Body-Headline Latent Dirichilet Allocation

Prateek Gupta (UNI: pg2455)

December 16, 2014

Abstract

Topic models have been successfully applied to various types of unstructured datasets. Specifically in the world of natural language processing these tools have found wide applications in uncovering hidden topic distributions in documents. The problem we have considered relates to uncovering hidden structure of a text article at much more granular level as compared to only the article body. We relate the topic distribution of article's headline to the topic distribution of its body. This method can be extended further to paragraph level making it easier to find the topic distributions at each level. We conduct experiments on *The New York Times* corpus to find out *how much the body speaks of its headline*. We also discuss the wide applicability of this model.

1 Introduction

Statistical topic models have been extensively applied to uncover the hidden insights from grouped data like text articles where each article is a collection of words. *Latent Dirichlet Allocation (LDA)* model can help in arranging large unstructured collection of articles according to inferred topics from the model [?].

Traditionally topic models have been applied separately on body [?, ?] and tweets [?] which are short collection of words like headlines. Since headline is crafted in a manner such that it represents main theme of the body content we expect that topics represented in headline should be a subset of topics represented in body. In this work we develop the *Body-Headline Latent Dirichlet Allocation* (BHLDA) model which captures such an influence of body topics on headline topics and vice versa. We expect to get body topics such that headline topics are subset of it. This *subset property* can not be confirmed in *normal LDA* on body and headline owing to the lack of correspondence between headline topics and body topics.

LDA computes article topics from body content but it can be computationally expensive owing to the large number of words in body. The *BHLDA* model can enable the user to compute document topics by using the words from headline, the vocabulary of which is small compared to that of the body, thus ensuring relatively cheap computations. Because of the way the model is designed, word distribution of body topics influences word distribution of headline topics and vice-versa. *Headline Only LDA* can also perform cheap computations but headline topic distribution will be completely independent of body topic distribution. We expect results of topic inference from headline in BHLDA to be more appropriate as compared to *Headline Only LDA*.

The idea of *BHLDA* model extends beyond just the headline and the body. For example, it can be used to infer topics at article's abstract and paragraph level too. Thus, we think of *BHLDA* as a special case of *Microlevel LDA* which can involve applying LDA at various levels of an article. LDA at various levels is possible when there is a structure in grouped data. For example, images and their captions is an example of grouped data with structure [?]. Thus, BHLDA model can also find its application in the domain out of text.

Although the idea of *BHLDA* model was conceived independently of any work done yet, we looked at work that uses similar ideas of graphical models. Literature survey indicates that there have been similar graphical model but applied in entirely different context.

Polylingual topic models [?] is a way to find topics aligned across various language versions of the same article. Graphical model used in the paper is similar to the *BHLDA* model. Structure in grouped data has lots of implication for modeling purposes. Research paper [?] considers this problem in depth and proposed various models along with their upside as well as downsides. Paper discusses various models for correlating topics in images through their pixel level information and image captions.

In this paper we discuss the *BHLDA* model and derive its inference procedure used for parameter estimation. We discuss results using The New York Times Corpus [?]. We compare its performance with

Body Only LDA and *Headline Only LDA*. We wind up by pointing to future work needed to verify the results even more concretely and discuss few applications that this model can promise. We find that *BHLDA* model gives direct correspondence between word distribution in headline topics and body topics which otherwise is not possible in *Headline Only LDA* and *Body Only LDA*. Also the word distribution of headline and body topics are more similar in *BHLDA* model as compared to *normal LDA* on headline and body. In order to enable the reader with access to the code to enable self-experiments we have shifted our work on Github repo 'vewpoint' of 'prateekpg2455'.

