An Empirical Bayes Approach to Topic Modeling

Anirban Gangopadhyay[1]

[1]Department of Computer Science
Columbia University

ICPR 2020

Outline

Problem Statement

- Consider the problem of modeling text corpora
  - Finding short descriptions of members of a collection (ie derive topics, communities, clusters, etc)
  - Scoring similarities across members (ie finding similar documents, topics)
- Many disparate models exist
- Derived from different mathematical disciplines that make different assumptions about data population

Solution

- We give an Empirical Bayes framework for choosing optimal algorithm
- Map each algorithm to a graphical model
- Choose optimal model given the observed variables in our data (based on the assumptions each model makes)

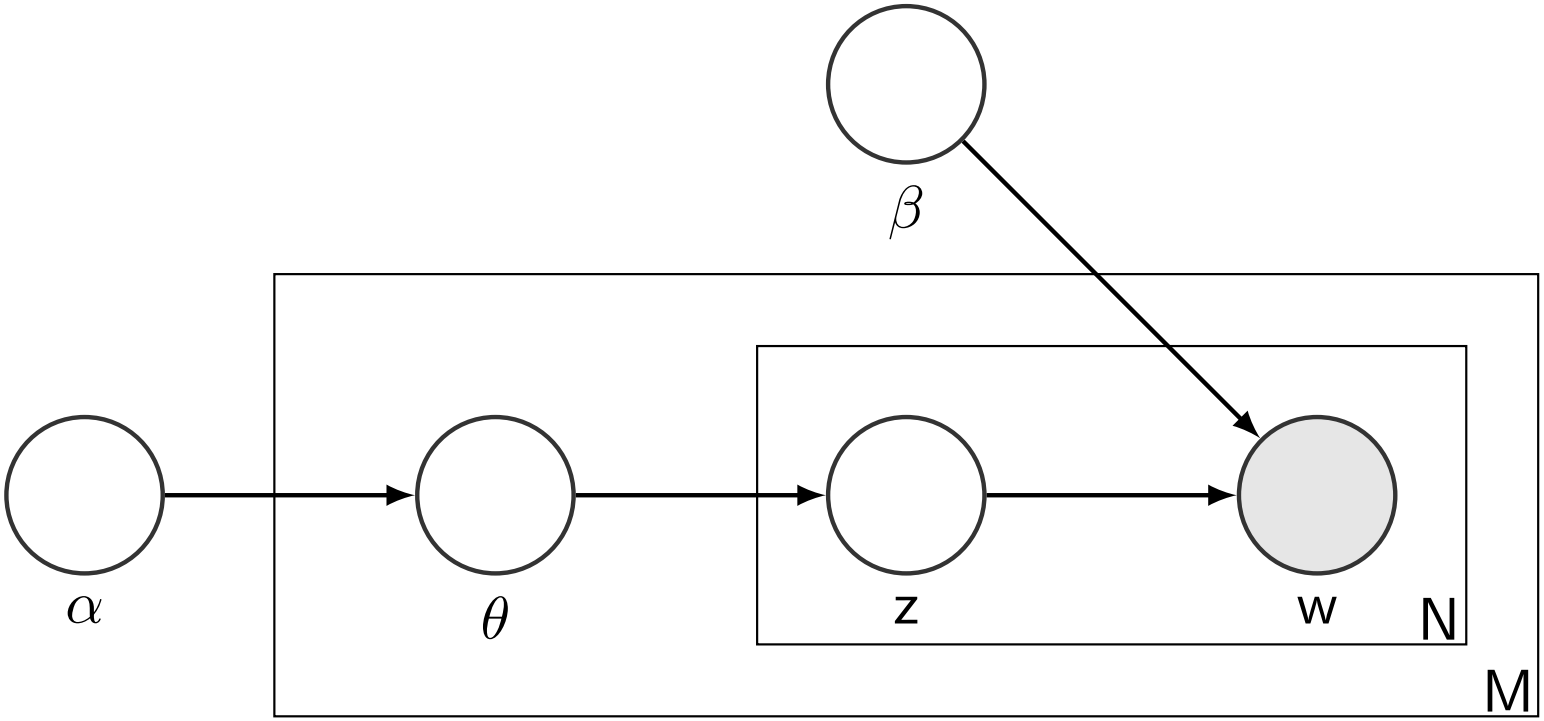
Outline

Generative Process

We outline the generative process for each document $\overrightarrow{w}$ in a corpus D:

1. Choose $N \sim Poisson(\xi)$
2. Choose $\Theta \sim Dir(\alpha)$
3. For each of the N words w_n:
   1. Choose a topic $z_n \sim Mult(\Theta)$
   2. Choose a word w_n from $p(w_n|z_n, \beta)$, a multinomial probability conditioned on topic z_n.

