UNSUPERVISED EMAIL LABELLING

by:

Dibya Ranjan (10243)

Vishal Kumavat(10818)

Advisor:
Prof. Amit Mukerjee

CS365- Artificial Intelligence

Dept. of Computer Science & Engineering

IIT Kanpur

IMPORTANCE OF PROBLEM

- Emails have become the basic part of one's life.
- One practically receives 50-60 emails per week.
- Most of the times it become difficult to manage these emails and one may sometimes cannot find

PAST WORKS & APPLICATIONS

- **Real-Time Topic Modelling of Microblogs**
  - *by Yogesh Tewari and Rajesh Kawad*

- **Topic Modelling the Sarah Palin Emails**
  - *from Edwin Chen's blog*

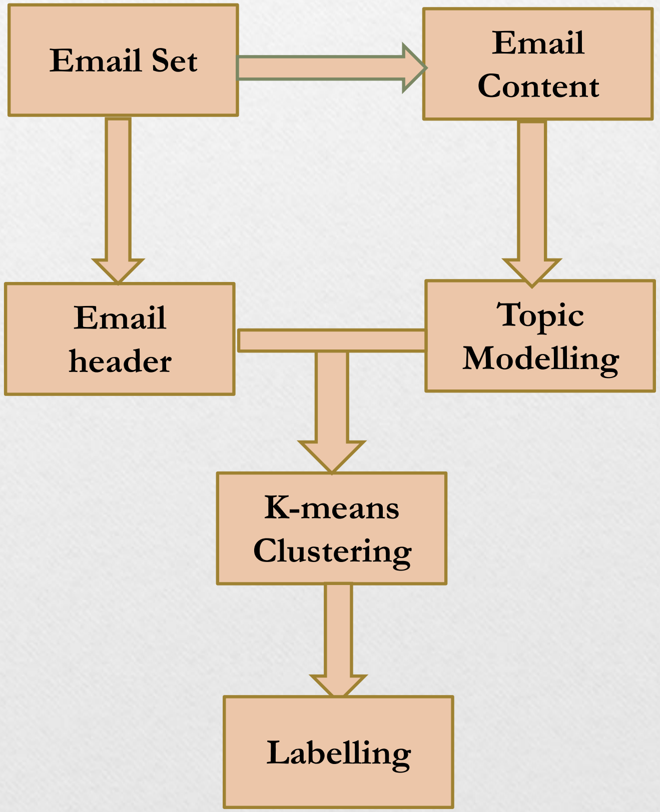
OUR APPROACH

1. Email is divided in two parts

- Email header
- Email content

2.Then topic modelling on email content is applied

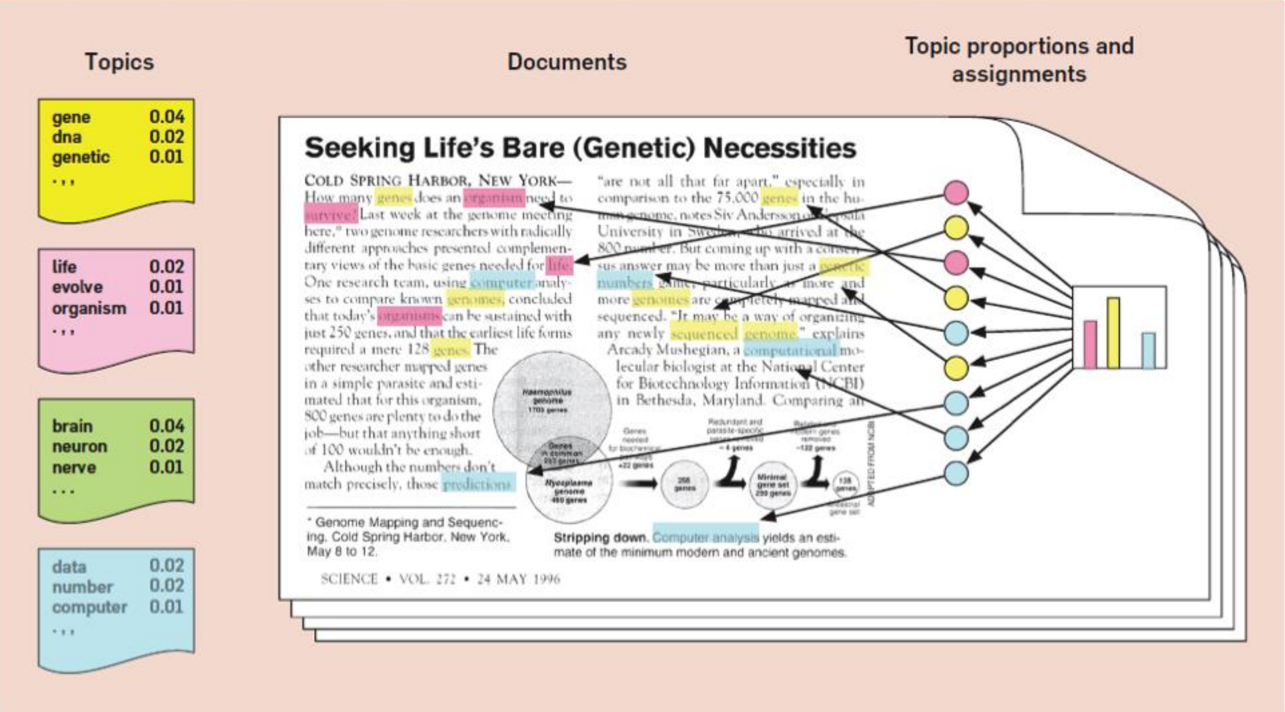
3. Combining the features of header and from topic modelling we apply K mean clustering



ALGORITHMS

Topic Modelling

