survey
2- SURVEY/INFORMATION EMAIL Re: 2- SURVEY - PHILLIP ALLEN 2- SURVEY/INFORMATION
EMAIL Behind the Stage Two Re: Enron's March Basdeload Fixed price physical deal
s as of 2/27
                /01 Attachment is free from viruses. Scan Mail RE: Jeff,

Here is the application from SPB.  I guess they want to use the same form as
a new loan application.  I have a call in to Lee O'Donnell to try to find out
if there is a shorter form.  What do I need to be providing to the buyer
according to the contract.  I was planning on bringing a copy of the survey
and a rentroll including deposits on Monday.  Please let me know this morning
what else I should be putting together.


Phillip


RE:
------------------------------------------------
Re: 2- SURVEY - PHILLIP ALLEN TIME SENSITIVE: Executive Impact & Influence Progr
am Survey . 2- SURVEY/INFORMATION EMAIL Re: 2- SURVEY - PHILLIP ALLEN 2- SURVEY/
INFORMATION EMAIL Behind the Stage Two Re: Enron's March Basdeload Fixed price p
```

```
Re: 2- SURVEY - PHILLIP ALLEN TIME SENSITIVE: Executive Impact & Influence Program S
urvey . 2- SURVEY/INFORMATION EMAIL Re: 2- SURVEY - PHILLIP ALLEN 2- SURVEY/INFORMAT
ION EMAIL Behind the Stage Two Re: Enron's March Basdeload Fixed price physical deal
s as of 2/27.
======================
Re: 2- SURVEY - PHILLIP ALLEN TIME SENSITIVE: Executive Impact & Influence Program S
urvey . 2- SURVEY/INFORMATION EMAIL Re: 2- SURVEY - PHILLIP ALLEN 2- SURVEY/INFORMAT
ION EMAIL Behind the Stage Two Re: Enron's March Basdeload Fixed price physical deal
s as of 2/27.
======================
Re: 2- SURVEY - PHILLIP ALLEN TIME SENSITIVE: Executive Impact & Influence Program S
urvey . 2- SURVEY/INFORMATION EMAIL Re: 2- SURVEY - PHILLIP ALLEN 2- SURVEY/INFORMAT
ION EMAIL Behind the Stage Two Re: Enron's March Basdeload Fixed price physical deal
s as of 2/27.
======================
Investment Structure Investment Structure Re: 2- SURVEY - PHILLIP ALLEN TIME SENSITI
VE: Executive Impact & Influence Program Survey 2- SURVEY/INFORMATION EMAIL Re: 2- S
URVEY - PHILLIP ALLEN 2- SURVEY/INFORMATION EMAIL Behind the Stage Two Re: Enron's M
arch Basdeload Fixed price physical deals as of 2/27 /01 Attachment is free from vir
uses.. Here is the application from SPB.. I guess they want to use the same form as
a new loan application.. I was planning on bringing a copy of the survey and a rentr
oll including deposits on Monday.. Please let me know this morning what else I shoul
d be putting together..
➜ cs671project git:(sandipan2) ✗
```

- LsaSummarizer [pyTLDR]

  This method reduces the dimensionality of the article into several "topic" clusters using singular value decomposition, and selects the sentences that are most relevant to these topics. This is a rather more abstract summarization algorithm.

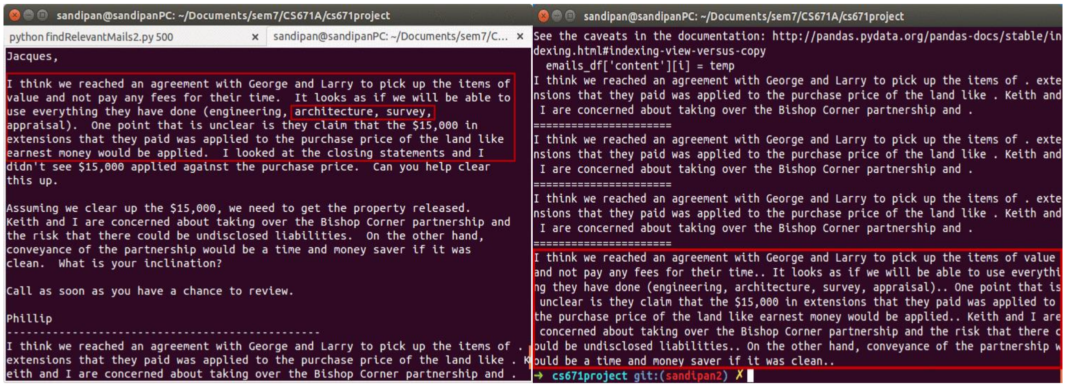

sandipan@sandipanPC: ~/Documents/sem7/CS671A/cs671project
python findRelevantMails2.py 500
sandipan@sandipanPC: ~/Documents/sem7/C...
sandipan@sandipanPC: ~/Documents/sem7/CS671A/cs671project

