Introduction to Probabilistic Topic Modeling

White Paper

December, 2012
www.mu-sigma.com

INNOVATION AND DEVELOPMENT LABS

Introduction to Probabilistic Topic Modeling

Ankit Sethi, Bharat Upadrasta,
Innovation and Development Group,
Mu Sigma Business Solutions
Bangalore, Karnataka

December 17, 2012

Abstract

Probabilistic topic models are a collection of algorithms used to discover the hidden thematic structure in large archives of documents based on unsupervised learning. This paper entails two popular algorithms in the realm of probabilistic topic modeling namely Latent Dirichlet Allocation and Correlated Topic Modeling. Each of the algorithms are described and discussed, followed by a comparison between them.

Contents

1 Introduction **4**

1.1 A generalized Example 4

2 Latent Dirichlet Allocation **5**

2.1 Statistical representation 6

3 Correlated Topic Modeling **6**

4 Posterior Estimation **8**

5 Comparison- LDA vs CTM **9**

5.1 Perplexity 9

5.2 Entropy 9

5.3 Mutual Information 10

6 Use Case **10**

6.1 Optimum number of Topics 10

7 Variation across topics **15**

7.1 Perplexity table 15

7.2 Entropy Table 15

7.3 Mutual Information Table 15

7.4 Perplexity Vs. Number of Topics 16

7.5 Entropy Vs. Number of Topics 16

7.6 Mutual Infomation Vs. Number of Topics 17

8 Conclusion and Future Work **17**

9 References **17**

List of Figures

1 [2] The plate graph representation of Latent Dirichlet Allocation. 6

2 [3] The plate graph representation of Correlated Topic Modeling. 7

3 A plot to find optimum number of topics 10

4 A plot to compare perplexity score for the three models 14

5 A plot of Perplexity Vs. Number of Topics 16

6 A plot of Entropy Vs. Number of Topics 16

7 A plot of Mutual Information Vs. Number of Topics 17

1 Introduction

With increased online user activity, companies have started to focus on large internet user base for segmented marketing. As more and more users have started using blogs and social networking sites to express views about various products and share their experiences, it has become important to analyse data from these blogs and tweets to find hidden themes or topics.

Probabilistic Topic Modeling algorithms help us find the hidden thematic structures in documents and thus, help us categorize these documents on the basis of their themes. It also helps us find relations between themes, topics and how they evolve over a period of time. Topic modeling algorithms do not require any prior annotations or labeling of the document. Topic modeling enables us to organize and summarize electronic archives at a scale that would be impossible by human annotation.

[1]In generative probabilistic modeling, we treat our data to be arising from a generative process that includes hidden variables. This generative process defines a joint probability distribution over both the observed and hidden random variables. We perform data analysis by using a joint distribution to compute the conditional distribution of hidden variables, given observed variables. This conditional distribution is also called the posterior distribution.

This paper discusses two popular topic modeling algorithms namely Latent Dirichlet Allocation and Correlated Topic Modeling

1.1 A generalized Example

The following paraphrased example explains the generative process of the probablistic topic modeling algorithms and applies equally well to LDA and CTM.

"More than 120,000 skins of a protected species of alligator were smuggled into Japan during the past seven months using stolen or falsified export documents, a wildlife protection organization said on Thursday. Traffic Japan, the wildlife trade monitoring group of the World Wide Fund for Nature, said the South American caiman skins were shipped by a complex route involving at least seven South American and Asian countries before they arrived in Japan. At least 46 tons of the skins from more than 120,000 alligators, entered Japan in the first seven months of this year through Thailand alone, the group said. It is believed that the skins were part of a larger shipment that was loaded onto Asia-bound ships off the coast of Uruguay at the end of last year. The declared customs value of the skins was about 427 million yen (about $3.2 million), but retail value would be four to five times more,spokeswoman Cecila Song said. The skins are used in Japan mainly for belts and watchbands. Permits are required for the export of South American caiman skins under the regulations of the Convention on International Trade in Endangered Species (CITES), an international treaty regulating trade in protected plants and animals, Song said."

In the above paragraph, the highlighted words constitute vocabulary that is distributed over various topics(β_k). Topics are chosen from terms in a document, illustrating in the above example that 'Nature', 'World', 'Wildlife' etc could constitute possible topics. (The names of themes are mentioned just for better understanding and are not the part of the output given by the algorithm.) The document contains these themes or topics in various proportions(θ_d), which are calculated randomly by using dirichlet parameter α in case of LDA and normal distribution parameter μ and $\sum$ in case of CTM.

2 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a statistical model of document collections that tries to capture the multiple topics in documents. It is most easily described by its generative process, the imaginary random process by which the model assumes the documents were produced. LDA assumes that documents are made up of words and that the order of words within a document is unimportant ("bag-of-words" assumption). LDA also assumes that every document is represented by a topic distribution and each topic defines an underlying distribution on words.The distinguishing characteristic of Latent Dirichlet Allocation is that all documents in the collection share the same set of topics, but each document exhibits those topics with different proportion.