2 BHLDA Model

2.1 Notations

Variables in bold letters denote matrix. N is number of words in a document. $\hat{N}$ is number of words in headline. D is number of documents in the corpus. V is number of words in vocabulary. K is number of topics. $\mathbf{W} = [w_{j,t}]_{j=1,t=1}^{D,N}$ represents body word matrix where $w_{j,t}$ is the t^{th} body word in j^{th} document. $\hat{\mathbf{W}} = [w_{j,\hat{t}}]_{j=1,\hat{t}=1}^{D,\hat{N}}$ where $w_{j,\hat{t}}$ is the $\hat{t}^{th}$ headline word in j^{th} document. $\mathbf{Z} = [z_{j,t}]_{j=1,t=1}^{D,N}$ represents matrix of topic allocation to words in the body where $z_{j,t}$ is topic allocation to t^{th} body word in j^{th} document. It is of same dimension as $\mathbf{W}$. Similar description holds for $\hat{\mathbf{Z}} = [z_{j,\hat{t}}]_{j=1,\hat{t}=1}^{D,\hat{N}}$. $\mathbf{\Theta} = [\theta_{j,i}]_{j=1,i=1}^{D,N}$ represents i^{th} topic proportion of j^{th} document/article. θ_j represents topic distribution vector for a document. $\mathbf{\Psi} = [\psi_{i,r}]_{i=1,r=1}^{K,V}$ where $\psi_{i,r}$ is probability of choosing r^{th} word for body from i^{th} body topic. ψ_i is the multinomial distribution vector for i^{th} topic. $\hat{\mathbf{\Psi}} = [\hat{\psi}_{i,r}]_{i=1,r=1}^{K,V}$ where $\hat{\psi}_{i,r}$ is probability of choosing r^{th} word for headline from i^{th} headline topic. Detailed description of variables, their size and representation in the model is given in Appendix 8.1.

2.2 Generative Model

Topic distribution for document j is generated from Dirichlet distribution with parameter α. Given the topic distribution θ_j, N topic allocations for body are sampled from multinomial distribution with parameter θ_j. Similarly, $\hat{N}$ topic allocations are sampled for headline. For each topic allocation sampled, a word is generated from that topic. Thus, this model captures the fact that topic allocations for body and headline can differ because of large number of words in the body as compared to the headline.

The *Body-Headline LDA* (BHLDA) model, shown in Figure 1 assumes following generative process:

1. Sample K Dirichlet random variables for body topic distribution, $\psi_i \sim$ *Dir(*β*)*
2. Sample K Dirichilet random variables for headline topic distribution, $\hat{\psi}_i \sim$ *Dir(*$\hat{\beta}$*)*
3. For each document j, sample a Dirichlet random variable, $\theta_j \sim$ *Dir(*α*)*
   (a) For each word $w_{j,t}$ in body
      i. Sample a body topic, $z_{j,t} \sim$ *Mult(*θ_j*)*
      ii. Sample a body word, $w_{j,t} \sim$ *Mult(*ψ_z*)*
   (b) For each word $\hat{w}_{j,\hat{t}}$ in headline
      i. Sample a headline topic, $\hat{z}_{j,\hat{t}} \sim$ *Mult(*θ_j*)*
      ii. Sample a headline word, $\hat{w}_{j,\hat{t}} \sim$ *Mult(*$\hat{\psi}_{\hat{z}}$*)*

2.3 Graphical Model

Model specifying the relation between variables is shown in Figure 1. Graphical model assumes that topic allocations $\hat{z}$ in headline and z in body influence each other through topic distribution θ. Sampling of words in body and headline are dependent on topic distribution θ_j. Thus topic allocations in the body are influenced by topic allocations in the headline and vice versa.

Topic distribution over one article does not depend on topic distribution of other documents implying exchangeability within θ_j vectors. Similarly, words within body are exchangeable and so are words within headline.

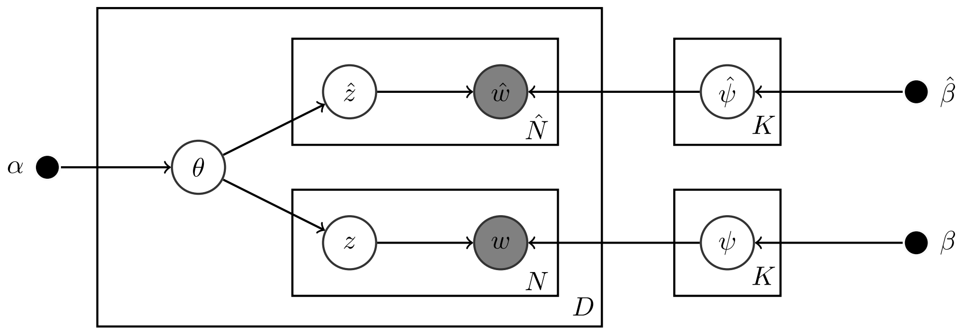

Figure 1: BHLDA Graphical Model

V-structure formed at ψ and $\hat{\psi}$ assures that influence from topic allocations flows from z or $\hat{z}$ to ψ and $\hat{\psi}$. Since there is an influence from z to $\hat{z}$, ψ and $\hat{\psi}$ influences each other which generates correspondence between headline topics and body topics.

