UNSUPERWISED LABELLING OF EMAILS

By:
Vishal Kumawat 10818
Dibya Ranjan 10243

MOTIVATION

- Classify a large number of emails
- Labelling the email according to their semantics.
- Many time we have large number of documents like articles in Wikipedia , then how to classify these articles according to semantics????

KEY ALGORITHMS

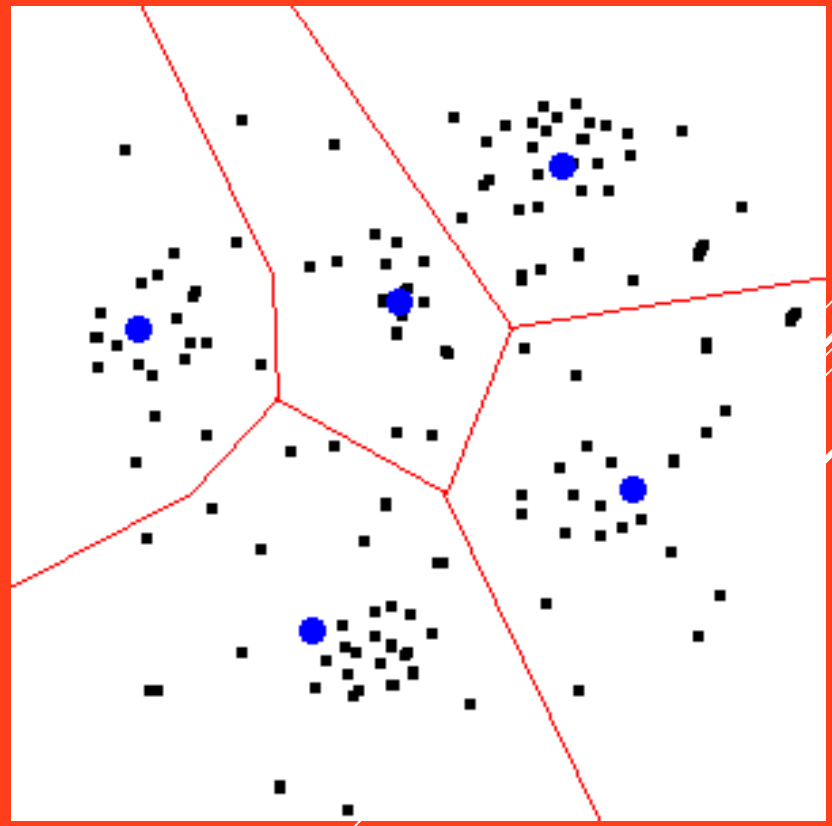
- Topic Modelling On Data
- K-mean Clustering.