An assumption is made that these topics are specified before any data has been generated. Now for each document in the collection, the words are generated in a two-stage process.

1. Randomly choose a distribution over topics.

2. For each word in the document,

   (a) Randomly choose a topic from the distribution over topics.

   (b) Randomly choose a word from the corresponding distribution over the vocabulary.

This statistical model reflects the intuition that documents exhibit multiple topics. Each document exhibits topics with a different proportion and each word in each document is drawn from one of the topics, where the selected topic is chosen from the per-document distribution over topics.

[1] The goal of topic modeling is to automatically discover the topics from a collection of documents. The documents themselves are observed, while the topic structure- the topics, per-document topic distributions, and the per-document per-word topic assignments are hidden structure. The central computational problem for topic modeling is to use the observed documents to infer the hidden topic structure. This can be thought of as "reversing" the generative process i.e. 'what is the hidden structure that probably generated the observed collection?'

In LDA, observed variables are words of the documents, hidden variables are the topic structure and the generative process is as described above. The computational problem of inferring the hidden topic structure from the documents is the problem of computing the posterior distribution, the conditional distribution of hidden variables, given documents.

2.1 Statistical representation

The generative presentation of LDA can be described by the following equation-

[1] $p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) (\prod_{n=1}^{N} p(z_{(d,n)}|\theta_d) p(w_{d,n}|\beta_{1:K}, z_{d,n}))(1)$

The generative process can be repesented by a plate graph in the following manner:

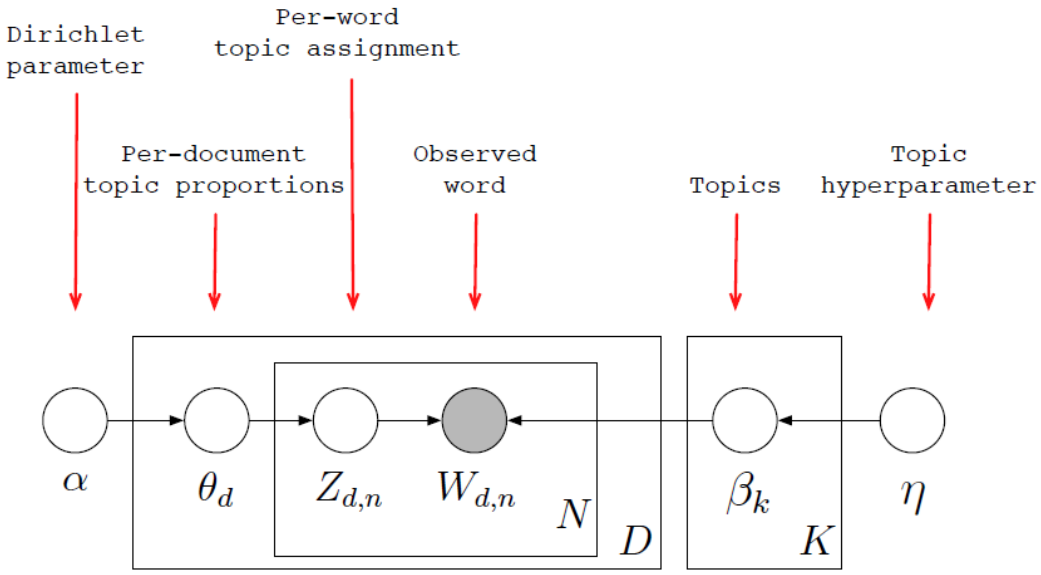

Figure 1: [2] The plate graph representation of Latent Dirichlet Allocation.

The shaded nodes are observed variables and unshaded nodes are latent or hidden variables. The rectangles denote "plate" which means replication. The D plate denotes a collection of documents(corpus) and N plate denotes words contained within documents. The K plate denotes words in vocabulary.

3 Correlated Topic Modeling

The correlated topic model (CTM) is a hierarchical model of document collections. It builds on LDA model by relaxing the assumption that a document is a "bag of words". It maintains a strong sequence of words occuring in the document.

Following terminology and notations are used to describe the data, latent variables and parameters in the CTM:

a. Words and documents: The only observable random variables that we consider are words that are organized into documents. Let $w_{d,n}$ denote the nth word in the dth document, which is an element in a V term vocabulary. Let w_d denote the vector of N_d words associated with document d.

b. Topics: A topic β is a distribution over the vocabulary. The model will contain K topics $\beta_{1:K}$.

c. Topic assignments: Each word is assumed to be drawn from one of the K topics. The topic assignment $z_{d,n}$ is associated with the nth word and dth document.

d. Topic proportions: Each document is associated with a set of topic proportions θ_d. Thus, θ_d is a distribution over topic indices and reflects the probabilities with which words are drawn from each topic in the collection.