2.4 Joint Distribution

The resulting joint distribution on body words, headline words and latent variables is given by

$$P(\mathbf{\Psi}, \hat{\mathbf{\Psi}}, \mathbf{\Theta}, \mathbf{Z}, \hat{\mathbf{Z}}, \mathbf{W}, \hat{\mathbf{W}}; \alpha, \beta) = \prod_{i=1}^{K} P(\psi_i \mid \beta) \prod_{i=1}^{K} P(\hat{\psi}_i \mid \hat{\beta}) \prod_{j=1}^{D} P(\theta_j \mid \alpha) \prod_{t=1}^{N} P(z_{j,t} \mid \theta_j) P(w_{j,t} \mid \psi_{z_{j,t}}) \prod_{\hat{t}=1}^{\hat{N}} P(\hat{z}_{j,t} \mid \theta_j) P(\hat{w}_{j,t} \mid \hat{\psi}_{\hat{z}_{j,t}}) \quad (1)$$

Reader is advised to refer Appendix 8.2 for detailed derivation of above equation.

3 Inference and Estimation

Collapsed Gibbs Sampling for the model is derived in Appendix 8.2. Update equations derived are as follows:

1. For body words,

$$P(z_{m,n} = k \mid \mathbf{W}, \hat{\mathbf{W}}, \mathbf{Z}_{-(\mathbf{m},\mathbf{n})}, \hat{\mathbf{Z}}; \alpha, \beta) \propto \left(\alpha_k + n_{m,(.)}^{k,-(m,n)} + \hat{n}_{m,(.)}^{k}\right) \times \left(\frac{\beta_\nu + n_{(.),\nu}^{k,-(m,n)}}{\sum_{r=1}^{V}(\beta_r + n_{(.),r}^{k,-(m,n)})}\right)$$

2. For headline words,

$$P(\hat{z}_{m,n} = k \mid \mathbf{W}, \hat{\mathbf{W}}, \mathbf{Z}, \hat{\mathbf{Z}}_{-(\mathbf{m},\mathbf{n})}; \alpha, \beta) \propto \left(\alpha_k + n_{m,(.)}^{k} + \hat{n}_{m,(.)}^{k,-(m,n)}\right) \times \left(\frac{\hat{\beta}_\nu + \hat{n}_{(.),\nu}^{k,-(m,n)}}{\sum_{r=1}^{V}(\hat{\beta}_r + \hat{n}_{(.),r}^{k,-(m,n)})}\right)$$

Above equations implies that conditional distribution of n^{th} word in the m^{th} document belonging to k^{th} topic is directly proportional to α_k and number of words other than $w_{m,n}$ that belong to k^{th} topic in m^{th} document and it befits intuition. Interestingly enough, conditional probability of $z_{m,n} = k$ is also influenced by number of headline words belonging to k^{th} topic. Thus, model supports the relation between headline topics and body topics as expected. Assuming $w_{m,n} = \nu$, conditional probability of $z_{m,n} = k$ is proportional to number of times ν has been allocated to k^{th} topic across the corpus except $w_{m,n}$. Similar conclusions can be drawn for conditional probability of $\hat{z}_{m,n} = k$.

Conditional probability of $z_{m,n} = k$ is in terms of present state of topic allocations except the $z_{m,n}^{th}$ term which is atypical of Gibbs Sampling method.

4 Data

We used *The New York Times Annotated Corpus* [?] to test our model. The corpus contains 1.8 million articles. Each article is a collection of article text and related metadata like author name, editor name, publication date, online section of published article, newspaper section as well as page number and column number of published article, manually annotated tags, headline and much more. Each article is referred to by its unique id. Detailed process of preprocessing steps and issues dealt with such a huge corpus are discussed in Appendix 8.3

5 Results

The *BHLDA* model was run on 100,000 articles of *The New York Times Corpus*. Numbers of articles were chosen such that the model can handle computations of Gibbs Sampling procedure. Appendix 8.5.1 shows the results of BHLDA model. Appendix 8.5.2 and 8.5.3 displays output of *Body Only LDA* and *Headline Only LDA*.