Graphical Model



Joint Distribution

- Joint distribution: $p(\theta, z, \vec{w}|\alpha, \beta) = p(\theta|\alpha)\prod_{n=1}^{N} p(z_n|\theta)p(w_n|z_n, \beta)$.
- To obtain the marginal distribution of a document, we integrate over θ and sum over z:
- $p(\vec{w}|\alpha, \beta) = \int p(\theta|\alpha) \cdot (\prod_{n=1}^{N} \sum_{z_n} p(z_n|\theta) \cdot p(w_n|z_n, \beta))d\theta$.
- Since we assume documents are i.i.d., we obtain the probability of the corpus by taking $\prod_{d=1}^{M} p(\vec{w}|\alpha, \beta)$.
- we wish to perform a maximum likelihood of
- $p(D|\alpha, \beta) = \prod_{i=1}^{M} \int p(\theta_d|\alpha) \cdot (\prod_{n=1}^{N_d} \sum_{z_{d_n}} p(z_{d_n}|\theta_d) \cdot p(w_{d_n}|z_{d_n}, \beta))d\theta_d$

Outline

Inference

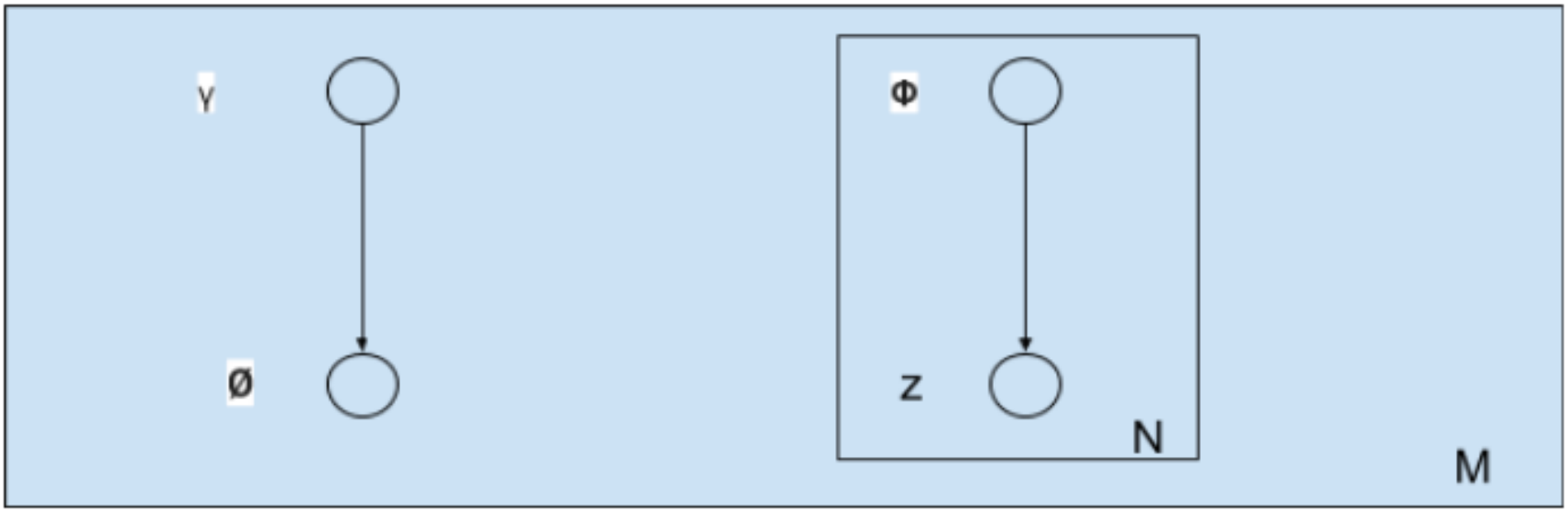
- Posterior distribution given a document is intractable to compute:
- $p(\theta, z|w, \alpha, \beta) = \frac{p(\theta,z,w|\alpha,\beta)}{p(w|\alpha,\beta)}$.
- We consider variational inference as an approximate inference method
- **Variational Inference** Derive parameters of a tractable model that has minimal KL divergence to LDA model

Variational Inference

- Using variational inference, we obtain a family of distributions on the latent variables $\gamma, \phi_1, ..., \phi_m$
- $q(\theta, z|\gamma, \phi) = q(\theta|\gamma) \cdot \prod_{n=1}^{N} q(z_n|\phi_n)$
- Dirichlet parameter γ, multinomial parameters $\phi_1, ..., \phi_m$ are free variational parameters
- The family of distributions are a family of lower bounds derived via Jensen's inequality.

Graphical Model

Variational Inference



Variational Inference

Continued

- We drop the edges between θ, z, w and the w nodes in the above model
- To find the optimal γ, ϕ, we minimize the KL divergence
- $(\gamma^*, \phi^*) = \text{argmin}_{\gamma,\phi} D(q(\theta, z|\gamma, \phi)||p(\theta, z|w, \alpha, \beta))$
- γ, ϕ are conditioned on $\overrightarrow{w}$ which implies that:
- $(\gamma^*, \phi^*) = \text{argmin}_{\gamma,\phi} D(q(\theta, z|\overrightarrow{w}(\gamma), \overrightarrow{w}(\phi))||p(\theta, z|\overrightarrow{w}, \alpha, \beta))$

Variational Inference

Continued Part II

Since α, β are hidden, we can use Expectation-Maximization to derive γ^*, ϕ^*.