[3]The correlated topic model assumes that an N-word document arises from the following generative process. Given topics $\beta_{1:K}$, a K-vector μ and a K*K covariance matrix $\sum$:

1. Draw $\eta_d|\{\mu, \sum\}$ ~ N(μ, $\sum$).
2. For n ϵ $\{1,\cdots, N_d\}$:

(a) Draw topic assignment $Z_{d,n}|\eta_d$ from Mult(f(η_d))

(b) Draw word $w_{d,n}|z_{d,n},\beta_{1:K}$ from Mult($\beta_{z_{d,n}}$)

The graphical representation of CTM can be done as follows:

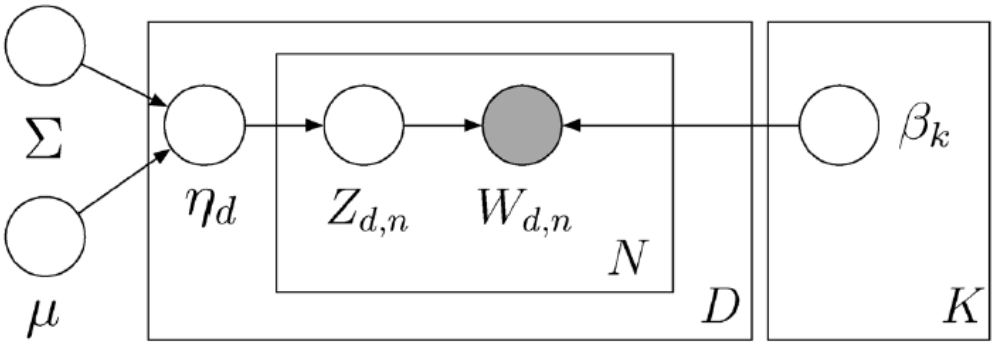

Figure 2: [3] The plate graph representation of Correlated Topic Modeling.

The D plate denotes a collection of documents(corpus) and N plate denotes words contained within documents. The K plate denotes words in vocabulary. η_d denotes the logistic normal distribution, used to model latent topic proportions of a document.(Notations have their usual meanings.)

The drawback of CTM model is that it has a complicated way to compute posterior as compared to LDA model but the output given by CTM model makes more sense than LDA model as it takes into account correlation amongst topics.

4 Posterior Estimation

According to the equation(1) we need to compute the posterior given by:

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}|w_{1:D}) = \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})}{p(w_{1:D})} \tag{2}$$

[1] The numerator is the joint distribution of all the random variables, which can be easily computed for any setting of the hidden variables. The denominator is the marginal probability of the observations, which is the probability of seeing the observed corpus under any topic model. In theory, it can be computed by summing the joint distribution over every possible instantiation of the hidden topic structure. That number of possible topic structures, however, is exponentially large; this sum is intractable to compute.

For the estimation of the posterior two approaches, namely, Variational inference method and Gibbs sampler method are used.

[1] In Gibbs sampling, a Markov chain is constructed that is a sequence of random variables, each dependent on the previous- whose limiting distribution is the posterior. The Markov chain is defined on the hidden topic variables for a particular corpus, and the algorithm is to run the chain for a long time, collect samples from the limiting distribution, and then approximate the distribution with the collected samples. (Often, just one sample is collected as an approximation of the topic structure with maximal probability.) See [4] for a good description of Gibbs sampling for LDA.

Variational methods are a deterministic alternative to sampling-based algorithms. Rather than approximating the posterior with samples, variational methods posit a parameterized family of distributions over the hidden structure and then find the member of that family that is closest to the posterior. Thus, the inference problem is transformed to an optimization problem. Variational methods open the door for innovations in optimization to have practical impact in probabilistic modeling.

5 Comparison- LDA vs CTM

The data used has been obtained from the following link http://www.cs.princeton.edu/ blei/lda-c/. It consists of 2246 documents from the Associated Press and the estimated number of topics are 100. The data looks like:

```
R Version:  R version 2.15.2 (2012-10-26)
```

```
chr [1:2243] "yearold student private baptist school teacher wounded firing
   schools pastor george pastor atlantic shores baptist church chris"|
   __truncated__ "bechtel israel discount promised proposed iraqi pipeline
   foreign ministry wednesday thenprime minister shimon peres bruce partn"|
   __truncated__ "gunman yearold woman hostage foiled steal jewelry belonging
   liberace police entertainers museum hostage margaret bloomberg poli"|
   __truncated__ " saturday reminder daylightsaving local clocks todays
   highlight history black tuesday descended stock exchange prices panic tho"|
   __truncated__ ...
NULL
```

Models are evaluated and assessed on the basis of characterizing measures such as perplexity, entropy and mutual information, concepts popular in information theory.

5.1 Perplexity

means how easily does a model react to test data. Statistically, it is equivalent to the geometric mean inverse per word likelihood.