One direct advantage of *BHLDA* model is that the output word distribution for each topic has one to one correspondence between body and headline. This is not possible in case of normal LDA run on only body or only headline. This result is shown in Figure 2.

```
                                  Topic 0
-----------------------------------------------------------------------------------------
                   BODY                  |                     HEADLINE
-----------------------------------------------------------------------------------------
       Word         |        Prob        i|!   Word             |        Prob
-----------------------------------------------------------------------------------------
       film         |        0.0234137   i|!   review/film      |        0.0300011
       television   |        0.011184    i|!   mail             |        0.0287416
       movie        |        0.00985363  i|!   answering        |        0.0243902
       race         |        0.00623159  i|!   review/television|        0.021642
       num          |        0.00584287  i|!   home             ||       0.0198099
       time         |        0.00445975  i|!   film             |        0.0153441
       video        |        0.00441605  i|!   movies           |        0.0127104
```

Figure 2: Snippet of Results

Uniqueness of topics in *BHLDA* is evident from the Table 1. For each topic, there exists unique topic in *BHLDA* while there are potential matches in *Body Only LDA* and *Headline Only LDA*. It is also evident from the results in section 8.5.1.1, 8.5.2, 8.5.3 by manually reading words in word distributions.

A sample of final topic allocations to articles is shown in Appendix 8.5.1.2. Topic allocations to headline words are sparse as compared to topic allocations to body words. It follows from intuition because of the presence of relatively fewer headline words as compared to the body words. Topic allocations in *Headline Only LDA* are also sparse but are completely independent of body words. As far as the sparsity is concerned headline topic distribution in *BHLDA* is not as sparse as in *Headline Only LDA*. As an evidence, across 100,000 articles, there were 91,268 articles in which number of non-zero entries in headline topic distribution for *BHLDA* was more than that for *Headline Only LDA*. This means *BHLDA* does loose on sparsity for headline topic distributions.

Kulback-Leibler Divergence of two multinomial distribution is defined as $KL(p||q) = \sum_{i=1} p_i \times ln(\frac{p_i}{q_i})$, where terms with $p_i = 0$ are simply put as 0 because $\lim_{x\to 0} x \times ln(x) = 0$. Please note that KL Divergence considers only the support of q for its computation. Only few topics from normal LDA on body and headline were found to be similar. For example, correspondence between body LDA and headline LDA can be drawn through topic 0 and topic 5 respectively. Similar topic can be found in BHLDA model at topic 14. Table 1 displays such pairs of topics and corresponding KL Divergence for both the models.

It is interesting to note that word distribution for body topics and headline topics in *BHLDA* are more closer i.e. they have lower average KL Divergence as compared to *normal LDA* on body and headline only. At the same time, uniqueness of topics in *BHLDA* can also be confirmed from the Table 1.

Majority of headline topics were found to be a subset of body topics. Results indicate that there were 2,658 articles in which headline topics were not subset of body topics. Thus, 97.342% times headline topics were subset of body topics.

	LDA		BHLDA	
Topic	p,q	$KL(body_p \| \| headline_q)$	z	$KL(body_z \| \| headline_z)$
National Budget	0,5	0.3599	14	0.1827
International Affairs	13,7	0.2749	7	0.2109
Dance/Theater/Music	14,0/14,6	0.2692/0.3037	6	0.0280
Art/Music/Travel	19,0/19,6	0.0598/0.0856	0	-0.0143
Markets/Finance	18,1/18,2/18,12	0.1498/0.3177/-0.0372	4/9	0.3614/0.4187
Sports	17,11/16,11	0.3976/0.3354	5/15	0.1963/0.2360
	Average	0.2103		0.1608

Table 1: KL Divergence between word distributions for LDA and BHLDA. p is topic number in *Body only LDA* and q is corresponding topic number in *Headline only LDA*. z is corresponding topic number in *BHLDA*. $body_*$ and $headline_*$ are word distribution of topics. Matching of topics in LDA is subjective. Average includes minimum values in each observation. Negative values of KL Divergence are because of inadequate storage of float values. Note that there can be multiple matches for each Topic. They are separated by '/'.

6 Future Work

Present results were obtained using uniform priors on β and α. More informative priors can improve the results drastically. As per the figure 3, it is quite evident that word distribution for headline and body differ a lot. As per the plot 3, cumulative distribution function differ a lot for body words and headline words. Thus, β and $\hat{\beta}$ should be initialized differently for better results.

Problem of aligning word distribution of *Body Only LDA* with *Headline Only LDA* can be addressed by seeding *Headline Only LDA* with the results from *Body Only LDA*. Thus, results on KL Divergence can be verified even further after running LDA this way.

Running BHLDA based on *stochastic variational inference* can speed up the computations and hence can be used to verify the results based on even larger number of articles. Present work uses Collapsed Gibbs Sampling method for inference and estimation and hence we cannot scale up the model to larger number of articles.

Generation of headline words just from body words and body words just from headline words can be another useful metric to compare normal LDA and BHLDA. Thus, mathematical formulation to compute maximum likelihood estimates of conditional distribution can aid in performing this analysis. Metrics like perplexity and posterior predictive checks can also help in verifying the results of *BHLDA* model.