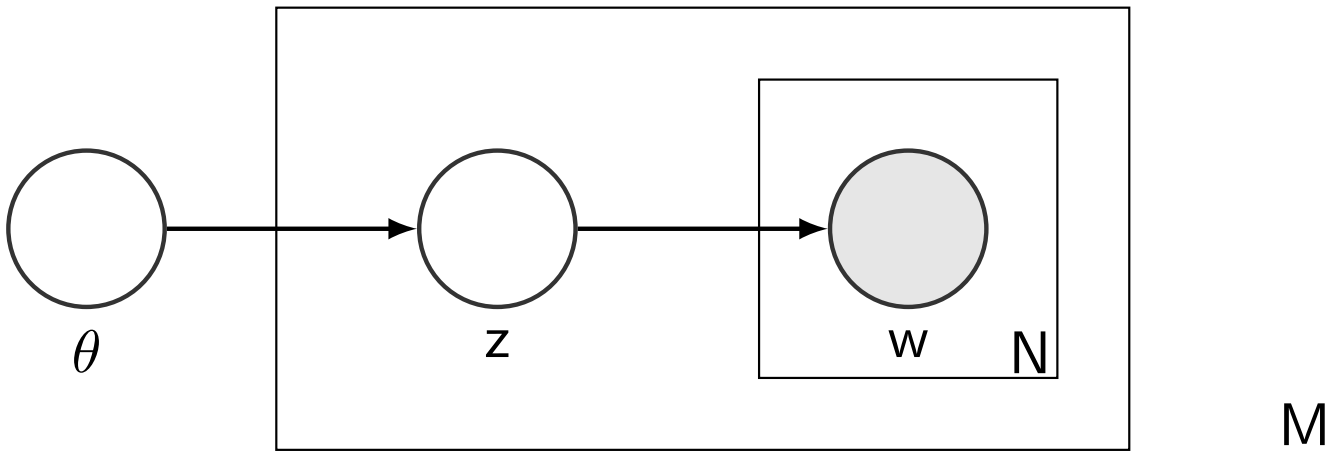
1. (E-step) For each document, find the optimizing values of the variational parameters $\{\gamma_d^*, \phi_d^* : d \in D\}$
2. (M-step) Maximize the resulting lower bound on the log likelihood w.r.t. α, β (corresponds to finding the MLE for each document under approximate posterior)

Outline

Statistical Assumptions

- **Graphical Model**: LDA allows each document to be generated from a distribution of topics whereas the two level Dirichlet-Multinomial clustering model would restrict a document to be associated with a single topic.
- **Exchangeability**: DeFinetti's theorem states that the joint distribution of an infinitely exchangeable sequence of random variables is as if a random parameter were drawn from some distribution and then the random variables in question were i.i.d. conditioned on the parameter.
- Allows us to assume a bag of words model for LDA (words are i.i.d. conditioned on the topic it is generated from)

Two Layer Model



Outline

Algorithmic Process

We outline the algorithmic process for deriving topics (communities) given a corpus:

1. Construct a network graph where each node represents a word w_i in our overall dictionary (constructed from D) and the edge weight e_{w_i,w_j} between two words represents the number of documents w_i, w_j appear in together.
2. We consider two metrics as a measure of goodness of community structure for the weighted, undirected graph constructed above: modularity Q and modularity density Q_{ds}.
   - Modularity: $Q(G) = \sum_{C_i \in C}[\frac{W_{c_i}^{in}}{W} - (\frac{W_{c_i}}{2W})^2]$ where W represents the sum of all edge weights in the entire graph, $W_{c_i}^{in}$ is the sum of weights of edges within community C_i, $W_{C_i} = 2W_{C_i}^{in} + W_{C_i}^{out}$ where $W_{C_i}^{out}$ is the sum of weights of edges with exactly one endpoint inside C_i.

Algorithmic Process

Continued

- We wish to maximize Q (or Q_{ds}) to obtain the optimal set of communities $C = \{C_i\}$. These communities (set of words) may represent topics.
- A document may be broken down into a set of topics based on the span of words in $\overrightarrow{w}$.

Generative Process

We outline the generative process (and underlying statistical assumptions) for generating a network graph, its underlying communities and the edges (based on the assumptions made in the case where modularity is maximized).