[5] Perplexity(w)= $\exp\left(-\frac{\log(p(w))}{\sum_{d=1}^{D}\sum_{j=1}^{V} n^{(jd)}}\right)$ (3)

5.2 Entropy

Entropy is a measure of uncertainty associated with random variables. In terms of information theory, it genrally refers to Shanon's Entropy(H). Higher values of entropy means the spread of topic distributions over topics is even [5].

$$H = -\sum_{n=1}^{N} p(\beta_{1:K}, \theta_{1:D}, z_{1:D}|w_{1:D}) \log p(\beta_{1:K}, \theta_{1:D}, z_{1:D}|w_{1:D}) \tag{4}$$

5.3 Mutual Information

Mutual Information(MI) is a measure of mutual dependency of random variables.

$$MI(X,Y) = H(X) + H(Y) - H(X,Y) \tag{5}$$

where, X and Y are random variables.

6 Use Case

Data for the use case is courtesy of the Associated Press. In light of model building and evaluation, the data is split into training and test data respectively.

6.1 Optimum number of Topics

The choice of optimum topics is essential in model building to strike a balance between the extremes of under-fitting and over-fitting a model. The most frequently occurring value of alpha (parameter of the dirichlet distribution) is chosen as the threshold corresponding to which the number of topics are fixed. The following graph illustrates the optimum number of topics chosen for this exercise.

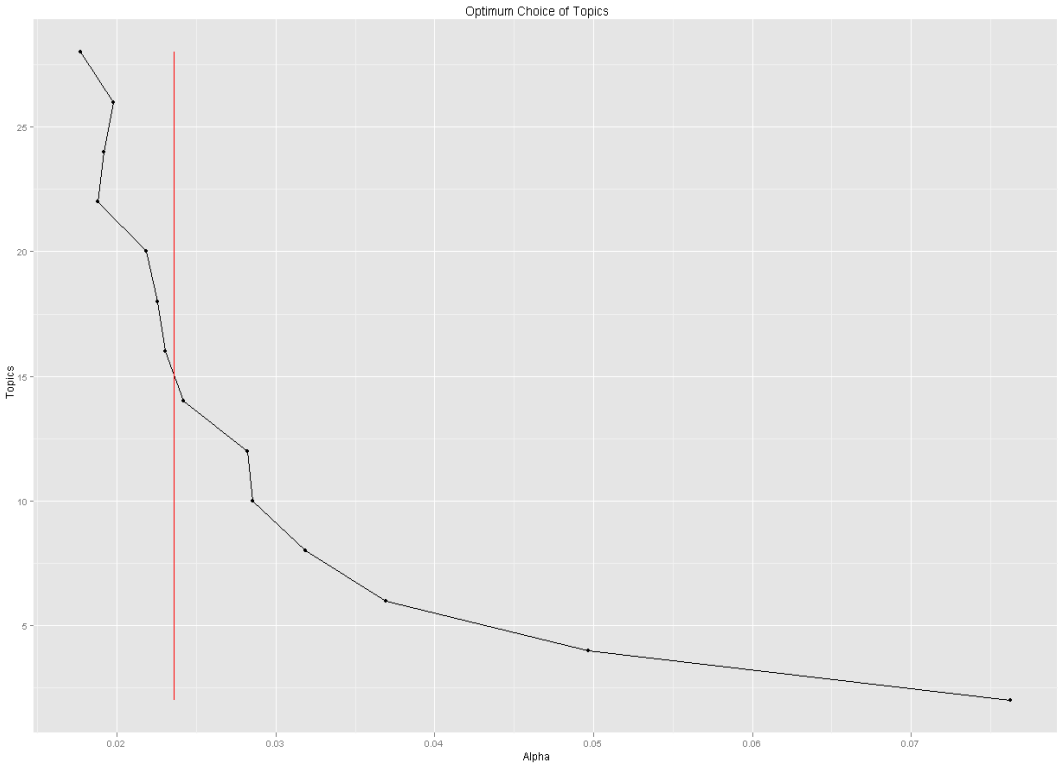

Figure 3: A plot to find optimum number of topics

LDA Model- The two approaches for computation of posterior probability used are Variational Expectation Maximization(VEM) and Collapsed Gibbs sampler(CGS).