```
Omber/NA/Enron@ENRON, Wes Dempsey/NA/Enron@Enron, Barry Feldman/NYC/MGUSA@M=
GUSA, Catherine Huynh/NA/Enron@Enron
cc:=09=20
Subject:=092- SURVEY/INFORMATION EMAIL 5-14- 01

Current Notes User:=20

To ensure that you experience a successful migration from Notes to Outlook,=
 it is necessary to gather individual user information prior to your date o=
f migration.  Please take a few minutes to completely fill out the followin=
g survey. Double Click on document to put it in "Edit" mode. When you finis=
h, simply click on the 'Reply With History' button then hit 'Send'  Your su=
rvey will automatically be sent to the Outlook 2000 Migration Mailbox.

Thank you.

Outlook 2000 Migration Team

---------------------------------------------------------------------------=
-----------------------------------------------------------------


Full Name:        Phillip Allen=09=09

Login ID:  =09pallen
```

```
A/Enron@ENRON, Frank Ermis/HOU/ECT@ECT, Jane M Tholt/HOU/ECT@ECT, Jay Reitm=. lin/HO
U/ECT@ECT, Mike Potter/NA/Enron@Enron, Natalie Baker/HOU/ECT@ECT, Suz=. ron@ENRON, L
am Nguyen/NA/Enron@Enron, Andy Pace/NA/Enron@Enron, Anna Santuc=. /NA/Enron@Enron, K
evin Brady/NA/Enron@Enron, Kirk Lenart/HOU/ECT@ECT, Lisa =. Tian/NA/Enron@ENRON, Fra
nk Ermis/HOU/ECT@ECT, Jane M Tholt/HOU/ECT@ECT, Jay .
======================
nXgate, Victor Guggenheim/HOU/ECT@ECT, Alejandra Chavez/NA/Enron@ENRON, Ann=. A/Enro
n@ENRON, Frank Ermis/HOU/ECT@ECT, Jane M Tholt/HOU/ECT@ECT, Jay Reitm=. H Fletcher/H
OU/ECT@ECT, Jeanne Wukasch/Corp/Enron@ENRON, Mary Theresa Frank=. lin/HOU/ECT@ECT, M
ike Potter/NA/Enron@Enron, Natalie Baker/HOU/ECT@ECT, Suz=. Tian/NA/Enron@ENRON, Fra
nk Ermis/HOU/ECT@ECT, Jane M Tholt/HOU/ECT@ECT, Jay .
======================
onXgate, Kevin Cline/Corp/Enron@Enron, Rika Imai/NA/Enron@Enron, Todd DeCoo=. h, sim
ply click on the 'Reply With History' button then hit 'Send'  Your su=. Please list
all Notes databases applications that you currently use: =20. Re: 2- SURVEY/INFORMAT
ION EMAIL 5-14- 01 Outlook Migration Team@ENRON. Tian/NA/Enron@ENRON, Frank Ermis/HO
U/ECT@ECT, Jane M Tholt/HOU/ECT@ECT, Jay .
======================
Please take a few minutes to completely fill out the followin= g survey.. Please lis
t all Notes databases applications that you currently use: =20. Please take a few mi
nutes to completely fill out the following survey.. Double Click on document to put
it in "Edit" mode.. Please list all Notes databases applications that you currently
use:.
➜ cs671project git:(sandipan2) ✗
```

- RelevanceSummarizer [pyTLDR]

  This method computes and ranks the cosine similarity between each sentence vector and the overall document, removing the most relevant sentence at each iteration.

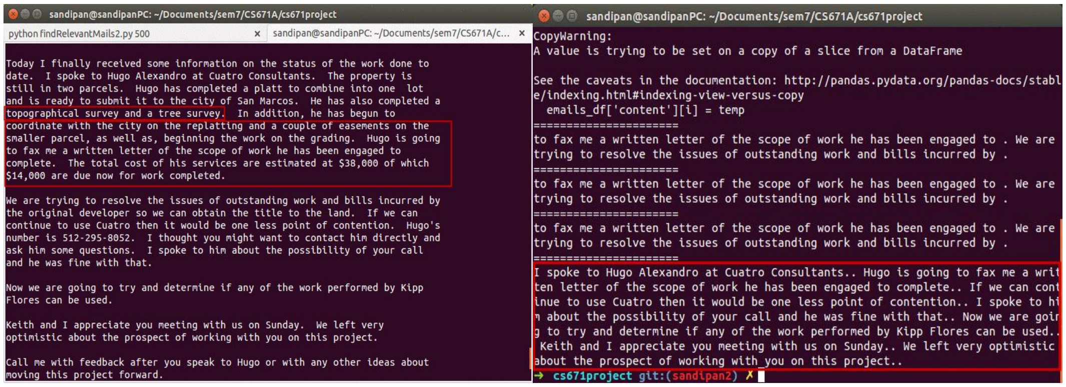

sandipan@sandipanPC: ~/Documents/sem7/CS671A/cs671project
python findRelevantMails2.py 500
sandipan@sandipanPC: ~/Documents/sem7/C...
sandipan@sandipanPC: ~/Documents/sem7/CS671A/cs671project