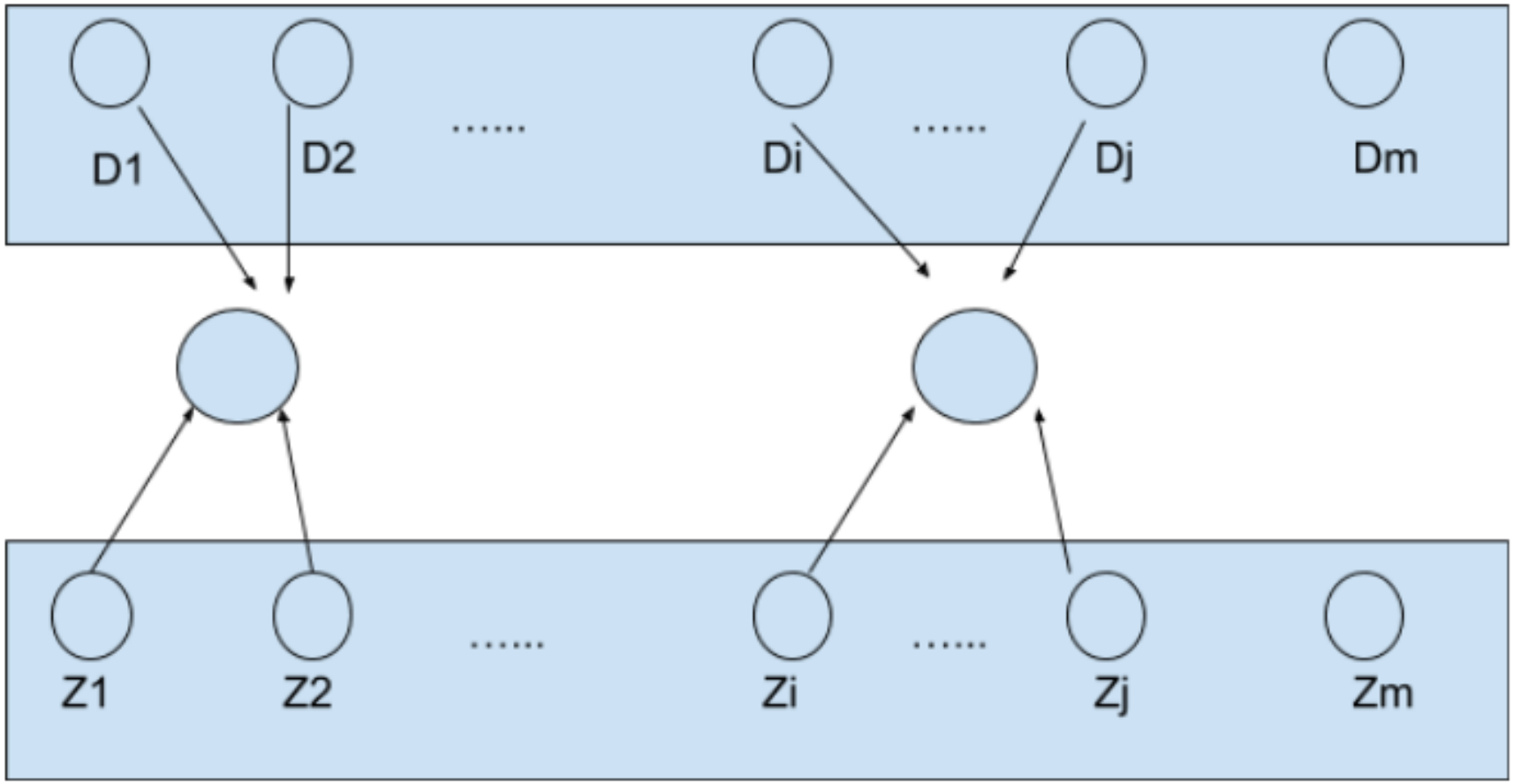
1. Choose the number of nodes $N \sim Dist(\alpha)$ where α is a hyperparameter and $Dist(\cdot)$ is a distribution TBD
2. For $i = 1$ to n:
   1. Choose $d_i \sim Dist2(\beta)$ where d_i represents the degree of node i, β is a hyperparameter, $Dist2(\cdot)$ is TBD.
   2. Choose the cluster $z_n \sim Mult(\Theta)$
3. Let us denote A_{ij} to be the edge weight between node i and node j. Then $A_{ij} \sim Dist3(\cdot)$ as a function of d_i, d_j, z_i, z_j

Single Layer Model

- We derive full joint probability, graphical model for the case where all priors are point parameters and only the edges are generative.
- **Goal**: choose $\text{argmax}_{\overrightarrow{z}} p(A|N, \overrightarrow{d}, \alpha, \beta; \overrightarrow{z})$
- Wish to choose $p(A|N, \overrightarrow{d}, \alpha, \beta; \overrightarrow{z})$ such that modularity Q is maximized.
- Can be shown that choosing $\text{argmax}_z L(\overrightarrow{z})$ is equivalent to choosing $\text{argmax}_z Q(\overrightarrow{z})$
- **Key insight**: By choosing Dist$(\cdot)$, Dist2$(\cdot)$ to be point parameters, Dist3$(\cdot) = \delta_{z_i=z_j} N(\mu_1, \sigma) + \delta_{z_i \neq z_j} N(\mu_2, \sigma)$, we have derived a one-layer generative process whose ML parameters will maximize modularity

Graphical Model

Single Layer



Three Layer Model

Derive joint probability, graphical model for the case where we presume a generative process for each derived parameter (only hyperparameters are point parameters).

- Choose Dist($\cdot$) appropriately
- Dist2($\cdot$) is traditionally Poisson and is a fixed observed variable in the configuration model. We derive the Poisson assumption below based on reasonable base assumptions.
- Dist3($\cdot$) was assumed to be a GMM in the single layer model and Poisson($d_i, d_j, \omega_{z_i,z_j}$) in the stochastic block model (which we use for inference).