1 Variational Expectation Maximization approach:

a The output of LDA is as follows:

```
      Topic 1         Topic 2       Topic 3      Topic 4        Topic 5
 [1,] "shearson"      "incentives"  "kiesner"    "proportion"   "duvalier"
 [2,] "subscriptions" "twomonth"    "atlantic"   "zionism"      "false"
 [3,] "jumped"        "photographs" "warnings"   "temporarily"  "create"
 [4,] "tasks"         "coniston"    "symbion"    "artist"       "sentenced"
 [5,] "worsening"     "leaned"      "serbian"    "import"       "sixtyfour"
 [6,] "fiberglas"     "listed"      "suing"      "searching"    "programs"
 [7,] "pfizer"        "tavarez"     "vladimir"   "combining"    "exhibition"
 [8,] "rates"         "transmit"    "leroy"      "productivity" "drugrelated"
 [9,] "broadcasters"  "cargo"       "restricts"  "chain"        "diplomats"
[10,] "parole"        "southwest"   "regulation" "involvement"  "fruehauf"
      Topic 6     Topic 7           Topic 8       Topic 9        Topic 10
 [1,] "morale"    "jackson"         "afterward"   "breach"       "nightmare"
 [2,] "murderer"  "white"           "financier"   "expenditures" "organ"
 [3,] "noticed"   "proindependence" "volume"      "guests"       "patient"
 [4,] "yielded"   "learn"           "peters"      "gorbachevs"   "duplicated"
 [5,] "clients"   "cosmetics"       "refrain"     "thousand"     "nikolai"
 [6,] "legalized" "widely"          "delhi"       "fighters"     "object"
 [7,] "spinal"    "hashimoto"       "radioactive" "understand"   "citizenship"
 [8,] "predicts"  "changed"         "platform"    "clinging"     "broiler"
 [9,] "bursts"    "kuwait"          "bankrupt"    "landslide"    "rockets"
[10,] "sample"    "anchored"        "fetus"       "circulation"  "fraud"
```

b Perplexity:

```
[1] 10252.3
```

c Entropy and Mutual Information:

	Entropy	MutualInformation
1	10.95	0.00

	Entropy	MutualInformation
1	11.48	0.00

2 Collapsed Gibbs Sampler approach:

a The output of LDA is as follows:

```
      Topic 1      Topic 2        Topic 3      Topic 4      Topic 5
 [1,] "abandoned"  "abctvs"       "abductors"  "abolition"  "abdelmeguid"
 [2,] "abdomen"    "acknowledge"  "abolish"    "absent"     "abductions"
 [3,] "abide"      "acquisition"  "abusers"    "actor"      "ability"
 [4,] "ablaze"     "administered" "achieved"   "admiration" "abound"
 [5,] "abolished"  "adoption"     "actions"    "advancing"  "absence"
 [6,] "abortions"  "adventure"    "activated"  "adviser"    "absolutely"
 [7,] "absorbed"   "aerialist"    "activist"   "aerial"     "abstentions"
 [8,] "achievable" "aeroflot"     "activities" "afraid"     "achievements"
 [9,] "achieve"    "aerospace"    "acute"      "agadir"     "acquired"
[10,] "achieving"  "agencys"      "adequately" "airman"     "acustar"
```

```
      Topic 6        Topic 7         Topic 8         Topic 9
 [1,] "abandon"      "abortion"      "aboard"        "abduction"
 [2,] "abctv"        "abstainers"    "abortive"      "aborted"
 [3,] "abducted"     "academics"     "abrupt"        "abrams"
 [4,] "abdul"        "acquire"       "absolute"      "abundant"
 [5,] "abroad"       "acting"        "abstained"     "academic"
 [6,] "abruptly"     "actors"        "abuse"         "acknowledgment"
 [7,] "abuses"       "adjust"        "academy"       "acreage"
 [8,] "acampora"     "administrator" "acknowledged"  "acted"
 [9,] "acquisitions" "adult"         "acknowledges"  "active"
[10,] "acres"        "adventures"    "acknowledging" "actively"
      Topic 10
 [1,] "abandoning"
 [2,] "abused"
 [3,] "acquiring"
 [4,] "adjacent"
 [5,] "adjournment"
 [6,] "advantage"
 [7,] "adversary"
 [8,] "adverse"
 [9,] "adversely"
[10,] "advertisement"
```

b Perplexity:

```
[1] 9685.442
```

c Entropy and Mutual Information:

	Entropy	MutualInformation
1	10.95	0.00

	Entropy	MutualInformation
1	11.48	0.00

CTM Model- This uses VEM method for computation of posterior probability.

a The output of CTM is as follows:

```
      Topic 1          Topic 2       Topic 3           Topic 4
 [1,] "decrease"       "warning"     "escape"          "arizona"
 [2,] "klein"          "tinker"      "morality"        "recruit"
 [3,] "peasants"       "crime"       "filipinos"       "grape"
 [4,] "floors"         "phnom"       "longterm"        "reinstein"
 [5,] "jumping"        "interviewed" "armsforhostages" "release"
 [6,] "plight"         "voters"      "elizabeth"       "firebombs"
 [7,] "chrysler"       "holidays"    "diagnosed"       "unloaded"
 [8,] "regret"         "contractor"  "denominations"   "abdelmeguid"
 [9,] "elect"          "recorded"    "jukebox"         "hudson"
[10,] "antidefamation" "urine"       "spectators"      "convicted"
      Topic 5      Topic 6      Topic 7       Topic 8       Topic 9
 [1,] "trail"      "nationals"  "reason"      "fenced"      "sheik"
 [2,] "doctorate"  "posts"      "patients"    "charging"    "egizio"
 [3,] "ministries" "republican" "control"     "rover"       "revealed"
 [4,] "gambling"   "loved"      "jersey"      "articulated" "smith"
 [5,] "rudmans"    "broderick"  "blackowned"  "glued"       "relied"
 [6,] "thousands"  "cochair"    "stage"       "factor"      "expectation"
 [7,] "blueprint"  "closures"   "boesky"      "corporate"   "sailors"
 [8,] "dolphin"    "halting"    "cochairman"  "conscious"   "malicious"
 [9,] "humphrey"   "directly"   "jeopardized" "client"      "utilities"
[10,] "outcome"    "seasons"    "ordeal"      "resolution"  "shrimp"
      Topic 10
 [1,] "birds"
 [2,] "lawandorder"
 [3,] "reserve"
 [4,] "elvis"
 [5,] "disciples"
 [6,] "prosperous"
 [7,] "broadcasting"
 [8,] "populace"
 [9,] "investigate"
[10,] "proud"
```