7 BHLDA: Applications and extension

News article recommendation is a common problem in many news organizations that helps them in keeping the reader engaged to their websites. It is a common behavior of any reader to infer topics from the headline and then decide whether to read an article or not. Thus, ability to infer topics at various levels of the text article which is facilitated by *BHLDA* model can be useful in recommendation engines. For example, a recommendation engine with features as topics inferred from headline text can take into account reading behavior of readers. As discussed, headline topics inferred in this manner are more appropriate as compared to *Headline Only LDA* which can enable the recommendation engine to incorporate features from headline topics that reflect body topics more closely. Thus, although recommendation engine includes features only from headline it is able to incorporate body features too and hence recommendation is influenced by body and headline topics using computations involving only the headline words.

Summarizing an article is the problem that deals with identifying main theme in the article and displaying that as description of the article. Some news organization just displays first 3 sentences of the article along with headlines to give the flavor of an article. An article contains multiple paragraphs. Each paragraph represents different topic distribution. *Microlevel LDA* can help in this problem. LDA at each level of article i.e. paragraphs can give topic allocations of all the paragraphs. Assuming that paragraph with topic allocations similar to headline topic allocations are more reflective of headline content is true, summary of article can be given by paragraph with similar topic distribution as that of headline.

It enable cheaper computation of article topics using only the headline words. Owing to the fewer number of headline words as compared to body words, computation of topics from headline words is relatively cheaper as compared to that from body words and *BHLDA* model enables appropriate representation of topics influenced also by the body words.

The *BHLDA* model can be useful in establishing credibility of news publications. Some news sources have headlines crafted in a manner to attract more readers and does not reflect properly the contents of an article. Average KL Divergence of news sources can be used as a metric to quantify quality of news sources.

8 Appendix

8.1 Notations

8.1.1 Observed Data

As discussed in Appendix 8.3 words are represented by unique numbers for modeling purpose. Thus observed data, for purpose of our modeling, is represented by vector $\mathbf{W}$ for body and $\hat{\mathbf{W}}$ for headline.

$$\mathbf{W} = \begin{bmatrix} w_{11} & w_{12} & ...w_{1N} \\ w_{21} & w_{22} & ...w_{2N} \\ ... & & \\ ... & & \\ ... & & \\ w_{D1} & w_{D2} & ...w_{DN} \end{bmatrix} \quad \hat{\mathbf{W}} = \begin{bmatrix} \hat{w}_{11} & \hat{w}_{12} & ...\hat{w}_{1\hat{N}} \\ \hat{w}_{21} & \hat{w}_{22} & ...\hat{w}_{2\hat{N}} \\ ... & & \\ ... & & \\ ... & & \\ \hat{w}_{D1} & \hat{w}_{D2} & ...\hat{w}_{D\hat{N}} \end{bmatrix}$$

where $w_{jt}, \hat{w}_{j\hat{t}} \in \{1, 2, 3...V\}$, V denotes number of words in vocabulary, N denotes number of words in document and $\hat{N}$ denotes number of words in headline.

8.1.2 Latent Variables

Figure 1 consists of various latent variables in the model. These variables are represented in a certain way for the purpose of our model representation. This section gives the view into how are the variables represented for the purpose of our code.

Latent Z variable represents association of words to a topic. It has same dimensions as of W but the values taken by its elements are one of the K topics.

$$\mathbf{Z} = \begin{bmatrix} z_{11} & z_{12} & ...z_{1N} \\ z_{21} & z_{22} & ...z_{2N} \\ ... & & \\ ... & & \\ ... & & \\ z_{D1} & z_{D2} & ...z_{DN} \end{bmatrix} \quad \hat{\mathbf{Z}} = \begin{bmatrix} \hat{z}_{11} & \hat{z}_{12} & ...\hat{z}_{1\hat{N}} \\ \hat{z}_{21} & \hat{z}_{22} & ...\hat{z}_{2\hat{N}} \\ ... & & \\ ... & & \\ ... & & \\ \hat{z}_{D1} & \hat{z}_{D2} & ...\hat{z}_{D\hat{N}} \end{bmatrix}$$

where $z_{jt}, \hat{z}_{j\hat{t}} \in \{1, 2, 3...K\}$, K denotes number of topics and is specified by the user.