```
Jacques,

I think we reached an agreement with George and Larry to pick up the items of
value and not pay any fees for their time.  It looks as if we will be able to
use everything they have done (engineering, architecture, survey,
appraisal).  One point that is unclear is they claim that the $15,000 in
extensions that they paid was applied to the purchase price of the land like
earnest money would be applied.  I looked at the closing statements and I
didn't see $15,000 applied against the purchase price.  Can you help clear
this up.

Assuming we clear up the $15,000, we need to get the property released.
Keith and I are concerned about taking over the Bishop Corner partnership and
the risk that there could be undisclosed liabilities.  On the other hand,
conveyance of the partnership would be a time and money saver if it was
clean.  What is your inclination?

Call as soon as you have a chance to review.

Phillip
------------------------------------------------
I think we reached an agreement with George and Larry to pick up the items of .
extensions that they paid was applied to the purchase price of the land like . K
eith and I are concerned about taking over the Bishop Corner partnership and .
```

```
See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/in
dexing.html#indexing-view-versus-copy
  emails_df['content'][i] = temp
I think we reached an agreement with George and Larry to pick up the items of . exte
nsions that they paid was applied to the purchase price of the land like . Keith and
 I are concerned about taking over the Bishop Corner partnership and .
======================
I think we reached an agreement with George and Larry to pick up the items of . exte
nsions that they paid was applied to the purchase price of the land like . Keith and
 I are concerned about taking over the Bishop Corner partnership and .
======================
I think we reached an agreement with George and Larry to pick up the items of . exte
nsions that they paid was applied to the purchase price of the land like . Keith and
 I are concerned about taking over the Bishop Corner partnership and .
======================
I think we reached an agreement with George and Larry to pick up the items of value
and not pay any fees for their time.. It looks as if we will be able to use everythi
ng they have done (engineering, architecture, survey, appraisal).. One point that is
 unclear is they claim that the $15,000 in extensions that they paid was applied to
the purchase price of the land like earnest money would be applied.. Keith and I are
 concerned about taking over the Bishop Corner partnership and the risk that there c
ould be undisclosed liabilities.. On the other hand, conveyance of the partnership w
ould be a time and money saver if it was clean..
➜  cs671project git:(sandipan2) ✗
```

- LsaSummarizer [sumy]

  Algorithm is same as second but is taken from different library.


sandipan@sandipanPC: ~/Documents/sem7/CS671A/cs671project
python findRelevantMails2.py 500
sandipan@sandipanPC: ~/Documents/sem7/CS671A/c...
sandipan@sandipanPC: ~/Documents/sem7/CS671A/cs671project

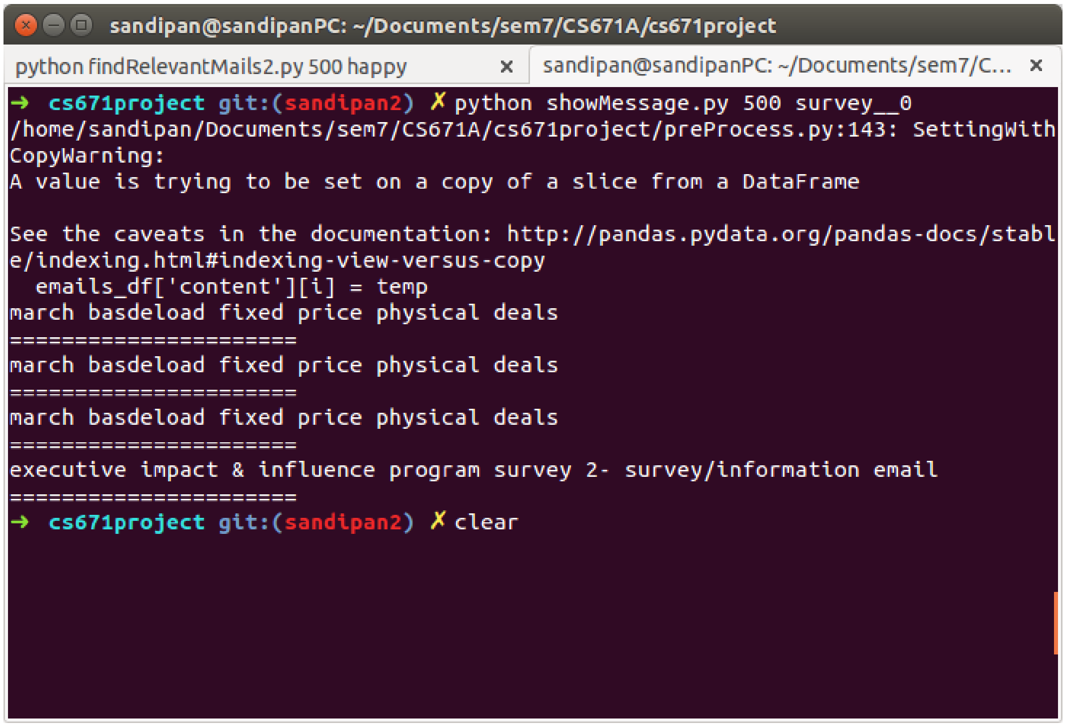
```
Today I finally received some information on the status of the work done to
date.  I spoke to Hugo Alexandro at Cuatro Consultants.  The property is
still in two parcels.  Hugo has completed a platt to combine into one  lot
and is ready to submit it to the city of San Marcos.  He has also completed a
topographical survey and a tree survey.  In addition, he has begun to
coordinate with the city on the replatting and a couple of easements on the
smaller parcel, as well as, beginning the work on the grading.  Hugo is going
to fax me a written letter of the scope of work he has been engaged to
complete.  The total cost of his services are estimated at $38,000 of which
$14,000 are due now for work completed.

We are trying to resolve the issues of outstanding work and bills incurred by
the original developer so we can obtain the title to the land.  If we can
continue to use Cuatro then it would be one less point of contention.  Hugo's
number is 512-295-8052.  I thought you might want to contact him directly and
ask him some questions.  I spoke to him about the possibility of your call
and he was fine with that.