Figure 1. The intuitions behind latent Dirichlet allocation. We assume that some number of "topics," which are distributions over words, exist for the whole collection (far left). Each document is assumed to be generated as follows. First choose a distribution over the topics (the histogram at right); then, for each word, choose a topic assignment (the colored coins) and choose the word from the corresponding topic. The topics and topic assignments in this figure are illustrative—they are not fit from real data. See Figure 2 for topics fit from data.


Topics
Documents
Topic proportions and assignments
gene 0.04
dna 0.02
genetic 0.01
...
life 0.02
evolve 0.01
organism 0.01
...
brain 0.04
neuron 0.02
nerve 0.01
...
data 0.02
number 0.02
computer 0.01
...
Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 468 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed −4 genes
Minimal gene set 250 genes
Related modern genes removed −122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

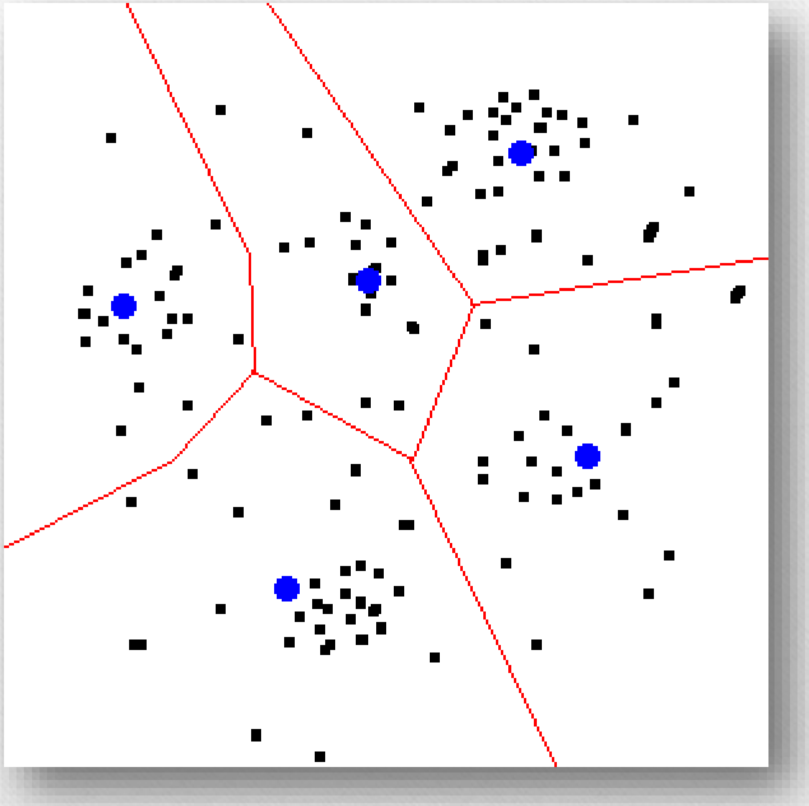
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