Topic distribution of each documentθ is the hidden structure and represents the multinomial probability associated with the document. It can be interpreted as proportion of document representing i^{th} topic.

$$\mathbf{\Theta} = \begin{bmatrix} \theta_{11} & \theta_{12} & ...\theta_{1K} \\ \theta_{21} & \theta_{22} & ...\theta_{2K} \\ ... & & \\ ... & & \\ ... & & \\ \theta_{D1} & \theta_{D2} & ...\theta_{DK} \end{bmatrix}$$

where $\theta_{ij} \in [0, 1], \sum_{j=1}^{K} \theta_{ij} = 1$, K denotes number of topics and is specified by the user.

Probability distribution over words ψ_i for each topic represents likelihood of word associated to that topic.

$$\mathbf{\Psi} = \begin{bmatrix} \psi_{11} & \psi_{12} & ...\psi_{1V} \\ \psi_{21} & \psi_{22} & ...\psi_{2V} \\ ... & & \\ ... & & \\ ... & & \\ \psi_{K1} & \psi_{K2} & ...\psi_{KV} \end{bmatrix} \quad \hat{\mathbf{\Psi}} = \begin{bmatrix} \hat{\psi}_{11} & \hat{\psi}_{12} & ...\hat{\psi}_{1V} \\ \hat{\psi}_{21} & \hat{\psi}_{22} & ...\hat{\psi}_{2V} \\ ... & & \\ ... & & \\ ... & & \\ \hat{\psi}_{D1} & \hat{\psi}_{D2} & ...\hat{\psi}_{DV} \end{bmatrix}$$

where $\psi_{ir}, \hat{\psi}_{ir} \in [0, 1], \sum_{r=1}^{V} \psi_{ir} = 1, \sum_{r=1}^{V} \hat{\psi}_{ir} = 1$, V denotes number of words in vocabulary.

Priors in the model are represented by α, β and $\hat{\beta}$. Each of these latent variables are represented by vectors.

$$\alpha = \begin{bmatrix} \alpha_1 & \alpha_2 & ... & \alpha_K \end{bmatrix}$$
$$\beta = \begin{bmatrix} \beta_1 & \beta_2 & ... & \beta_V \end{bmatrix}$$
$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 & \hat{\beta}_2 & ... & \hat{\beta}_V \end{bmatrix}$$

8.2 Model Derivation

Throughout the derivation following notations are used:

- i for i^{th} topic, $i \in \{1, 2, 3...K\}$
- j for j^{th} document, $j \in \{1, 2, 3...D\}$
- r for r^{th} word in vocabulary, $r \in \{1, 2, 3...V\}$
- t for t^{th} word in a document, $t \in \{1, 2, 3...N\}$

Graphical model in Figure 1 suggests following joint distribution,

$$P(\mathbf{\Psi}, \hat{\mathbf{\Psi}}, \mathbf{\Theta}, \mathbf{Z}, \hat{\mathbf{Z}}, \mathbf{W}, \hat{\mathbf{W}}; \alpha, \beta) = P(\mathbf{\Psi} \mid \beta) P(\hat{\mathbf{\Psi}} \mid \hat{\beta}) P(\mathbf{\Theta} \mid \alpha) P(\mathbf{Z} \mid \mathbf{\Theta}) P(\mathbf{W} \mid \mathbf{Z}) P(\hat{\mathbf{Z}} \mid \hat{\mathbf{\Theta}}) P(\hat{\mathbf{W}} \mid \hat{\mathbf{Z}})$$

Following the independence assumptions implicit in the graphical model,

$$P(\mathbf{\Psi}, \hat{\mathbf{\Psi}}, \mathbf{\Theta}, \mathbf{Z}, \hat{\mathbf{Z}}, \mathbf{W}, \hat{\mathbf{W}}; \alpha, \beta) = \prod_{i=1}^{K} P(\psi_i \mid \beta) \times \prod_{i=1}^{K} P(\hat{\psi}_i \mid \hat{\beta}) \times \prod_{j=1}^{D} P(\theta_j \mid \alpha) \times \prod_{t=1}^{N} P(z_{j,t} \mid \theta_j) P(w_{j,t} \mid \psi_{z_{j,t}})$$
$$\times \prod_{\hat{t}=1}^{\hat{N}} P(\hat{z}_{j,t} \mid \theta_j) P(\hat{w}_{j,t} \mid \hat{\psi}_{\hat{z}_{j,t}})$$

