Project 18-Topic Modelling

By-Team 18
Akhil Singh, Ishan Bansal,
Vikrant Goyal, Paawan Gupta

Supervised Learning

- Input is a pair consisting of an input object and a desired output value.

- It learns a function to map input object to the desired output value which can be used in predicting the output values for new input object.

- Examples: Based on past information about spams, filtering out a new incoming email into Inbox (normal) or Junk folder (Spam), classification of objects etc.

Unsupervised Learning

- Unsupervised learning studies how systems can learn to represent particular input patterns in a way that reflects the statistical structure of the overall collection of input patterns.
- There are no explicit target outputs rather the unsupervised learner brings to bear prior biases as to what aspects of the structure of the input should be captured in the output.
- It is often easier to obtain unlabeled data from a lab instrument or a computer than labeled data, which can require human intervention.
- Examples: Clustering, PCA etc.

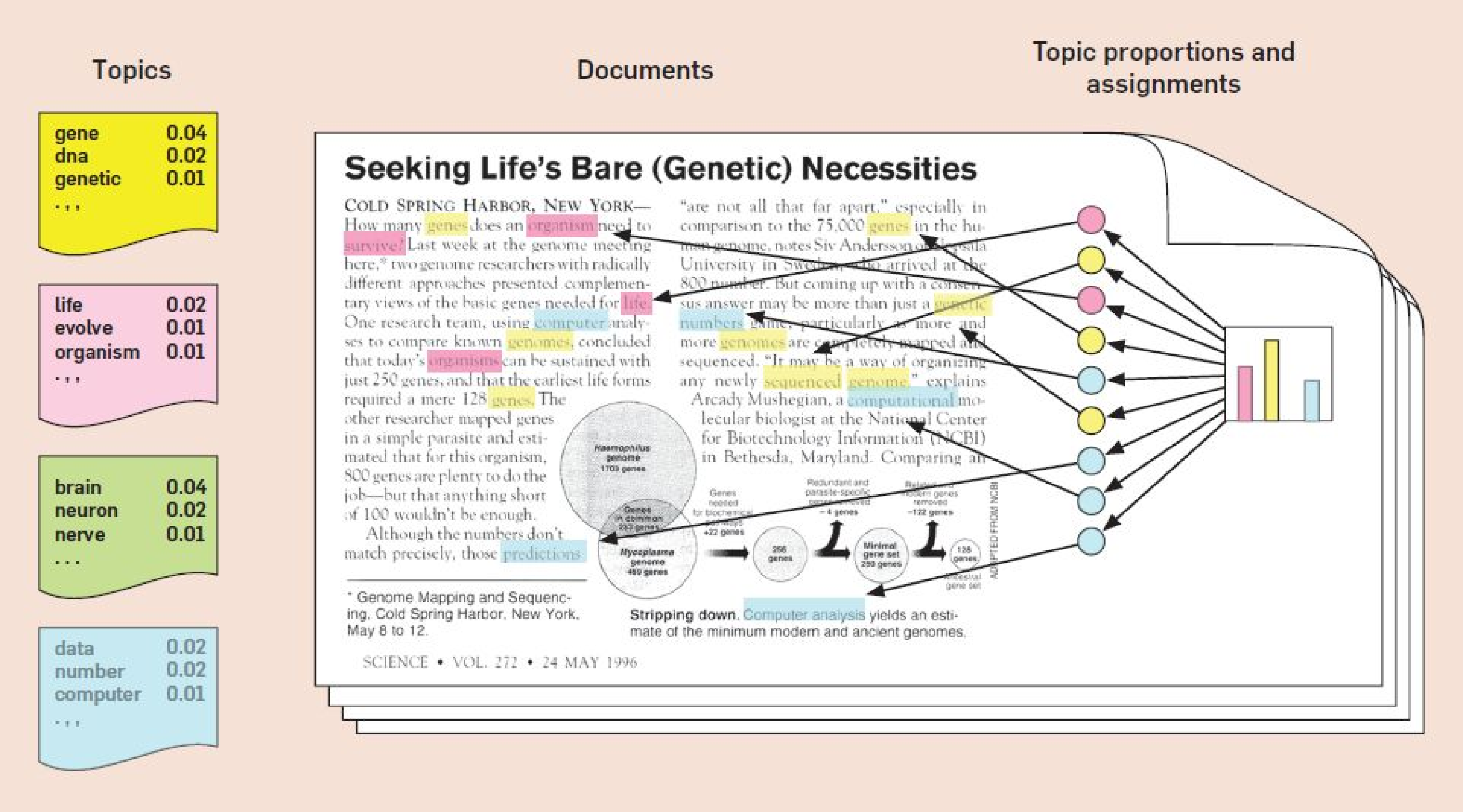
Topic Modelling

❏ Motivation

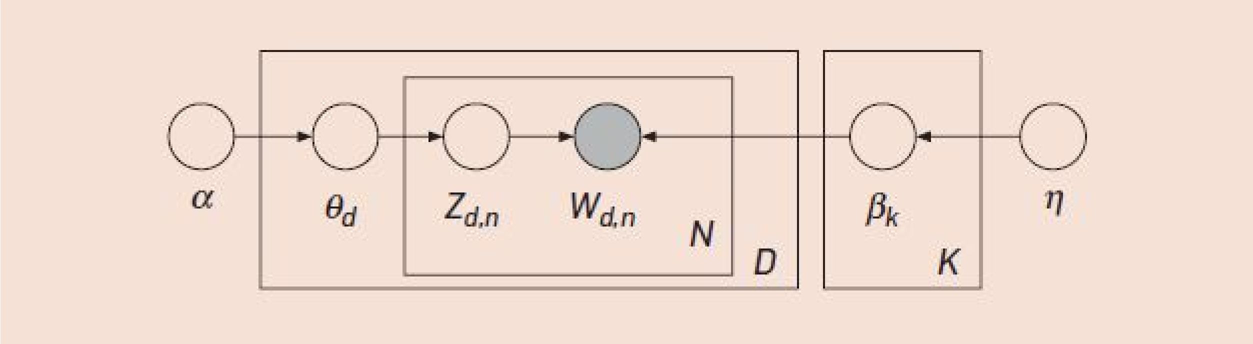
- Large unstructured collection of documents.
- Discover set of topics that generated the documents.
- Annotate documents with topics and it’s topic distributions.

Latent Dirichlet allocation (LDA)

Intuition behind LDA - Generative model



Model of LDA



- Each node is a random variable and is labeled according to its role in the generative process.
- The topics are $b_{1:k}$, where each b_k is a distribution over the vocabulary.
- θ_d is the topic proportion for the d^{th} document.
- Z_d is the topic assignment for the d^{th} document.
- W_d is the observed word for d^{th} document.

Dirichlet Priors α and β

- α is a force on the topic combinations.
