Topic Modelling using Latent Dirichlet Allocation

Students Involved: Akshat Karani 170010003
Hritik Kumar 170010013
Arpit Agrawal 170010040

Description

Topic Modeling is a type of statistical modeling for discovering the abstract “topics” that occur in a collection of documents. Latent Dirichlet Allocation(LDA) is an example of a topic model and is used to classify text in a document to a particular topic. In this project, we are going to apply LDA to a set of documents and split them into topics.

Topic Modelling of 20-newsgroups data using Gensim Model

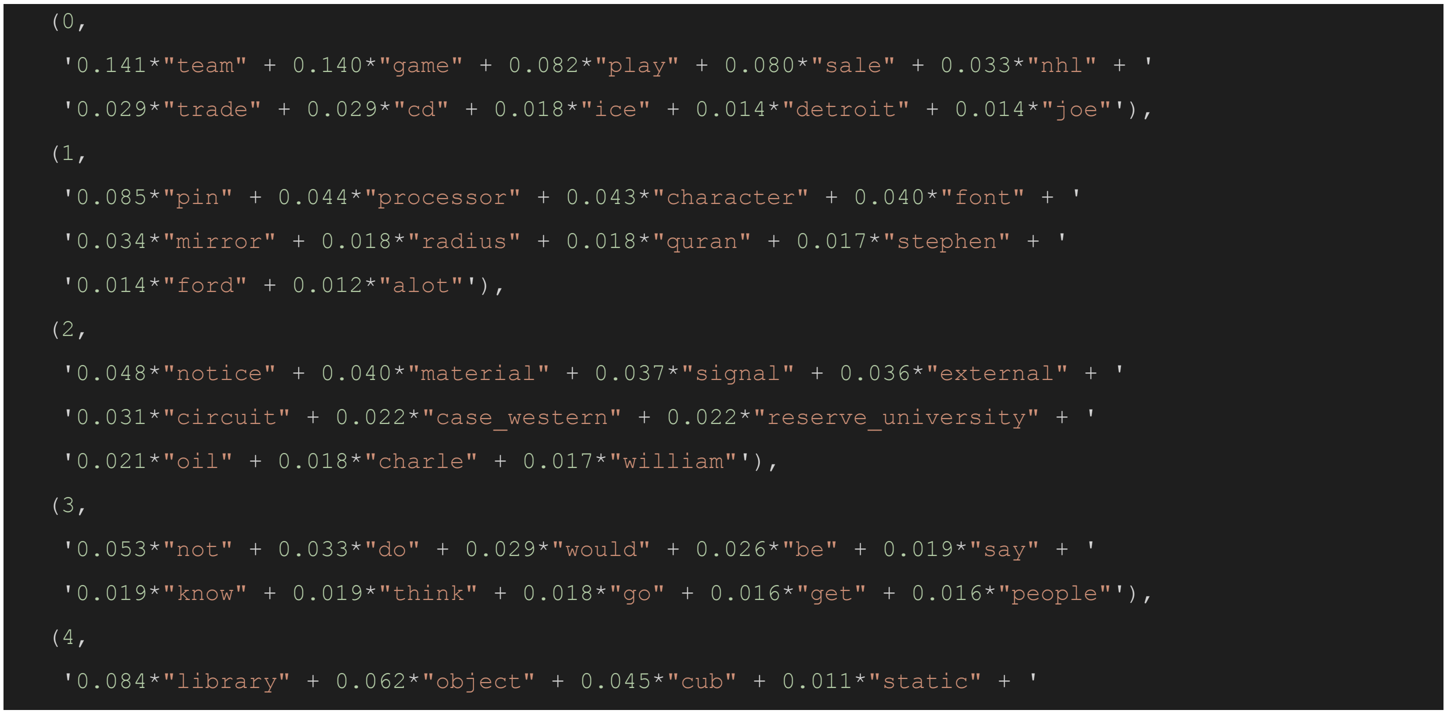
Dataset - *Link to dataset*

No. of documents in the corpus:

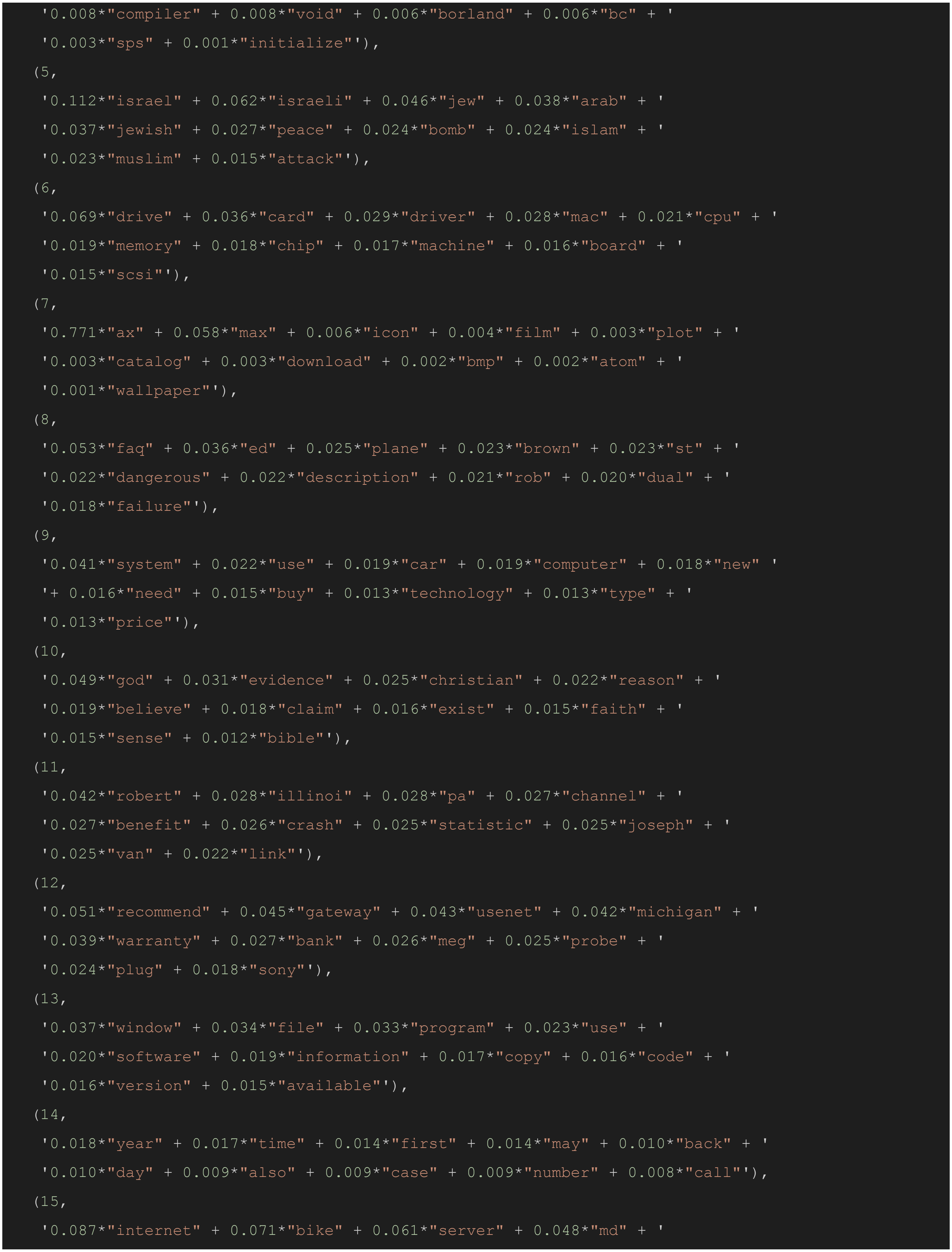
Total Number of Documents: 11314
Size of Vocabulary: 71002