From: journalofdigitalhumanities.org/wp-content/uploads/2013/02/blei_lda_illustration.png

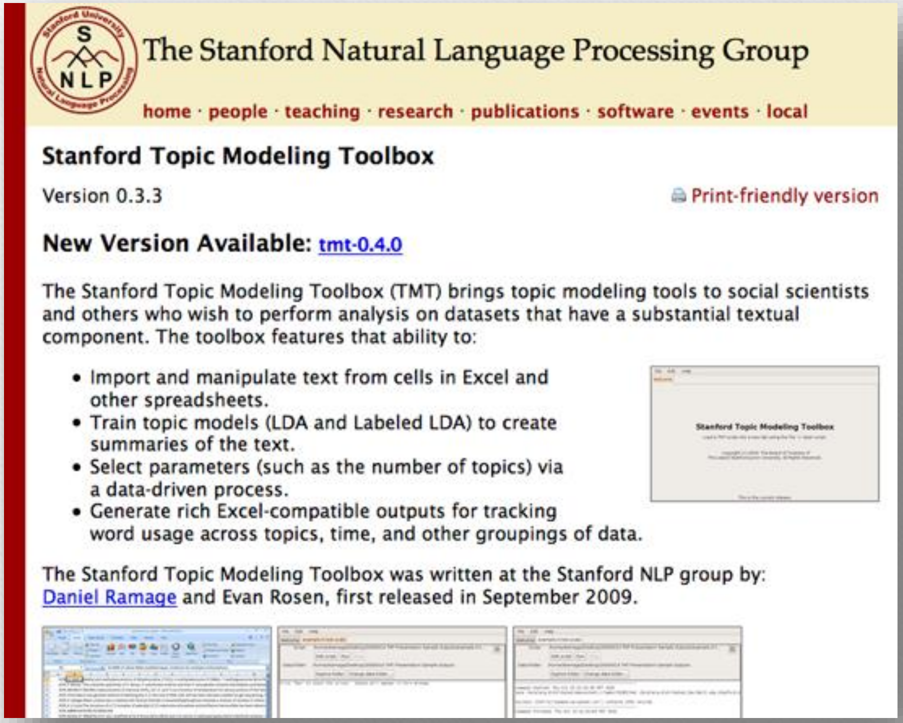
ALGORITHMS

K Means Clustering



TOOLS USED

- Topic Modelling Tool
  - ✓ It divides documents under different topics
  - ✓ It uses LDA(Latent Dirichlet Allocaton)



The Stanford Natural Language Processing Group
home · people · teaching · research · publications · software · events · local