Now we are going to try and determine if any of the work performed by Kipp
Flores can be used.

Keith and I appreciate you meeting with us on Sunday.  We left very
optimistic about the prospect of working with you on this project.

Call me with feedback after you speak to Hugo or with any other ideas about
moving this project forward.
```

```
CopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stabl
e/indexing.html#indexing-view-versus-copy
  emails_df['content'][i] = temp
======================
to fax me a written letter of the scope of work he has been engaged to . We are
trying to resolve the issues of outstanding work and bills incurred by .
======================
to fax me a written letter of the scope of work he has been engaged to . We are
trying to resolve the issues of outstanding work and bills incurred by .
======================
to fax me a written letter of the scope of work he has been engaged to . We are
trying to resolve the issues of outstanding work and bills incurred by .
======================
I spoke to Hugo Alexandro at Cuatro Consultants.. Hugo is going to fax me a writ
ten letter of the scope of work he has been engaged to complete.. If we can cont
inue to use Cuatro then it would be one less point of contention.. I spoke to hi
m about the possibility of your call and he was fine with that.. Now we are goin
g to try and determine if any of the work performed by Kipp Flores can be used..
 Keith and I appreciate you meeting with us on Sunday.. We left very optimistic
about the prospect of working with you on this project..
➜  cs671project git:(sandipan2) ✗
```

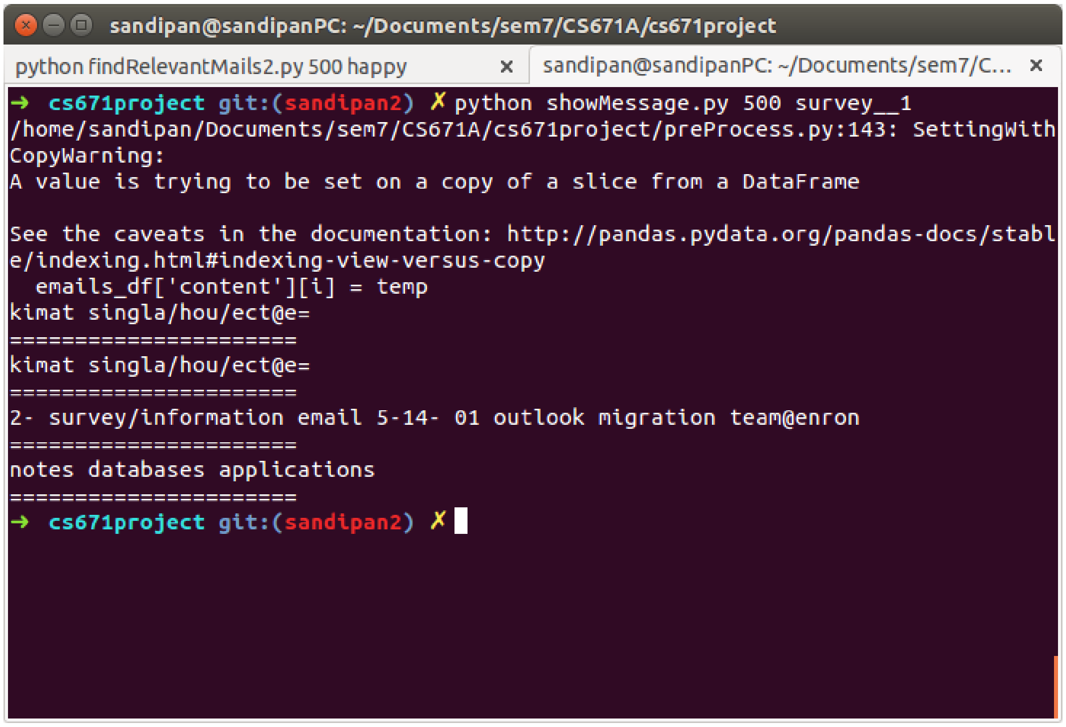
4.2.5 Topic Extraction

Next we extracted the most probable topic/keyphrase from the summaries corresponding to each cluster. We used 'RAKE' [6] library for this purpose. This was done to identify the clusters. We used the highest scoring keyphrase as topic for the cluster.

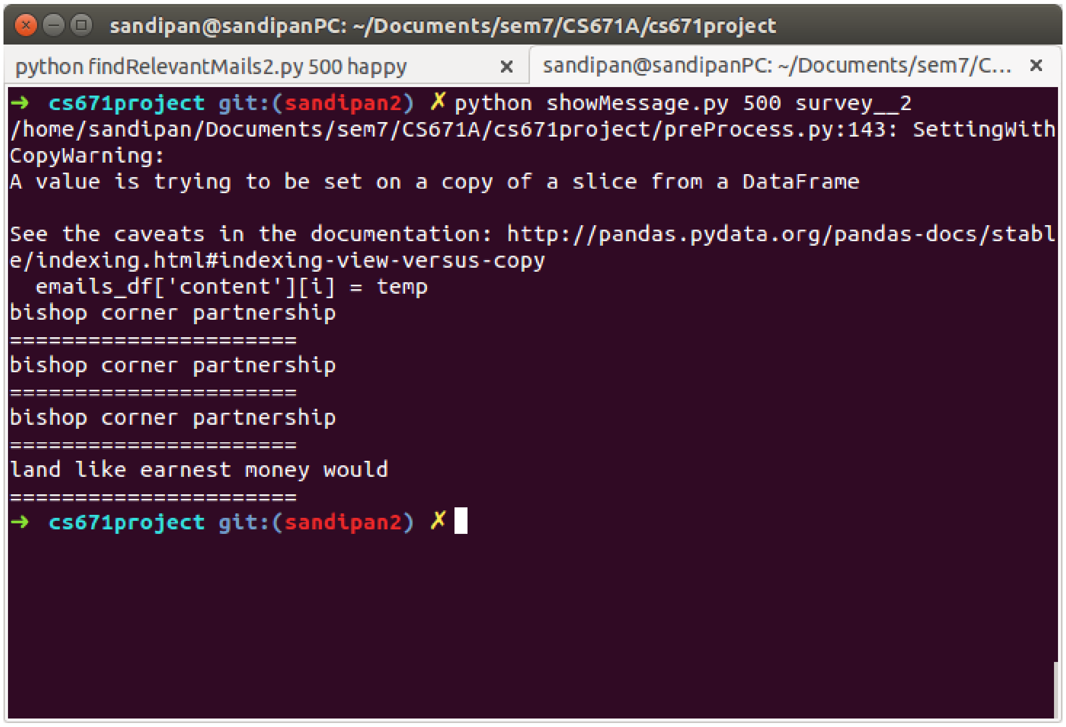
From the results, we observed that the topic extractor was not performing well. In many cases, the topic was unable to describe the content. It was happening because often mail contents spans over multiple small topics. Hence it becomes difficult to extract a single topic to describe the content.