Results Obtained

The newsgroups data is built with 20 different topics(the numbers from 0 to 19 represent the index of the topic) where each topic is a combination of keywords and the coefficient of each word represents the probability of the word occurring in that topic.



```
 (0,
  '0.141*"team" + 0.140*"game" + 0.082*"play" + 0.080*"sale" + 0.033*"nhl" + '
  '0.029*"trade" + 0.029*"cd" + 0.018*"ice" + 0.014*"detroit" + 0.014*"joe"'),
 (1,
  '0.085*"pin" + 0.044*"processor" + 0.043*"character" + 0.040*"font" + '
  '0.034*"mirror" + 0.018*"radius" + 0.018*"quran" + 0.017*"stephen" + '
  '0.014*"ford" + 0.012*"alot"'),
 (2,
  '0.048*"notice" + 0.040*"material" + 0.037*"signal" + 0.036*"external" + '
  '0.031*"circuit" + 0.022*"case_western" + 0.022*"reserve_university" + '
  '0.021*"oil" + 0.018*"charle" + 0.017*"william"'),
 (3,
  '0.053*"not" + 0.033*"do" + 0.029*"would" + 0.026*"be" + 0.019*"say" + '
  '0.019*"know" + 0.019*"think" + 0.018*"go" + 0.016*"get" + 0.016*"people"'),
 (4,
  '0.084*"library" + 0.062*"object" + 0.045*"cub" + 0.011*"static" + '
```



```
  '0.008*"compiler" + 0.008*"void" + 0.006*"borland" + 0.006*"bc" + '
  '0.003*"sps" + 0.001*"initialize"'),
 (5,
  '0.112*"israel" + 0.062*"israeli" + 0.046*"jew" + 0.038*"arab" + '
  '0.037*"jewish" + 0.027*"peace" + 0.024*"bomb" + 0.024*"islam" + '
  '0.023*"muslim" + 0.015*"attack"'),
 (6,
  '0.069*"drive" + 0.036*"card" + 0.029*"driver" + 0.028*"mac" + 0.021*"cpu" + '
  '0.019*"memory" + 0.018*"chip" + 0.017*"machine" + 0.016*"board" + '
  '0.015*"scsi"'),
 (7,
  '0.771*"ax" + 0.058*"max" + 0.006*"icon" + 0.004*"film" + 0.003*"plot" + '
  '0.003*"catalog" + 0.003*"download" + 0.002*"bmp" + 0.002*"atom" + '
  '0.001*"wallpaper"'),
 (8,
  '0.053*"faq" + 0.036*"ed" + 0.025*"plane" + 0.023*"brown" + 0.023*"st" + '
  '0.022*"dangerous" + 0.022*"description" + 0.021*"rob" + 0.020*"dual" + '
  '0.018*"failure"'),
 (9,
  '0.041*"system" + 0.022*"use" + 0.019*"car" + 0.019*"computer" + 0.018*"new" '