Stanford Topic Modeling Toolbox

Version 0.3.3

New Version Available: tmt-0.4.0

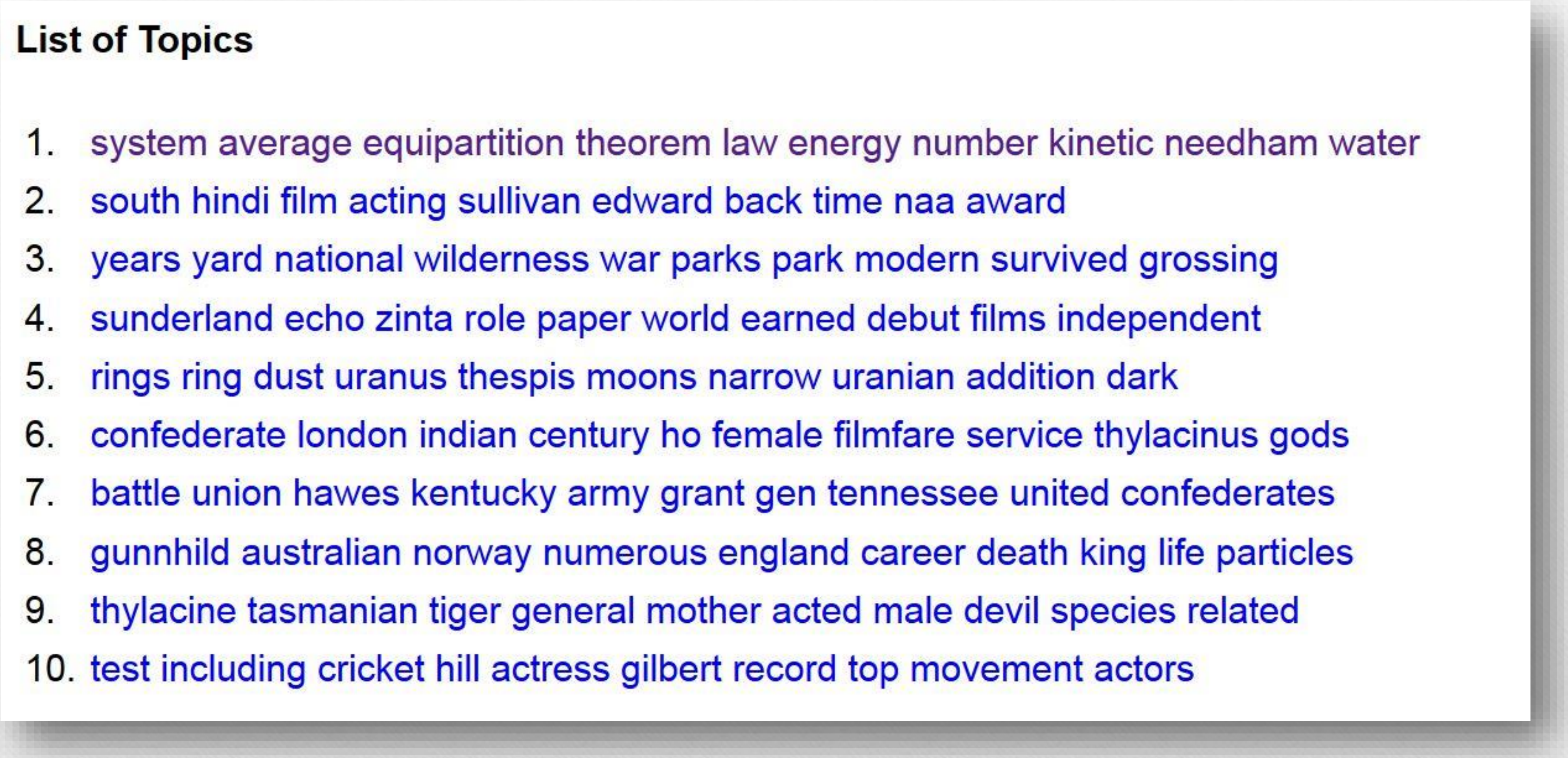
The Stanford Topic Modeling Toolbox (TMT) brings topic modeling tools to social scientists and others who wish to perform analysis on datasets that have a substantial textual component. The toolbox features that ability to:

- Import and manipulate text from cells in Excel and other spreadsheets.
- Train topic models (LDA and Labeled LDA) to create summaries of the text.
- Select parameters (such as the number of topics) via a data-driven process.
- Generate rich Excel-compatible outputs for tracking word usage across topics, time, and other groupings of data.