TOPIC MODELLING

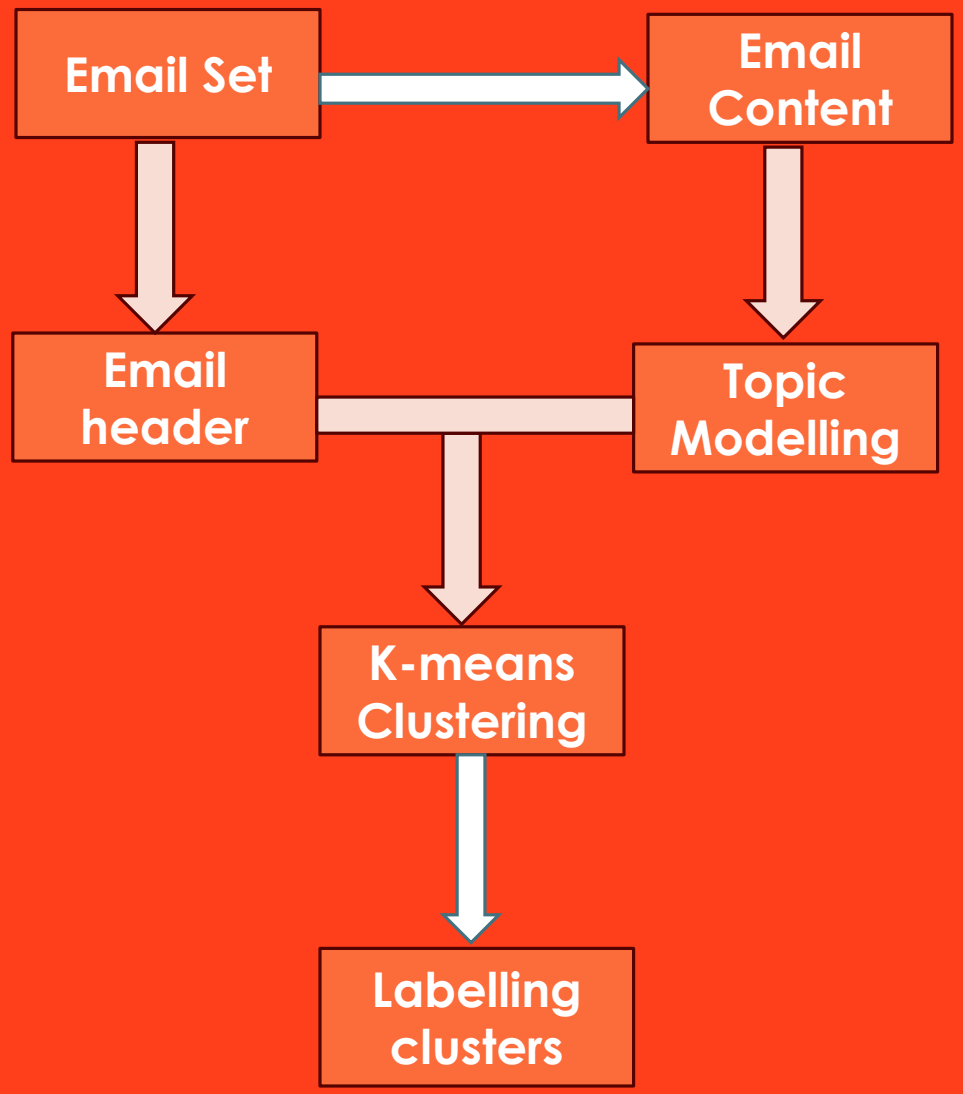
- Divide Each Document As Distribution Of Topics.
- Based On probabilistic model.
- Find Hidden Pattern In Data.

K –MEAN CLUSTERING

- one of the simplest unsupervised learning algorithms
- classify a given data set through a certain number of k clusters
- main idea is to define k centroids and better choice is to place them as far away as possible.
- take each point belonging to a given data set and associate it to the nearest centroid



OUR APPROACH



RESULT OF TOPIC MODELLING

List of topics in Test data set[R]

- system average equipartition theorem law energy number kinetic nedham water
- South hindi film acting Sullivan Edward back time naa award
- Years yard national wilderness war parks park modern survived grossing
- Sunderland echo zinta role paper world earned debut film independent
- Rings ring dust Uranus thespis moons narrow uranian addition dark
- Confederate London indian century ho filmfare service thylacinus gods
- Battle union hawes Kentucky army grant gen Tennessee united confederaters
- Gunnhild Australian Norway numerous England creer death king life particles
- Thyacine Tasmanian tiger general mother acted mail devil species related
- test including cricket hill actress gilbert record top movement actors

Ref of data set- http://dhhumanist.org/Archives/Current/

RESULT OF TOPIC MODELLING

Test Doc

Elizabeth Needham(died 3 May, 1731), also known as Mother Needham, was an English procuress and Brothel-keeper of greeting Moll Hackabout in the first plate of William Hogarth's series of satirical etchings, A Harlot's Progress. Although Needham was notorious in London at the time, little is recorded of her life, and no genuine potraits of her survive. Her house was the mst exclusive ijn London and her customers came from the highest stra…

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

- (20%) confederate london indian century ho female filmfare service thylacinus gods ...
- (13%) thylacine tasmanian tiger general mother acted male devil species related …
- (13%) system average equipartition theorem law energy number kinetic needham water ...
- (11%) gunnhild australian norway numerous england career death king life particles …
- (9%) rings ring dust uranus thespis moons narrow uranian addition dark …
- (9%) sunderland echo zinta role paper world earned debut films independent …
- (7%) test including cricket hill actress gilbert record top movement actors ...
- (7%) years yard national wilderness war parks park modern survived grossing …
- (6%) battle union hawes kentucky army grant gen tennessee united confederates …

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/

THANKS!!!

Questions