  '+ 0.016*"need" + 0.015*"buy" + 0.013*"technology" + 0.013*"type" + '
  '0.013*"price"'),
 (10,
  '0.049*"god" + 0.031*"evidence" + 0.025*"christian" + 0.022*"reason" + '
  '0.019*"believe" + 0.018*"claim" + 0.016*"exist" + 0.015*"faith" + '
  '0.015*"sense" + 0.012*"bible"'),
 (11,
  '0.042*"robert" + 0.028*"illinoi" + 0.028*"pa" + 0.027*"channel" + '
  '0.027*"benefit" + 0.026*"crash" + 0.025*"statistic" + 0.025*"joseph" + '
  '0.025*"van" + 0.022*"link"'),
 (12,
  '0.051*"recommend" + 0.045*"gateway" + 0.043*"usenet" + 0.042*"michigan" + '
  '0.039*"warranty" + 0.027*"bank" + 0.026*"meg" + 0.025*"probe" + '
  '0.024*"plug" + 0.018*"sony"'),
 (13,
  '0.037*"window" + 0.034*"file" + 0.033*"program" + 0.023*"use" + '
  '0.020*"software" + 0.019*"information" + 0.017*"copy" + 0.016*"code" + '
  '0.016*"version" + 0.015*"available"'),
 (14,
  '0.018*"year" + 0.017*"time" + 0.014*"first" + 0.014*"may" + 0.010*"back" + '
  '0.010*"day" + 0.009*"also" + 0.009*"case" + 0.009*"number" + 0.008*"call"'),
 (15,
  '0.087*"internet" + 0.071*"bike" + 0.061*"server" + 0.048*"md" + '
```

```
    '0.046*"engine" + 0.038*"ride" + 0.035*"route" + 0.031*"dod" + '
    '0.029*"announcement" + 0.021*"zone"'),
   (16,
    '0.019*"law" + 0.019*"state" + 0.018*"gun" + 0.017*"people" + '
    '0.017*"government" + 0.015*"kill" + 0.013*"death" + 0.012*"american" + '
    '0.010*"child" + 0.010*"country"'),
   (17,
    '0.089*"line" + 0.081*"organization" + 0.043*"write" + 0.040*"article" + '
    '0.033*"university" + 0.031*"host" + 0.019*"reply" + 0.018*"nntp_poste" + '
    '0.016*"thank" + 0.016*"nntp_posting"'),
   (18,
    '0.067*"space" + 0.019*"earth" + 0.018*"mount" + 0.017*"launch" + '
    '0.015*"moon" + 0.015*"research" + 0.015*"mission" + 0.014*"orbit" + '
    '0.014*"nasa" + 0.012*"satellite"'),
   (19,
    '0.148*"key" + 0.043*"encryption" + 0.036*"public" + 0.036*"security" + '
    '0.033*"chip" + 0.033*"clipper" + 0.026*"government" + 0.025*"secure" + '
    '0.023*"tap" + 0.021*"pgp"')
```