The Stanford Topic Modeling Toolbox was written at the Stanford NLP group by: Daniel Ramage and Evan Rosen, first released in September 2009.

From http://nlp.stanford.edu/software/tmt/tmt-0.4/

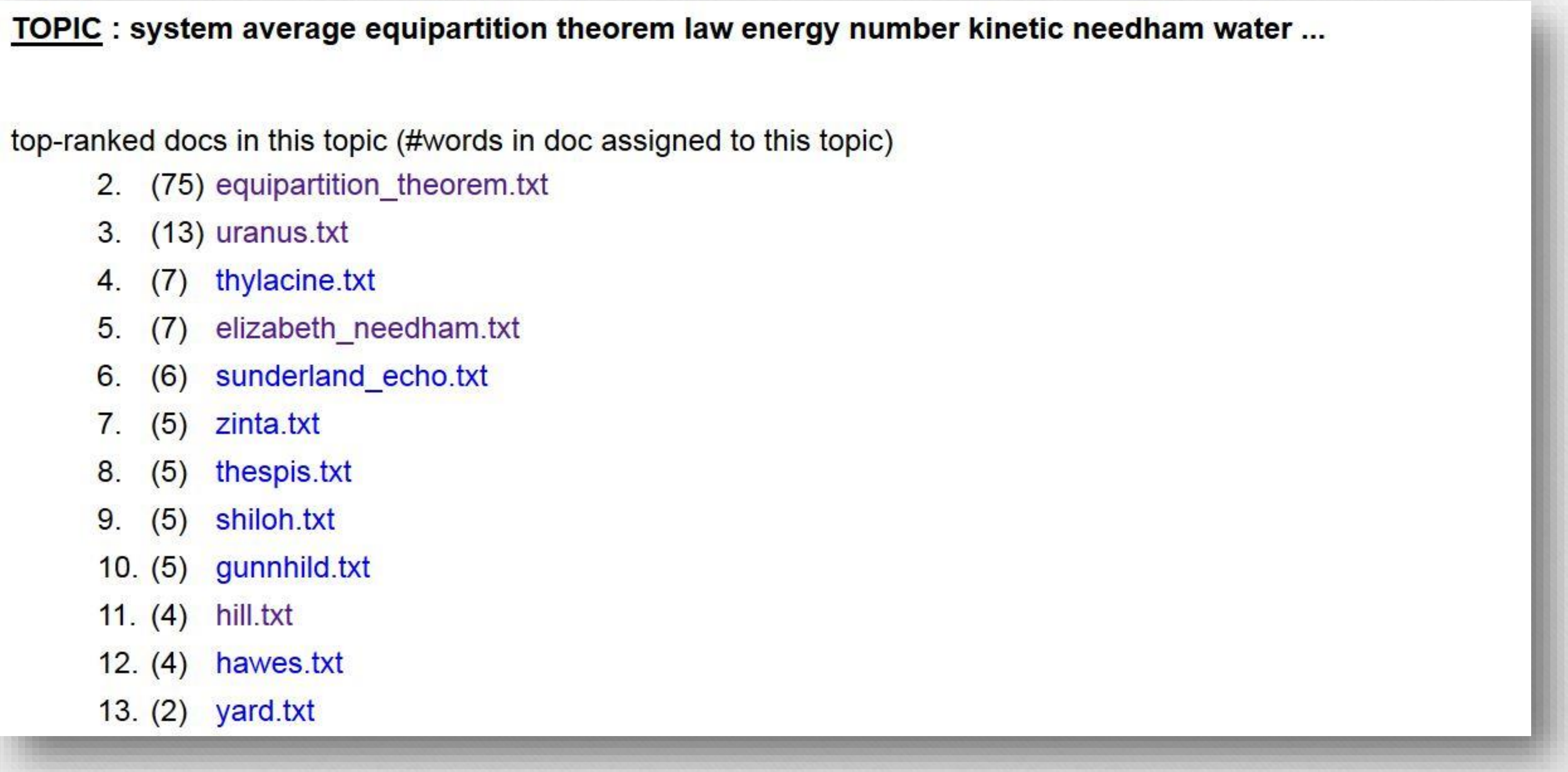
RESULTS



List of Topics

1. system average equipartition theorem law energy number kinetic needham water
2. south hindi film acting sullivan edward back time naa award
3. years yard national wilderness war parks park modern survived grossing
4. sunderland echo zinta role paper world earned debut films independent
5. rings ring dust uranus thespis moons narrow uranian addition dark
6. confederate london indian century ho female filmfare service thylacinus gods
7. battle union hawes kentucky army grant gen tennessee united confederates
8. gunnhild australian norway numerous england career death king life particles
9. thylacine tasmanian tiger general mother acted male devil species related
10. test including cricket hill actress gilbert record top movement actors