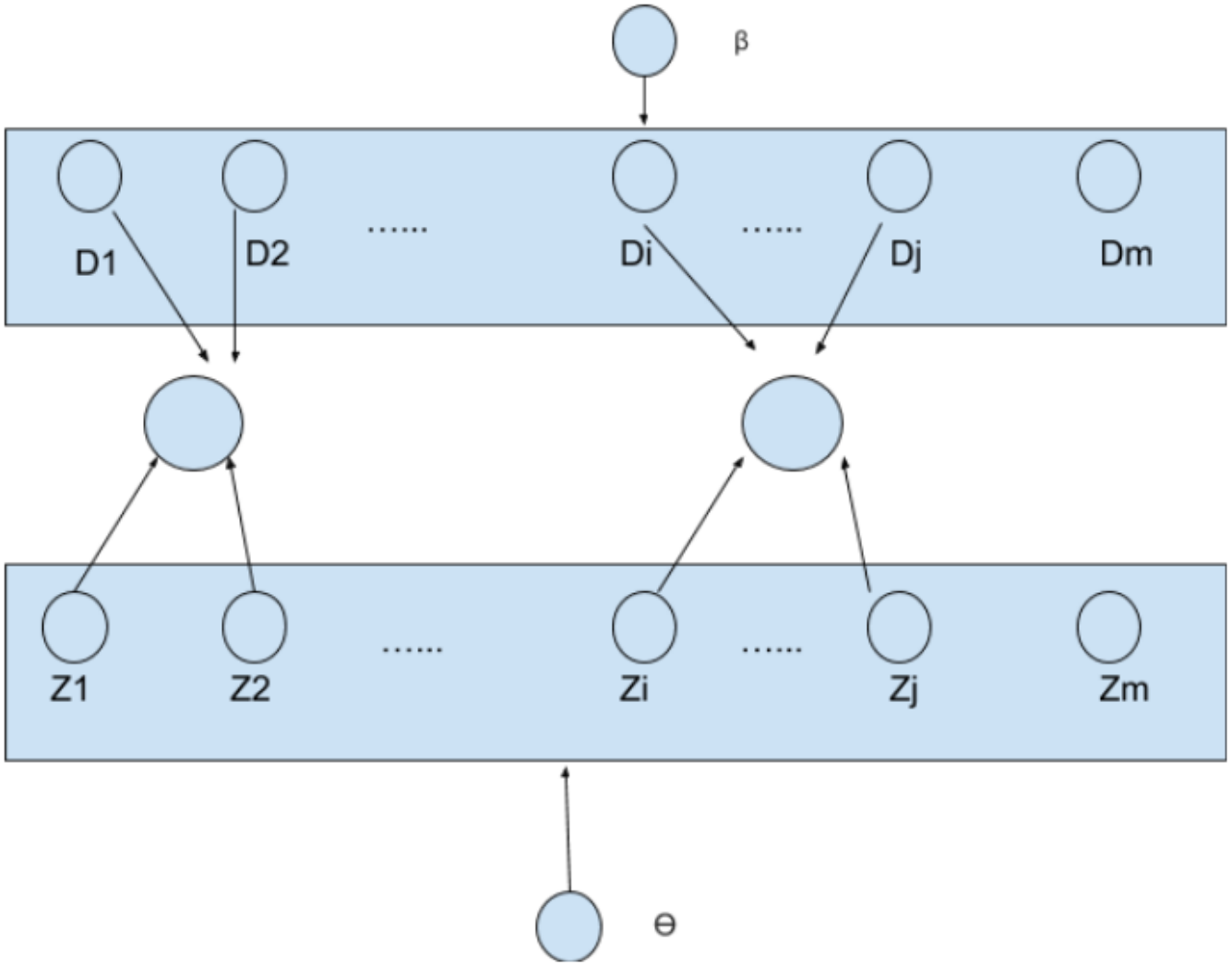
Three Layer Model

Dist2(·) derivation

1. Assume the probability of an edge between two nodes is represented by p.
2. Furthermore the probability of an edge between two nodes is assumed to be independent of the existence of any other edges in the graph
3. Let z be the average number of edges a given node is connected to. Then $p = \frac{z}{N-1}$ where N is the total number of nodes.
4. Let us denote p_k to be the probability a given node has degree k. Then p_k can be derived as follows: $p_k = \binom{N}{k} p^k \cdot (1-p)^{N-k} \approx \frac{z^k \cdot e^{-z}}{k!}$. Note the equality becomes exact in the limit of large N.
5. We see p_k follows a Poisson distribution.