sandipan@sandipanPC: ~/Documents/sem7/CS671A/cs671project
python findRelevantMails2.py 500 happy
sandipan@sandipanPC: ~/Documents/sem7/C...

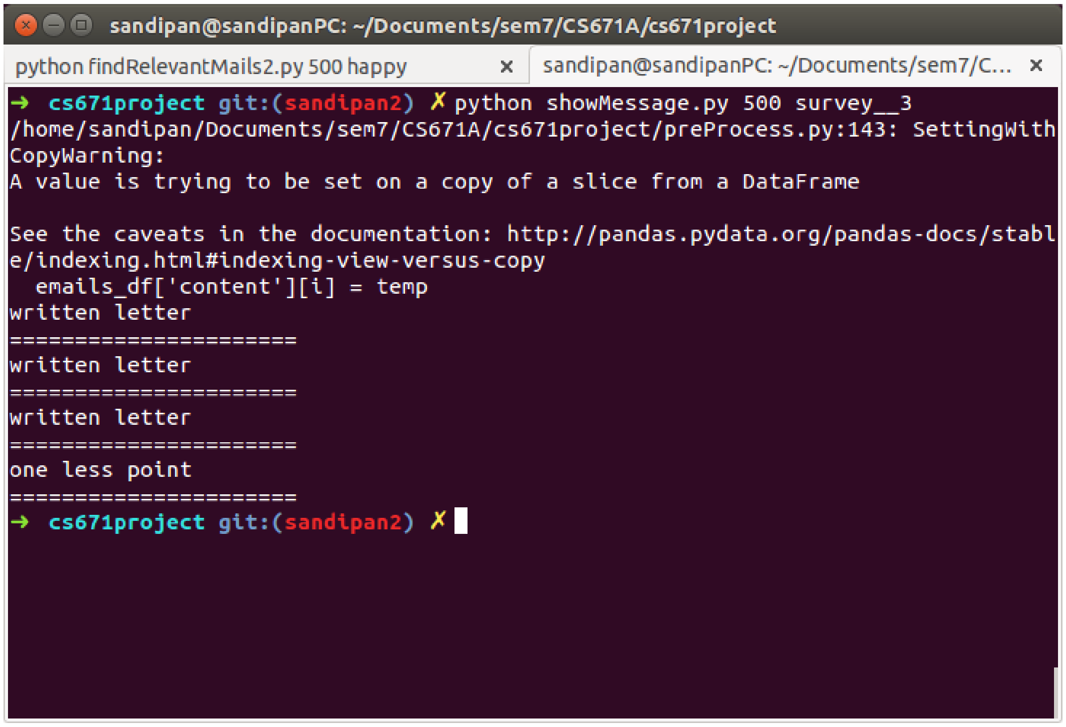
```
→ cs671project git:(sandipan2) ✗ python showMessage.py 500 survey__0
/home/sandipan/Documents/sem7/CS671A/cs671project/preProcess.py:143: SettingWith
CopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stabl
e/indexing.html#indexing-view-versus-copy
  emails_df['content'][i] = temp
march basdeload fixed price physical deals
======================
march basdeload fixed price physical deals
======================
march basdeload fixed price physical deals
======================
executive impact & influence program survey 2- survey/information email
======================
→ cs671project git:(sandipan2) ✗ clear
```


sandipan@sandipanPC: ~/Documents/sem7/CS671A/cs671project
python findRelevantMails2.py 500 happy
sandipan@sandipanPC: ~/Documents/sem7/C...