b Perplexity:

```
[1] 9925.727
```

c Entropy and Mutual Information:

	Entropy	MutualInformation
1	11.48	0.00

The table showing perplexity, entropy and mutual entropy for LDA using Variational Expectation Maximization(VEM) and Collapsed Gibbs Sampling(CGS), CTM using VEM approach:

	Entropy	MutualInformation
1	11.48	0.00

	Perplexity	Entropy	MutualInformation
LDA_VEM	10252.30	11.48	0.00
LDA_CGS	9685.44	11.48	0.00
CTM	9925.73	11.48	0.00

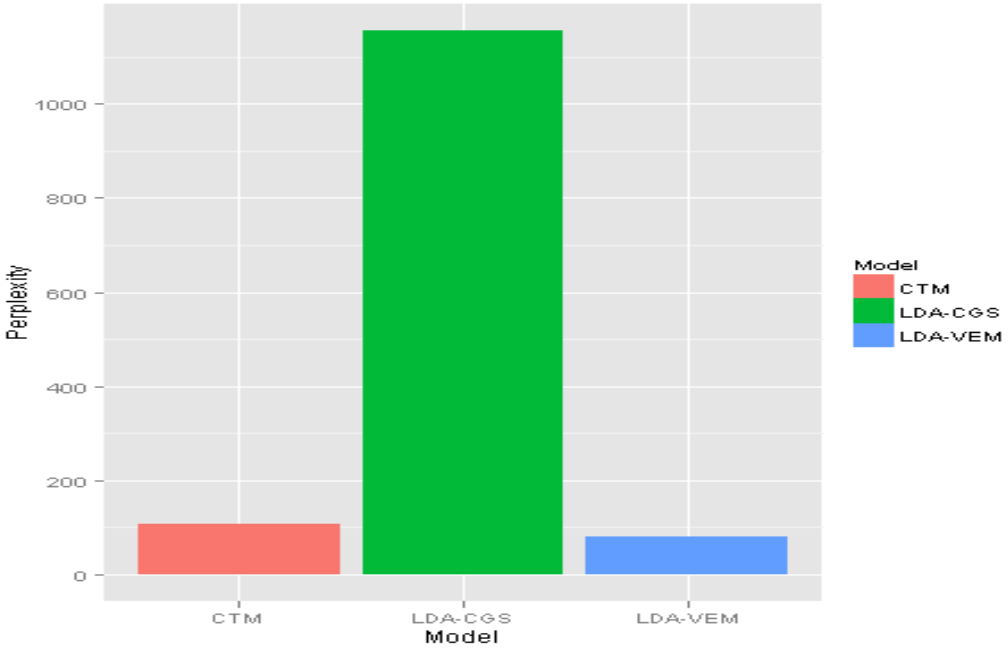

Figure 4: A plot to compare perplexity score for the three models

7 Variation across topics

7.1 Perplexity table

	Topics	LDA-VEM	LDA-CGS	CTM
1	3.00	315.41	1250.92	356.46
2	8.00	141.28	1208.58	184.15
3	13.00	96.02	1188.14	126.02
4	18.00	71.72	1140.84	97.92
5	23.00	60.05	1125.60	81.39
6	28.00	48.50	1099.46	70.35

7.2 Entropy Table

	Topics	LDA-VEM	LDA-CGS	CTM
1	3.00	6.82	6.84	7.72
2	8.00	7.31	7.82	8.74
3	13.00	8.30	8.31	9.24
4	18.00	8.63	8.63	9.58
5	23.00	8.87	8.88	9.84
6	28.00	9.07	9.08	10.04

7.3 Mutual Information Table

	Topics	LDA-VEM	LDA-CGS	CTM
1	3.00	0.02	0.00	0.12
2	8.00	0.01	0.00	0.09
3	13.00	0.01	0.00	0.07
4	18.00	0.01	0.00	0.05
5	23.00	0.01	0.00	0.04
6	28.00	0.01	0.00	0.04