Graphical Model

Three Layer Model



Outline

Spectral Clustering

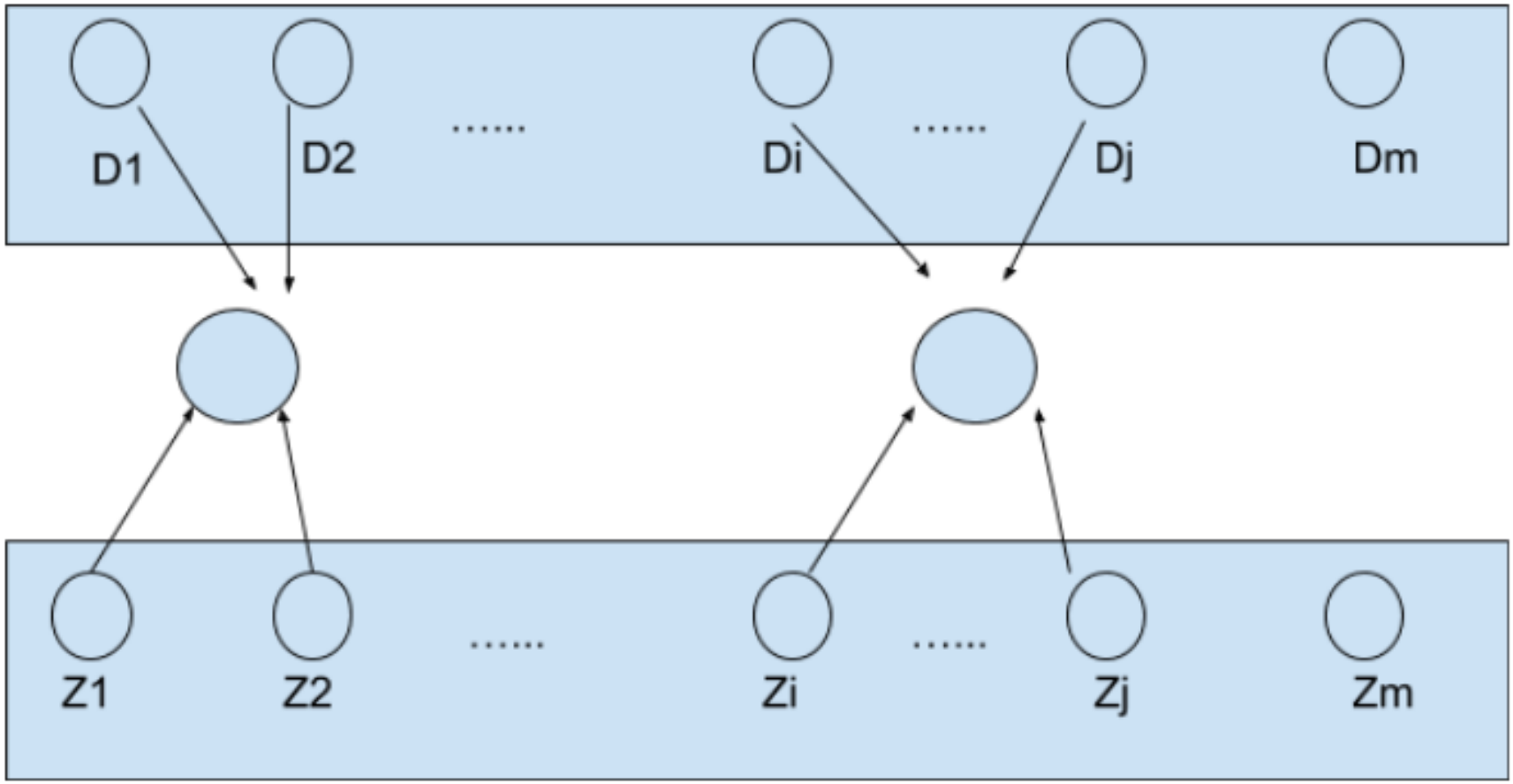
- Consider deriving communities that maximize modularity via a spectral algorithm
- Can show equivalence of this inference method to the single layer model for the case of two communities [Newman]
- **Key Insight:** Can directly compare the single layer model to other derived graphical models when choosing the optimal model for topic modeling.

Inference Algorithm

- Deriving $\overrightarrow{z}$ by performing maximum likelihood on the joint probability of the degree corrected stochastic block model is roughly equivalent to deriving communities (again represented by $\overrightarrow{z}$) that maximize modularity using spectral clustering.
- We use the degree corrected block model as our derived graphical model for approximate inference

Graphical Model

Spectral Clustering



Inference Algorithm

Spectral Clustering

- Construct $L = D^{-1/2}AD^{-1/2}$ where A is the adjacency matrix, D is the diagonal matrix of vertex degrees.
- Derive the eigenvalues and eigenvectors $\lambda_1, \lambda_2, ..., \lambda_k, s_1, ..., s_k$ of L (using Lanczos algorithm)
- Consider s_2 and partition s_2 into k groups clustered by distance, for predetermined cluster size k
- $\overrightarrow{z}$ can be represented as the cluster assignment for node i

Outline

Statistical Assumptions

- We explore the relevance of the bag of words assumption for our given graph based model
  1. Our model makes the assumption that d_i is i.i.d. (for i = 1 to n). Furthermore A_{ij} is i.i.d. $\forall i, j$ conditioned on $\alpha, \beta, \overrightarrow{d}$.
  2. Di Finetti's theorem can be applied with β being the hyperparameter, w_i, w_j co-occuring i.i.d. of other co-occurences
  3. A_{ij} can be treated as an infinitely exchangeable sequence
- Hence, we see words don't co-occur in any specific order across our corpus of documents. Therefore we can assume our model follows the bag of words assumption.

Outline

Algorithm

1. Assume we are given a term document matrix D $(n \times m)$ where n is the size of the vocabulary and m is the number of documents.
2. We assume there exists a matrix $\hat{D}$ that can be decomposed as such: $\hat{D} = AW$ where A is a $n \times r$ matrix with columns representing the topics and w is the $r \times m$ matrix representing topic weights for the given m documents.
3. We attempt to find A, W such that $||D - \hat{D}||_F < \epsilon$ for some predefined ϵ. We assume A, W are non-negative matrices.
4. Hence we recover our topic matrix A and given a new document $\hat{d}_i$, A we can recover the topic weights $\hat{w}_i$ (since $\hat{d}_i = A \cdot \hat{w}_i$).