For the purpose of Gibbs sampling required distribution of words and associated topics which is given by,

$$P(\mathbf{Z}, \hat{\mathbf{Z}}, \mathbf{W}, \hat{\mathbf{W}}; \alpha, \beta) = \int_{\mathbf{\Theta}} \int_{\mathbf{\Psi}} \int_{\hat{\mathbf{\Psi}}} \prod_{i=1}^{K} P(\psi_i \mid \beta) \prod_{i=1}^{K} P(\hat{\psi}_i \mid \hat{\beta}) \prod_{j=1}^{D} P(\theta_j \mid \alpha) \prod_{t=1}^{N} P(z_{j,t} \mid \theta_j) P(w_{j,t} \mid \psi_{z_{j,t}})$$
$$\times \prod_{\hat{t}=1}^{\hat{N}} P(\hat{z}_{j,t} \mid \theta_j) P(\hat{w}_{j,t} \mid \hat{\psi}_{\hat{z}_{j,t}}) \, d\mathbf{\Theta} \, d\mathbf{\Psi} \, d\hat{\mathbf{\Psi}}$$

$$= \int_{\mathbf{\Psi}} \prod_{i=1}^{K} P(\psi_i \mid \beta) \prod_{j=1}^{D} \prod_{t=1}^{N} P(w_{j,t} \mid \psi_{z_{j,t}}) \, d\mathbf{\Psi} \int_{\hat{\mathbf{\Psi}}} \prod_{i=1}^{K} P(\hat{\psi}_i \mid \hat{\beta}) \prod_{j=1}^{D} \prod_{\hat{t}=1}^{\hat{N}} P(\hat{w}_{j,\hat{t}} \mid \hat{\psi}_{\hat{z}_{j,\hat{t}}}) \, d\hat{\mathbf{\Psi}}$$
$$\int_{\mathbf{\Theta}} \prod_{j=1}^{D} P(\theta_j \mid \alpha) \prod_{t=1}^{N} P(Z_{j,t} \mid \theta_j) \prod_{\hat{t}=1}^{\hat{N}} P(\hat{z}_{j,\hat{t}} \mid \theta_j) \, d\mathbf{\Theta}$$

Using the argument of exchangeability we can move product outside integration.

$$= \prod_{i=1}^{K} \int_{\psi_i} P(\psi_i \mid \beta) \prod_{j=1}^{D} \prod_{t=1}^{N} P(w_{j,t} \mid \psi_{z_{j,t}})\, d\psi_i \times \prod_{i=1}^{K} \int_{\hat{\psi}_i} P(\hat{\psi}_i \mid \hat{\beta}) \prod_{j=1}^{D} \prod_{\hat{t}=1}^{\hat{N}} P(\hat{w}_{j,\hat{t}} \mid \hat{\psi}_{\hat{z}_{j,\hat{t}}})\, d\hat{\psi}_i$$

$$\times \prod_{j=1}^{D} \int_{\theta_j} P(\theta_j \mid \alpha) \prod_{t=1}^{N} P(Z_{j,t} \mid \theta_j) \prod_{\hat{t}=1}^{\hat{N}} P(\hat{z}_{j,\hat{t}} \mid \theta_j)\, d\theta_j$$

Following step relies on the fact that Dirichlet distribution is *conjugate prior* to multinomial distribution. Also it uses the fact that integral of Dirichlet distribution is 1.

$$\begin{aligned}
\int_{\psi_i} P(\psi_i \mid \beta) \prod_{j=1}^{D} \prod_{t=1}^{N} P(w_{j,t} \mid \psi_{z_{j,t}})\, d\psi_i &= \int_{\psi_i} \frac{\Gamma(\sum_{r=1}^{V} \beta_r)}{\prod_{r=1}^{V} \Gamma(\beta_r)} \prod_{r=1}^{V} (\psi_{i,r})^{\beta_r - 1} \prod_{r=1}^{V} (\psi_{i,r})^{n^i_{(.),r}}\, d\psi_i \\
&= \int_{\psi_i} \frac{\Gamma(\sum_{r=1}^{V} \beta_r)}{\prod_{r=1}^{V} \Gamma(\beta_r)} \prod_{r=1}^{V} (\psi_{i,r})^{n^i_{(.),r} + \beta_r - 1}\, d\psi_i \\
&= \frac{\Gamma(\sum_{r=1}^{V} \beta_r)}{\prod_{r=1}^{V} \Gamma(\beta_r)} \times \frac{\prod_{r=1}^{V} \Gamma(n^i_{(.),r} + \beta_r)}{\Gamma(\sum_{r=1}^{V} n^i_{(.),r} + \beta_r)}
\end{aligned}$$

where $n^i_{j,r}$ denotes number of r^{th} body word in j^{th} document that were assigned to i^{th} topic. $n^i_{(.),r}$ denotes number of r^{th} body word across the corpus that were assigned to i^{th} topic.