```
→ cs671project git:(sandipan2) ✗ python showMessage.py 500 survey__1
/home/sandipan/Documents/sem7/CS671A/cs671project/preProcess.py:143: SettingWith
CopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stabl
e/indexing.html#indexing-view-versus-copy
  emails_df['content'][i] = temp
kimat singla/hou/ect@e=
======================
kimat singla/hou/ect@e=
======================
2- survey/information email 5-14- 01 outlook migration team@enron
======================
notes databases applications
======================
→ cs671project git:(sandipan2) ✗
```


sandipan@sandipanPC: ~/Documents/sem7/CS671A/cs671project
python findRelevantMails2.py 500 happy
sandipan@sandipanPC: ~/Documents/sem7/C...

```
→ cs671project git:(sandipan2) ✗ python showMessage.py 500 survey__2
/home/sandipan/Documents/sem7/CS671A/cs671project/preProcess.py:143: SettingWith
CopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stabl
e/indexing.html#indexing-view-versus-copy
  emails_df['content'][i] = temp
bishop corner partnership
======================
bishop corner partnership
======================
bishop corner partnership
======================
land like earnest money would
======================
→ cs671project git:(sandipan2) ✗
```


sandipan@sandipanPC: ~/Documents/sem7/CS671A/cs671project
python findRelevantMails2.py 500 happy
sandipan@sandipanPC: ~/Documents/sem7/C...