Generative Process

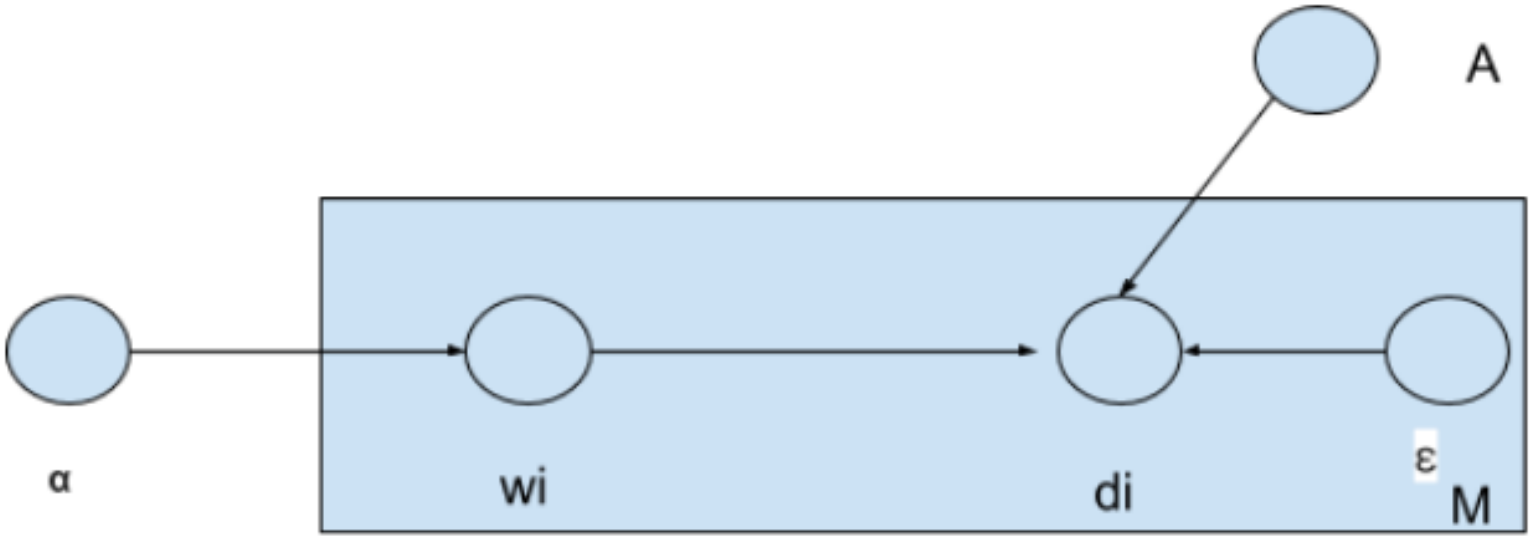
Let A, n (size of the vocabulary), r (number of topics), α (Dirichlet parameter) be fixed.

1. for $i = 1$ to m:
    1. $w_i \sim \text{Dir}(\alpha)$
    2. $\epsilon_i \sim N(0, \sigma^2)$
    3. $d_i = A \cdot w_i + \epsilon_i$

The reason we add noise in our generative process is because the observed term document matrix D may not be factorizable.

Graphical Model

NMF



Outline

Statistical Assumptions

- We assume our topic matrix A is $p-$seperable
- **Seperability** - r rows out of n have only one non-zero value such that if permuted form the $r \times r$ identity matrix (as a submatrix within A)
- Discuss how a non-negative matrix A is invertible (relevant for Algorithm, Inference algorithm)
- Discuss how bag of words assumption holds here

Outline

Problem Constraints

- Recovering non-negative matrices A, W is called NMF and is NP-hard.
- A given observed document d_i is a noisy approximation to the true document $A \cdot w_i$ which represents the weighted sum of topics.

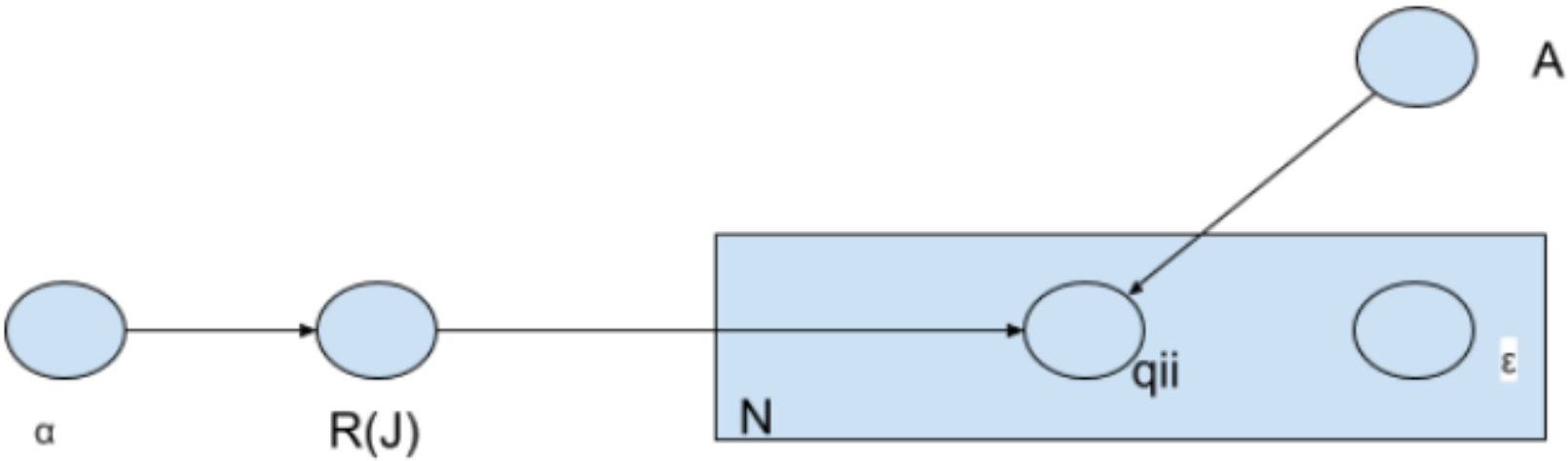
Inference Algorithm

Given non-noisy D

1. We are given D, for which we assume there exists A, W s.t. $D = AW$
2. Let us consider
   $Q = \frac{1}{m} D \cdot D^T \rightarrow Q = \frac{1}{m} AW \cdot W^T A \rightarrow A \cdot \frac{1}{m} WW^T A = AR(\tau)A$
3. Note A is p-seperable and Q is a product of two non-negative matrices namely $A, R(\tau)A^T$
4. We can rewrite $Q = AB$ where $B = R(\tau)A^T$.
5. Let us assume we figure out the r anchor words of A. Then a simple transformation yields $\hat{Q}$ where we permute the the rows of A s.t. the top r rows form the identity matrix, we permute the columns of $R(\tau)A^T$ to match.
6. This allows us to derive $\hat{Q}, \hat{R(\tau)A^T}$ which are permuted versions of the original matrices.
7. We can hence recover A from these matrices.