RESULTS

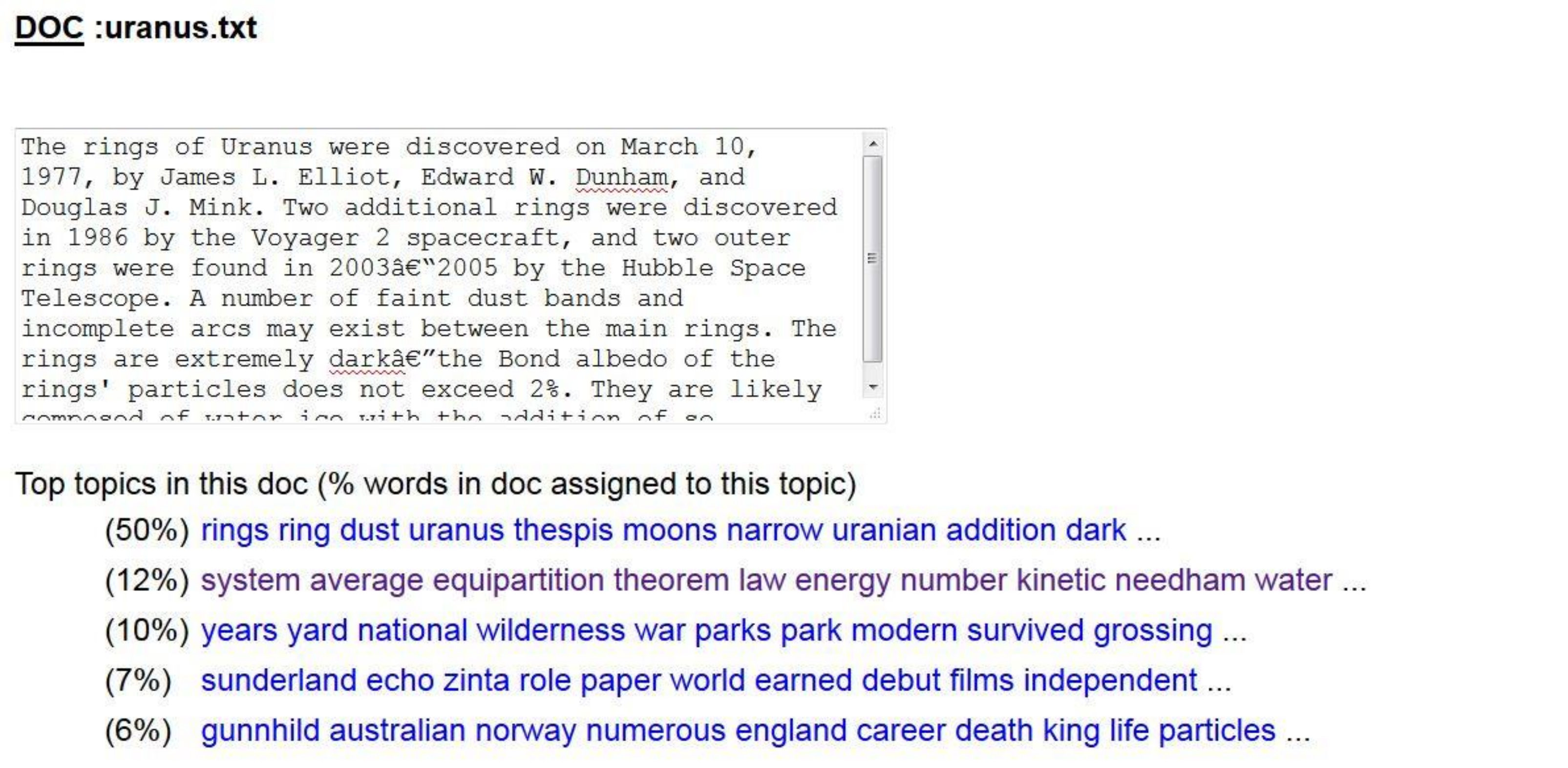


TOPIC : **system average equipartition theorem law energy number kinetic needham water ...**

top-ranked docs in this topic (#words in doc assigned to this topic)

2. (75) equipartition_theorem.txt
3. (13) uranus.txt
4. (7) thylacine.txt
5. (7) elizabeth_needham.txt
6. (6) sunderland_echo.txt
7. (5) zinta.txt
8. (5) thespis.txt
9. (5) shiloh.txt
10. (5) gunnhild.txt
11. (4) hill.txt
12. (4) hawes.txt
13. (2) yard.txt

RESULTS



DOC :uranus.txt

```
The rings of Uranus were discovered on March 10,
1977, by James L. Elliot, Edward W. Dunham, and
Douglas J. Mink. Two additional rings were discovered
in 1986 by the Voyager 2 spacecraft, and two outer
rings were found in 2003â€“2005 by the Hubble Space
Telescope. A number of faint dust bands and
incomplete arcs may exist between the main rings. The
rings are extremely darkâ€”the Bond albedo of the
rings' particles does not exceed 2%. They are likely
composed of water ice with the addition of so
```

Top topics in this doc (% words in doc assigned to this topic)

(50%) rings ring dust uranus thespis moons narrow uranian addition dark ...

(12%) system average equipartition theorem law energy number kinetic needham water ...

(10%) years yard national wilderness war parks park modern survived grossing ...

(7%) sunderland echo zinta role paper world earned debut films independent ...

(6%) gunnhild australian norway numerous england career death king life particles ...

DATASET

We have used the topic modelling on online available dataset on around 2000 emails

Created the data set of around 400-500 emails from personal inbox and used them for the unsupervised labelling

-

IMPROVEMENTS

We can also use more header information like subject and whether the email was sent to single or multiple receiver.

CONCLUSION

- Classified the emails based on balance between email header information and email semantics information.
- A Parameter was found which give priority to email header or email semantics

REFERENCES

- Mtech thesis on Email Classification Ozcaglar, Cagri. (2008)
- Topic Modelling theory-
  - http://clc.yale.edu/2011/10/07/how-to-do-your-own-topic-modeling/
  - http://www.fredgibbs.net/clio3workspace/blog/topic-modeling/
  - http://miriamposner.com/blog/?p=1335
  - http://blog.echen.me/2011/06/27/topic-modeling-the-sarah-palin-emails/
- Topic Modelling Tool-
  - http://nlp.stanford.edu/software/tmt/tmt-0.4/
- Dataset-
  - http://dhhumanist.org/Archives/Current/