7.4 Perplexity Vs. Number of Topics

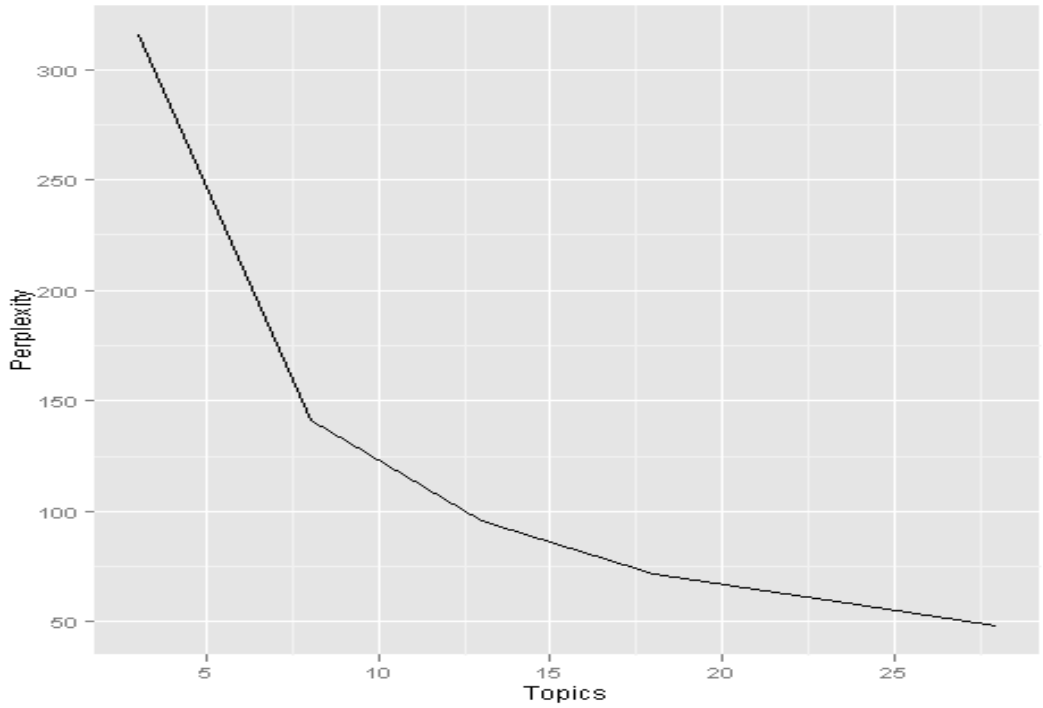

Figure 5: A plot of Perplexity Vs. Number of Topics

7.5 Entropy Vs. Number of Topics

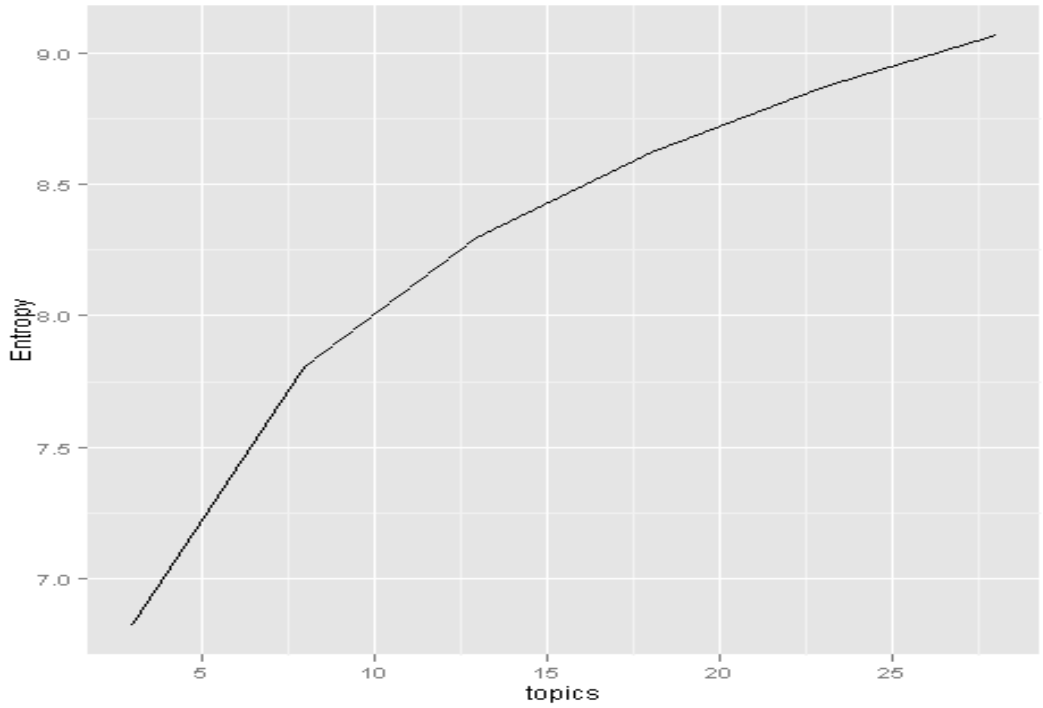

Figure 6: A plot of Entropy Vs. Number of Topics

7.6 Mutual Infomation Vs. Number of Topics

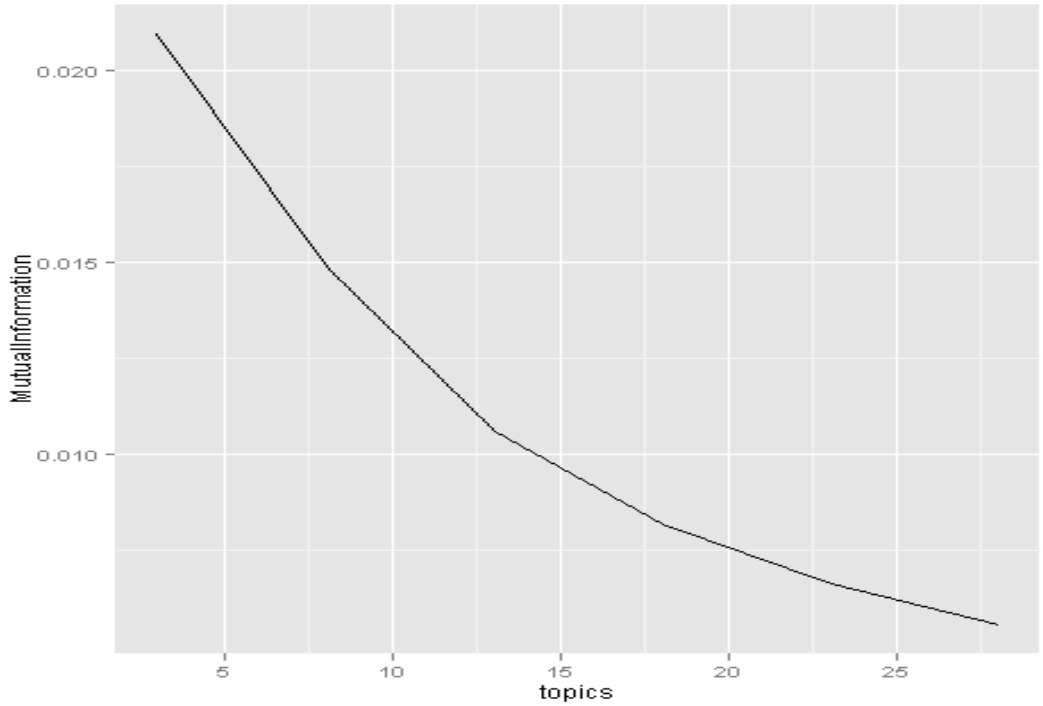

Figure 7: A plot of Mutual Information Vs. Number of Topics

8 Conclusion and Future Work

The LDA and CTM models give best results when the optimum number of topics is known i.e. if the model is not overfitted or underfitted. To find the optimum number of topics, calculation of max loglikelihood for a range of topic values is done iteratively and the topic number that gives us the maximum loglikelihood is chosen. This approach needs to be made more intuitive which can be done using Bayesian Non-parametric approaches.

9 References

[1] Blei, D.(2011) Introduction to Probabilistic Topic Models, Princeton University

[2] Blei, D. and Lafferty J. (2008) Modeling Science, Princeton University and Carnegie Mellon University

[3] Blei, D. and Lafferty J. (2007) A correlated topic model of Science. Annals of Applied Statistics

[4] M. Steyvers and T. Grifiths. Probabilistic topic models. In T. Landauer, D. McNamara, S. Dennis, and W. Kintsch, editors, Latent Semantic Analysis: A Road to Meaning. Laurence Erlbaum, 2006.

[5] Bettina Gruen and Kurt Hornik topicmodels: An R Package for Fitting Topic Models Johannes Kepler Universitaet and Linz WU Wirtschaftsuniversitaet Wien

Mu Sigma is a leading provider of decision sciences and analytics solutions, helping companies institutionalize data-driven decision making. We work with market-leading companies across multiple verticals, solving high impact business problems in the areas of Marketing, Supply Chain and Risk analytics. For these clients we have built an integrated decision support ecosystem of people, processes, methodologies & proprietary IP and technology assets that serve as a platform for cross-pollination and innovation. Mu Sigma has driven disruptive innovation in the analytics industry by integrating the disciplines of business, math, and technology in a sustainable model. With over 75 Fortune 500 clients and over 2000 decision science professionals we are one of the largest pure-play decision sciences and analytics companies.

Learn more at *http://www.mu-sigma.com/contact.html* us for further information:

Mu Sigma Inc., 3400 Dundee Rd, Suite 160, Northbrook, IL – 60062 www.mu-sigma.com