  - Low α forces to pick for each doc a topic distribution which favors few topics.
  - High α allows documents to have similar, smooth topic proportions.
- β is a force on the word combinations.
  - Low β forces each topic to favors few words.
  - High β allows topics to be less distinct.

Posterior Probability for LDA

- $p(\beta_{1:k},\theta_{1:D},z_{1:D}|w_{1:D})$ = $$\frac{p(\beta_{1:K},\theta_{1:D},z_{1:D},w_{1:D})}{p(w_{1:D})}.$$

- Where, $$p(\beta_{1:K},\theta_{1:D},z_{1:D},w_{1:D}) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \left(\prod_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d,n}|\beta_{1:K}, z_{d,n})\right).$$

Steps Followed

Data Preprocessing:

1) Dataset used: bbc news, bbc sports
2) Various NLP preprocessing techniques like removing stop words,punctuation marks were applied to get a cleaned corpus.
3) Words with POS tags as NN,NNP and NNS would be the only one that contribute to find good topic distributions.
4) Used nltk tokenizer and stemmer(Porter's algorithm) to get a clean unbiased list of words.

Steps Followed...Continued

Implementation of the LDA model:

1)Initialized the topic assignment matrix with random topics and populated the word-topic matrix and document-topic matrix according to this random assignments.

2)The further process follows the **Expectation-Maximization** algorithm in which we iteratively estimate the new probabilities with this initial matrices and then change each of the matrices according to the new probabilities calculated in the expectation step and used Gibbs sampling too.

Steps Followed...Continued

3) Then we finally find the word probability distribution in each topic and the topic distribution in each document.

Implementation of Text Classification(Application of LDA):

Used the topic distribution for each document as its feature vector and did document classification with the help of SVM.