Result for a particular document

Contents of document

```
From: lerxst@wam.umd.edu (where's my thing)
Subject: WHAT car is this!?
Nntp-Posting-Host: rac3.wam.umd.edu
Organization: University of Maryland, College Park
Lines: 15
 I was wondering if anyone out there could enlighten me on this car I saw
the other day. It was a 2-door sports car, looked to be from the late 60s/
early 70s. It was called a Bricklin. The doors were really small. In addition,
the front bumper was separate from the rest of the body. This is
all I know. If anyone can tellme a model name, engine specs, years
of production, where this car is made, history, or whatever info you
have on this funky looking car, please e-mail.

Thanks,
- IL
```

Topic distribution of the above document (*as obtained using the model*)

```
(3, 0.28364155), (9, 0.12898049), (10, 0.04176736), (13, 0.011838715), (14, 0.14811127), (15,
0.017890228), (16, 0.013242425), (17, 0.32608366)
```

Below is the result of querying on an unseen document:

Content of the document

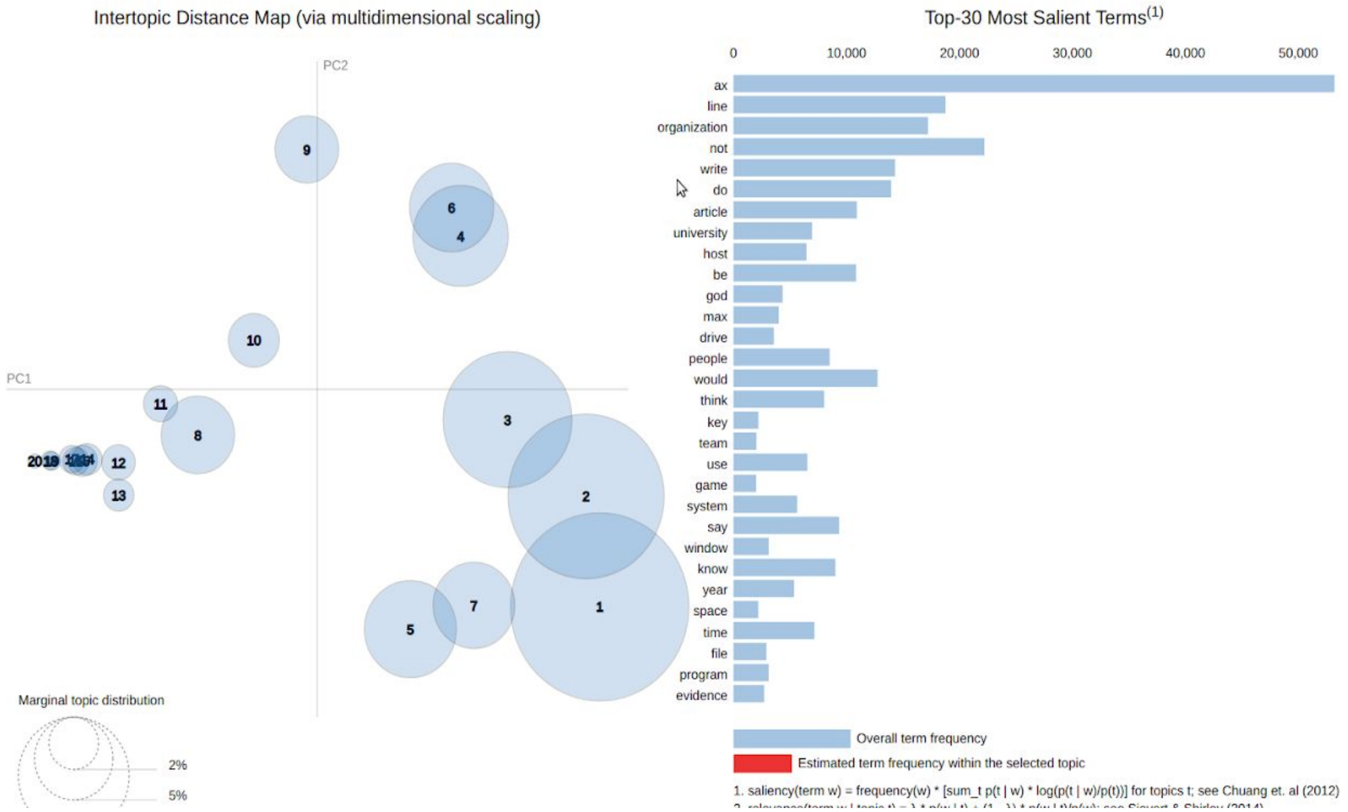
```
['computer', 'processor', 'graphic', 'network', 'jew', 'gun', 'politics']
```

Result Obtained (Topic Distribution)

```
(1, 0.051447194), (3, 0.18724184), (5, 0.053672783), (6, 0.027362216), (9, 0.15580104), (10,
0.03153057), (13, 0.047364414), (14, 0.13654219), (16, 0.07919185), (17, 0.17183337), (18,
0.012116399)
```

Topics visualized in 2-dimensions

The area of circle represents the importance of each topic over the entire corpus, the distance between the center of circles indicate the similarity between topics. For each topic, the histogram on the right side lists the top 30 most relevant terms.