Graphical Model

NMF Inference Algorithm



Empirical Bayes Approach

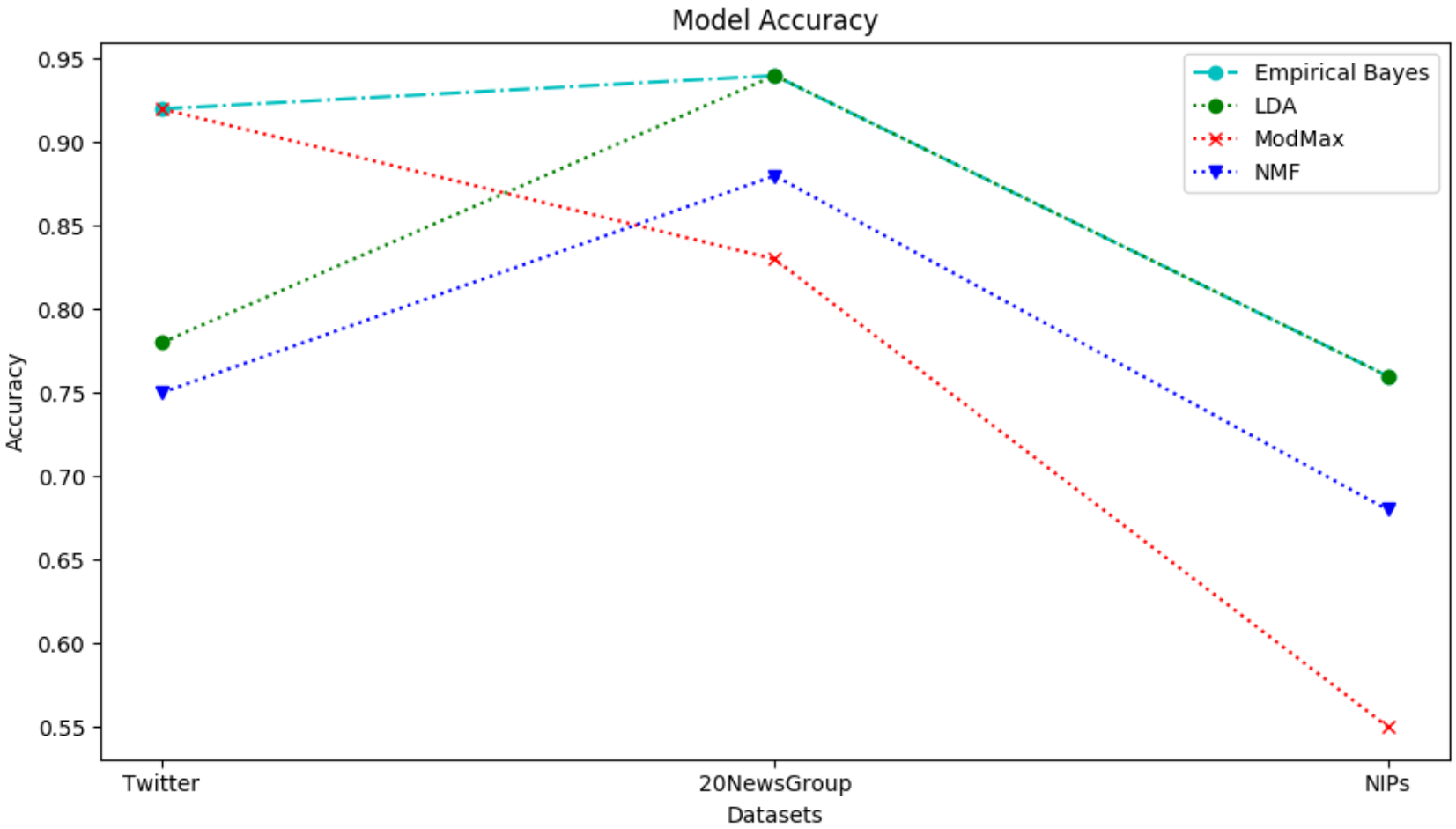
Here we outline how to choose the right algorithm (including the inference process) given the observed variables in our data.

- LDA & variational inference graphical model
- Modularity maximization & spectral clustering graphical model
- Plain LDA graphical model (for theoretical purposes)
- One layer Modularity maximization generative model (for theoretical purposes)
- NMF
- NMF & noise & seperability assumption

Applications & Results

- Use model selection framework to evaluate which of the three models is most suitable given a corpus of data D.
- Model evaluation is done on labeled data using a random forest
- Data
  - Twitter
  - 20 newsgroup
  - NIPs

Applications & Results



Applications & Results