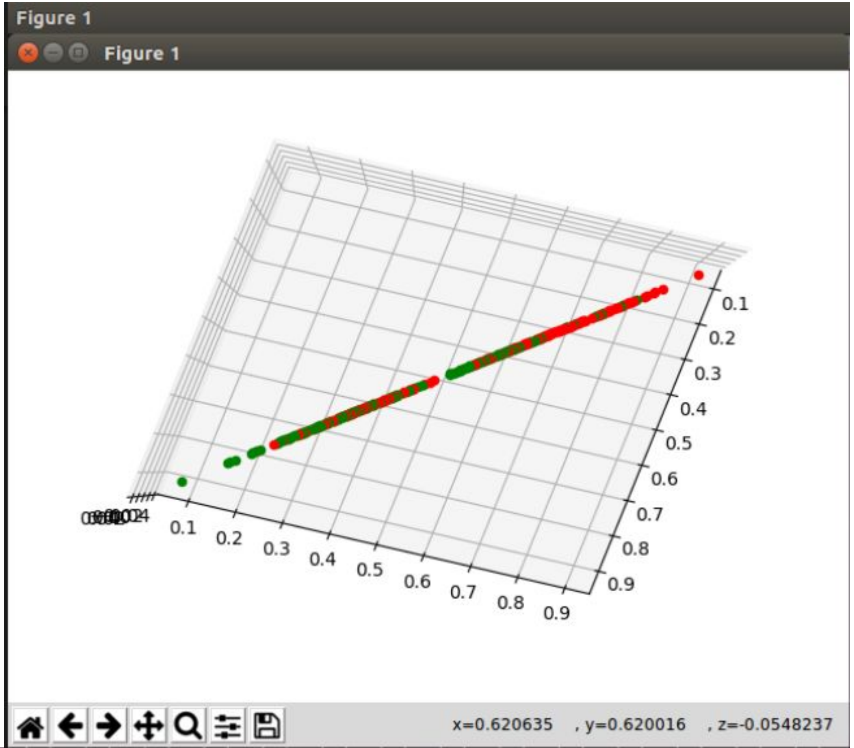
Results and Simulations

No of Topics	Labels	Accuracy
2	Sports,Entertainment	64.5%
3	Tennis,Athletic,Cricket	49.6%
3	Tennis,Football,Rugby	44.4%

Topics - 2

Labels - Sports,Entertainment

Accuracy - 64.5%

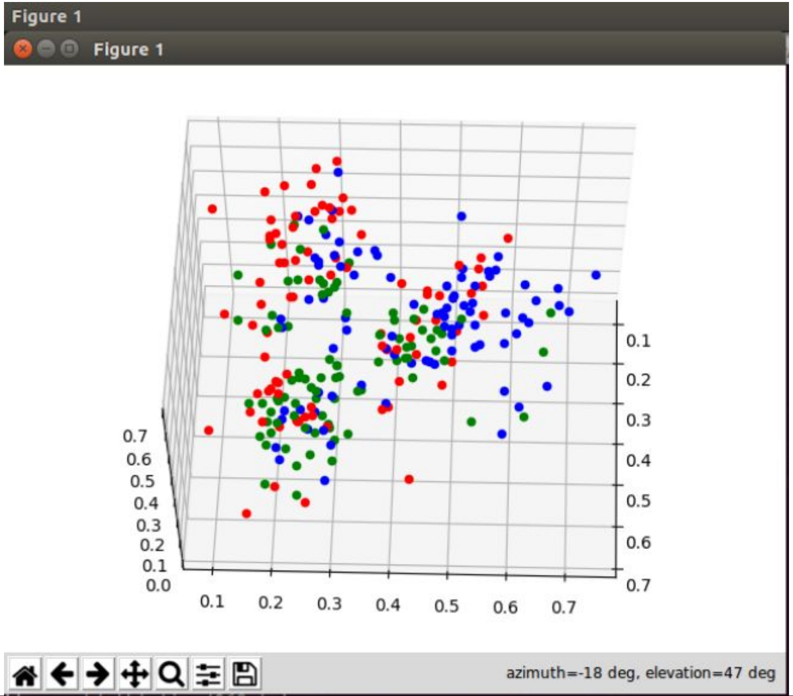



```
0.6458333333333334
topic2 topics =>  film year people number years awards world team play song record show side wales band place weeks group comedy films
topic1 topics =>  music time years film game players star world season home match actor england award club games france victory career injury
```