```
→ cs671project git:(sandipan2) ✗ python showMessage.py 500 survey__3
/home/sandipan/Documents/sem7/CS671A/cs671project/preProcess.py:143: SettingWith
CopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stabl
e/indexing.html#indexing-view-versus-copy
  emails_df['content'][i] = temp
written letter
======================
written letter
======================
written letter
======================
one less point
======================
→ cs671project git:(sandipan2) ✗
```

Search Term	Number of relevant mail	Number of clusters	Purity
Survey	27	4	0.94
Deal	38	14	0.90
Project	21	7	0.90
Birthday	12	5	0.97
Meeting	51	17	0.92

Table 1: Performance of clustering algorithm for various search terms

Search Term	Algorithm	Important Points	Redundancy	Coherence
Survey	TextRank[pyTLDR]	1	1	1
Survey	LSA[pyTLDR]	1	1	1
Survey	RelevanceScore[pyTLDR]	1	1	1
Survey	LSA[sumy]	1	1	1
Deal	TextRank[pyTLDR]	0	1	1
Deal	LSA[pyTLDR]	0	1	0
Deal	RelevanceScore[pyTLDR]	0	1	0
Deal	LSA[sumy]	1	1	1
Project	TextRank[pyTLDR]	1	1	0
Project	LSA[pyTLDR]	1	1	0
Project	RelevanceScore[pyTLDR]	0	1	0
Project	LSA[sumy]	1	0	0

Table 2: Performance of Summarization algorithms for various search terms

5 Evaluation of Algorithm

5.1 Evaluation of Clustering Algorithm

We evaluated the performance of Clustering algorithm using **Purity**[3] of clusters.
Some results are tabulated in Table 1.

As we can see from Table 1, even though the number of clusters is quite small the **Purity** is very close to **1** in every case.
From this we can conclude that the clustering algorithm is performing pretty well in clustering the mails by different instances of the search term.

5.2 Evaluation of Summarization

We evaluated the summarization algorithms w.r.t to **inclusion of important points, non-redundancy and coherence**.
But our principal concern in this project is **inclusion of important points.**
We had 5 human annotators evaluate our summary on a scale of 0-1 (whether it is unacceptable or acceptable respectively) under each point. The majority opinion is tabulated in Table 2.

From the Table 2, if we consider the column corresponding to **Important Points**, we observe that LSA from sumy library has a score of 1 (Acceptable) for all the three search terms. Since our main concern was inclusion of important points, we can conclude that LSA algorithm from sumy is performing better than others in the sense that it is always capturing the main points from the cluster.

5.3 Evaluation of Topic Extraction

We did not explicitly evaluate the performance of Topic Extraction because we could see from the screenshots (of topic extraction), the performance is very poor.

6 Problems faced

- No general format of writing mails, depends on user to user.
- No systematic way to keep track of threads.
- Modifying subjects while continuing on the same thread - becomes difficult to keep track of the threads by the Subject.
- Lots of stuffs from various domains - becomes difficult to extract a unique topic from the mails.

7 Suggested Improvements

- Semantic analysis may be used to extract the main information asked for by the user.
- Some sort of data structure can be used to properly organize the data.
- All the pre-processing can be done before hand to efficiently reply to the queries.

References

[1] *BIRCH.* URL: https://en.wikipedia.org/wiki/BIRCH.

[2] Edwin Chen. *Introduction to Latent Dirichlet Allocation.* URL: http://blog.echen.me/2011/08/22/introduction-to-latent-dirichlet-allocation/.

[3] *Evaluation of clustering.* URL: http://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-clustering-1.html.

[4] *Gensim Doc2Vec.* URL: https://radimrehurek.com/gensim/models/doc2vec.html.

[5] Larry Page. *Page Rank Algorithm.* URL: https://en.wikipedia.org/wiki/PageRank.

[6] *Python RAKE.* URL: https://pypi.python.org/pypi/python-rake/1.0.5.

[7] *Python sumy.* URL: https://pypi.python.org/pypi/sumy.

[8] *Python TLDR.* URL: https://pypi.python.org/pypi/PyTLDR.