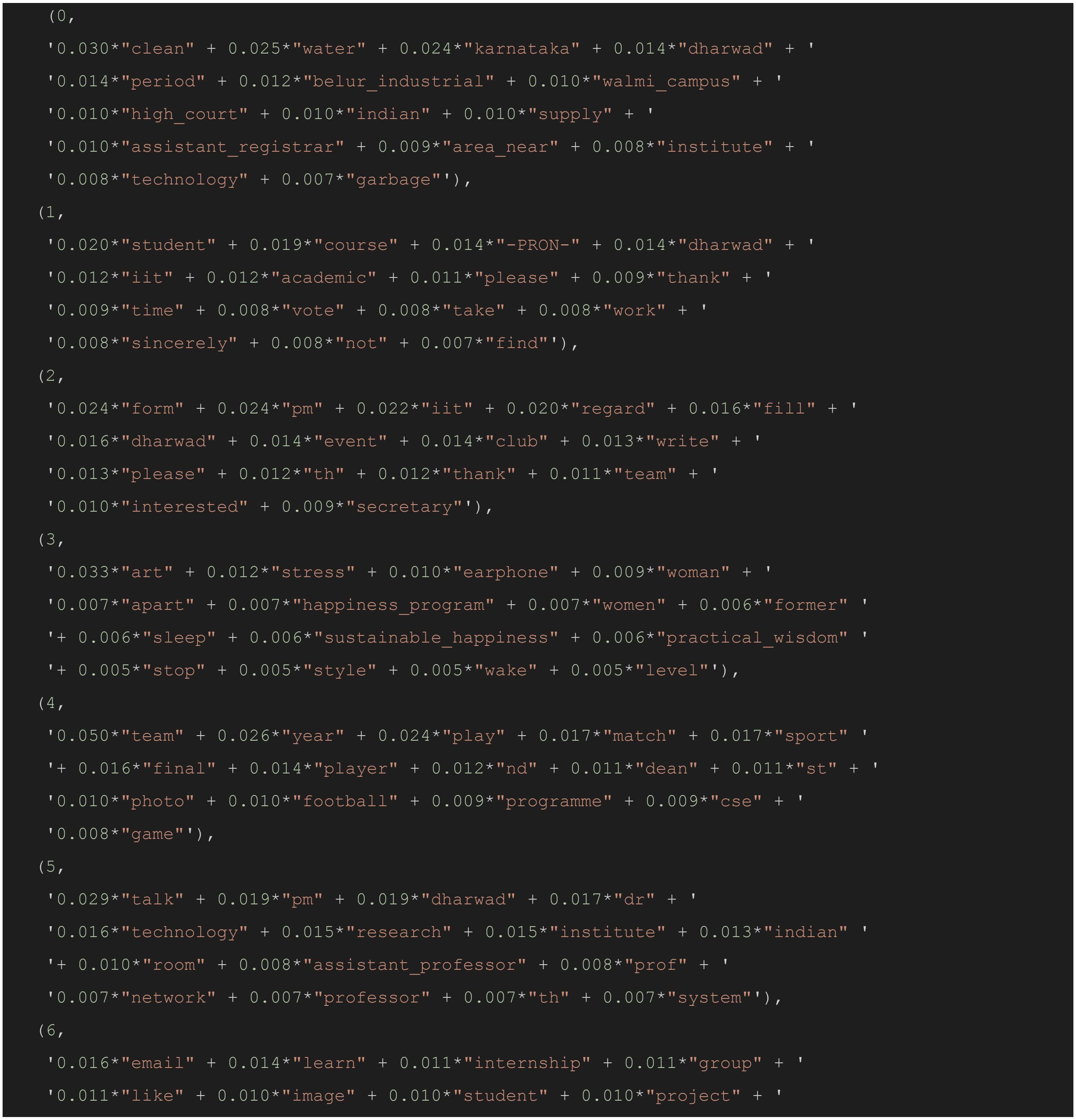
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)

Topic Modelling of IITDh Broadcast Email data using Gensim Model

No. of documents in the corpus:

Total Number of Documents: 1387
Size of Vocabulary: 9475

Results Obtained



```
 (0,
  '0.030*"clean" + 0.025*"water" + 0.024*"karnataka" + 0.014*"dharwad" + '
  '0.014*"period" + 0.012*"belur_industrial" + 0.010*"walmi_campus" + '
  '0.010*"high_court" + 0.010*"indian" + 0.010*"supply" + '
  '0.010*"assistant_registrar" + 0.009*"area_near" + 0.008*"institute" + '
  '0.008*"technology" + 0.007*"garbage"'),
 (1,
  '0.020*"student" + 0.019*"course" + 0.014*"-PRON-" + 0.014*"dharwad" + '
  '0.012*"iit" + 0.012*"academic" + 0.011*"please" + 0.009*"thank" + '
  '0.009*"time" + 0.008*"vote" + 0.008*"take" + 0.008*"work" + '
  '0.008*"sincerely" + 0.008*"not" + 0.007*"find"'),
 (2,
  '0.024*"form" + 0.024*"pm" + 0.022*"iit" + 0.020*"regard" + 0.016*"fill" + '
  '0.016*"dharwad" + 0.014*"event" + 0.014*"club" + 0.013*"write" + '
  '0.013*"please" + 0.012*"th" + 0.012*"thank" + 0.011*"team" + '
  '0.010*"interested" + 0.009*"secretary"'),
 (3,
  '0.033*"art" + 0.012*"stress" + 0.010*"earphone" + 0.009*"woman" + '
  '0.007*"apart" + 0.007*"happiness_program" + 0.007*"women" + 0.006*"former" '
  '+ 0.006*"sleep" + 0.006*"sustainable_happiness" + 0.006*"practical_wisdom" '
  '+ 0.005*"stop" + 0.005*"style" + 0.005*"wake" + 0.005*"level"'),
 (4,
  '0.050*"team" + 0.026*"year" + 0.024*"play" + 0.017*"match" + 0.017*"sport" '
  '+ 0.016*"final" + 0.014*"player" + 0.012*"nd" + 0.011*"dean" + 0.011*"st" + '
  '0.010*"photo" + 0.010*"football" + 0.009*"programme" + 0.009*"cse" + '
  '0.008*"game"'),
 (5,
  '0.029*"talk" + 0.019*"pm" + 0.019*"dharwad" + 0.017*"dr" + '
  '0.016*"technology" + 0.015*"research" + 0.015*"institute" + 0.013*"indian" '
  '+ 0.010*"room" + 0.008*"assistant_professor" + 0.008*"prof" + '
  '0.007*"network" + 0.007*"professor" + 0.007*"th" + 0.007*"system"'),
 (6,
  '0.016*"email" + 0.014*"learn" + 0.011*"internship" + 0.011*"group" + '
  '0.011*"like" + 0.010*"image" + 0.010*"student" + 0.010*"project" + '
```

```
    '0.009*"share" + 0.008*"get" + 0.008*"would" + 0.007*"use" + '
    '0.007*"password" + 0.007*"view" + 0.007*"may"'),
   (7,
    '0.038*"guy" + 0.029*"gen_secy" + 0.029*"ashrith_adepu" + 0.022*"regard" + '
    '0.021*"pm" + 0.021*"app" + 0.021*"hostel_affairs" + 0.020*"quiz" + '
    '0.020*"hostel" + 0.017*"hostel_affair" + 0.016*"campus" + 0.014*"please" + '
    '0.014*"post" + 0.010*"write" + 0.010*"general_secretary"'),
   (8,
    '0.014*"student" + 0.011*"india" + 0.010*"iit" + 0.009*"date" + 0.009*"th" + '
    '0.008*"please" + 0.007*"technology" + 0.007*"program" + 0.006*"workshop" + '
    '0.006*"information" + 0.006*"design" + 0.006*"opportunity" + 0.006*"year" + '
    '0.006*"regard" + 0.006*"dharwad"')
```

Result for a particular document

Content of a document

```
Subject: Fluid Mechanics Research Work on Musical Instruments
Text: Hello folks,

Ever wondered how 'Flute' or 'Jal-Tarang' create such melodious sounds? If
these phenomena fascinate you and if you are looking forward to get to know
the Science behind it, grab this opportunity, register here
<https://goo.gl/forms/GFi8pOQkpKOZvBOT2> and begin your research work on
Fluid Mechanics/Acoustics/Physics/Digital Signal Processing under the
valuable guidance of *Prof. Dhiraj Patil* and *Prof. Mahadeva Prasanna*.

Yours truly,
Saurav Dosi


--
Thanks and Regards,
Saurav Dosi
Institute Music Secretary
IIT Dharwad
+919819582916
```

Topic distribution of the above document (*as obtained using the model*)

```
(1, 0.066245176), (2, 0.19156684), (4, 0.44773257), (5, 0.17656894), (8, 0.10472268)
```

Below is the result of querying on an unseen document:

Content of the document

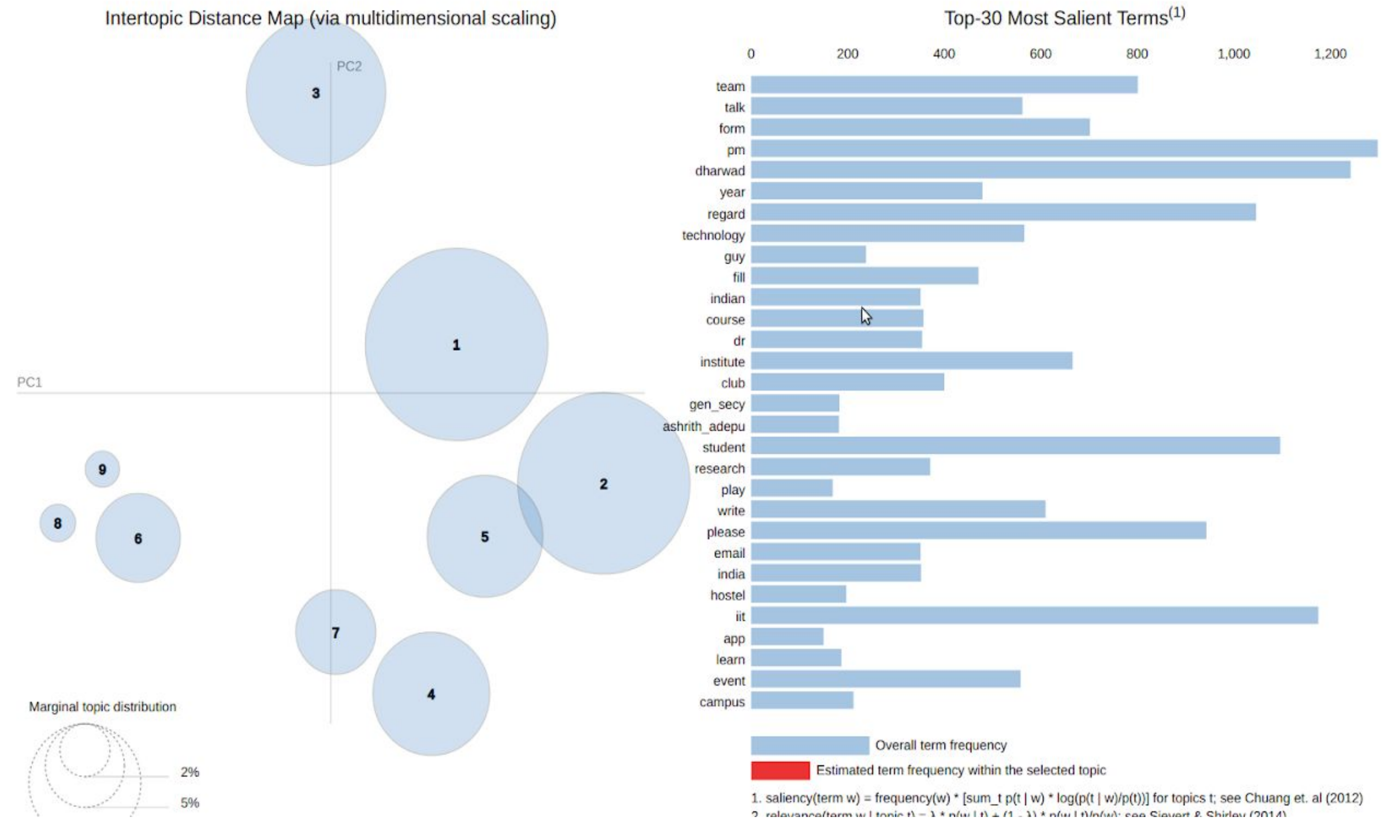
```
['event', 'talk', 'invited', 'guest', 'professor', 'snacks', 'gentle']
```

Result Obtained (Topic Distribution)

```
(1, 0.04725744), (2, 0.22576483), (4, 0.185834), (5, 0.21489222), (6, 0.04462768), (7, 0.023874043), (8, 0.24777623)
```

Topics visualized in 2-dimensions

The area of circle represents the importance of each topic over the entire corpus, the distance between the center of circles indicate the similarity between topics. For each topic, the histogram on the right side lists the top 30 most relevant terms.


Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
3
1
2
5
9
8
6
7
4
Marginal topic distribution
2%
5%
Top-30 Most Salient Terms(1)
0
200
400
600
800
1,000
1,200
team
talk
form
pm
dharwad
year
regard
technology
guy
fill
indian
course
dr
institute
club
gen_secy
ashrith_adepu
student
research
play
write
please
email
india
hostel
iit
app
learn
event
campus
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Future Plan:

- Improve the result obtained on 20-newsgroups data and IITDh broadcast mail by better preprocessing of data which includes removal of stop words and most frequently occurring words.
- Add the ability to retrieve documents for a query.
- Explore other use cases of LDA like using LDA on images.
- Get a better understanding of LDA.
- Come up with better visualization methods.