Topics - 3

Labels - Tennis,Athletic,Cricket

Accuracy - 49.6%

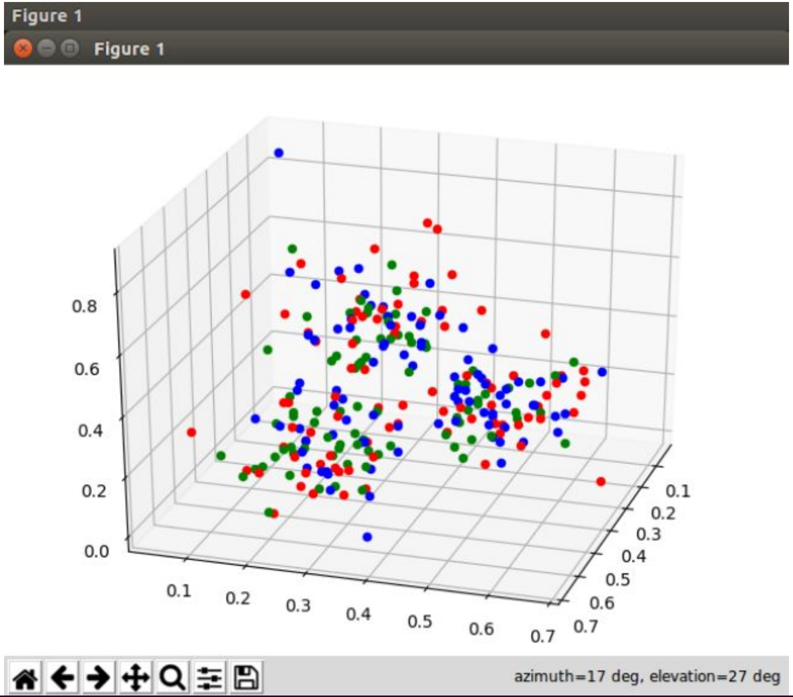



```
0.49673202614379086
topic3 topics =>  game side games match mark season team year overs dont football club title action start record player athletes race weeks
topic2 topics =>  world time years champion captain ball race place coach year michael manager team athletics olympic championships victory part peopl
e athens
topic1 topics =>  cricket players test home series play england sport chelsea tests chance decision team minutes jones days injury drugs balls austral
ia
```

Topics - 3

Labels - Tennis,Football,Rugby

Accuracy - 44.4%




```
0.444444444444444
topic1 topics =>  game team france injury games players victory nations chance world ireland league home week points start title year wales match
topic2 topics =>  time game years rugby coach players number year minutes wales beat world match roddick champion tournament weeks williams captain te
nnis
topic3 topics =>  game play world players seed matches football dont point match zealand goal side things player england penalty return squad somethin
g
```

Interclass and Intraclass Distances

Dataset used : BBC sports

```
InterClass Distance within different classes
-------------------------------------------
[[ 0.          0.2792921   0.10256726  0.14542142  0.0773072 ]
 [ 0.2792921   0.          0.19292387  0.14744622  0.23679979]
 [ 0.10256726  0.19292387  0.          0.07930518  0.08474343]
 [ 0.14542142  0.14744622  0.07930518  0.          0.09251124]
 [ 0.0773072   0.23679979  0.08474343  0.09251124  0.        ]]
IntraClass Distance for each Class
----------------------------------
[[ 0.12860021]
 [ 0.03709405]
 [ 0.06114653]
 [ 0.02556365]
 [ 0.03250214]]
```

Milestones achieved

1) Successfully implemented topic modelling with LDA from scratch and also applied Gibbs sampling to the same.
2) For larger number of topics, we also tried using SVD for dimensionality reduction to extract the important features to run classifier on the same.
3) Used the topic distribution of each document as its feature vector and used these feature vectors to do document classification using SVM resulting in a good accuracy.
4) We also evaluated our LDA model by calculating interclass and intraclass separation.
5) We also implemented LSA(Latent Semantic Analysis) model using SVD and k-means clustering to find the similarity of word which is an another approach to do topic modelling.

Thanks...