Similar derivation can be done for headline topic distribution. Thus,

$$\int_{\hat{\psi}_i} P(\hat{\psi}_i \mid \hat{\beta}) \prod_{j=1}^{D} \prod_{\hat{t}=1}^{\hat{N}} P(\hat{w}_{j,t} \mid \hat{\psi}_{\hat{z}_{j,\hat{t}}})\, d\hat{\psi}_i = \frac{\Gamma(\sum_{r=1}^{V} \hat{\beta}_r)}{\prod_{r=1}^{V} \Gamma(\hat{\beta}_r)} \times \frac{\prod_{r=1}^{V} \Gamma(\hat{n}^i_{(.),r} + \hat{\beta}_r)}{\Gamma(\sum_{r=1}^{V} \hat{n}^i_{(.),r} + \hat{\beta}_r)}$$

where $\hat{n}^i_{j,r}$ denotes number of r^{th} headline word in j^{th} document that were assigned to i^{th} topic. $\hat{n}^i_{(.),r}$ denotes number of r^{th} headline word across the corpus that were assigned to i^{th} topic.

Applying similar logic to the third integral, we get

$$\begin{aligned}
&\int_{\theta_j} P(\theta_j \mid \alpha) \prod_{t=1}^{N} P(z_{j,t} \mid \theta_j) \prod_{\hat{t}=1}^{\hat{N}} P(\hat{z}_{j,\hat{t}} \mid \theta_j) \\
&= \int_{\theta_j} \frac{\Gamma(\sum_{i=1}^{K} \alpha_i)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \prod_{i=1}^{K} \theta_{j,i}^{\alpha_i - 1} \prod_{i=1}^{K} \theta_{j,i}^{n^i_{j,(.)}} \prod_{i=1}^{K} \theta_{j,i}^{\hat{n}^i_{j,(.)}} \\
&= \int_{\theta_j} \frac{\Gamma(\sum_{i=1}^{K} \alpha_i)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \prod_{i=1}^{K} \theta_{j,i}^{n^i_{j,(.)} + \hat{n}^i_{j,(.)} + \alpha_i - 1} \\
&= \frac{\Gamma(\sum_{i=1}^{K} \alpha_i)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \times \frac{\prod_{i=1}^{K} \Gamma(n^i_{j,(.)} + \hat{n}^i_{j,(.)} + \alpha_i)}{\Gamma(\sum_{i=1}^{K} n^i_{j,(.)} + \hat{n}^i_{j,(.)} + \alpha_i)}
\end{aligned}$$

where $n^i_{j,(.)}$ denotes total number of body words in j^{th} document that are assigned to i^{th} body topic and $\hat{n}^i_{j,(.)}$ denotes total number of headline words in j^{th} document that are assigned to i^{th} headline topic.

Putting everything together,

$$\begin{aligned}
P(\mathbf{Z}, \hat{\mathbf{Z}}, \mathbf{W}, \hat{\mathbf{W}}; \alpha, \beta) &= \prod_{i=1}^{K} \left[\frac{\Gamma(\sum_{r=1}^{V} \beta_r)}{\prod_{r=1}^{V} \Gamma(\beta_r)} \times \frac{\prod_{r=1}^{V} \Gamma(n^i_{(.),r} + \beta_r)}{\Gamma(\sum_{r=1}^{V} n^i_{(.),r} + \beta_r)} \right] \prod_{i=1}^{K} \left[\frac{\Gamma(\sum_{r=1}^{V} \hat{\beta}_r)}{\prod_{r=1}^{V} \Gamma(\hat{\beta}_r)} \times \frac{\prod_{r=1}^{V} \Gamma(\hat{n}^i_{(.),r} + \hat{\beta}_r)}{\Gamma(\sum_{r=1}^{V} \hat{n}^i_{(.),r} + \hat{\beta}_r)} \right] \\
&\quad \times \prod_{j=1}^{D} \left[\frac{\Gamma(\sum_{i=1}^{K} \alpha_i)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \times \frac{\prod_{i=1}^{K} \Gamma(n^i_{j,(.)} + \hat{n}^i_{j,(.)} + \alpha_i)}{\Gamma(\sum_{i=1}^{K} n^i_{j,(.)} + \hat{n}^i_{j,(.)} + \alpha_i)} \right] \\
&= \left[\frac{\Gamma(\sum_{r=1}^{V} \beta_r)}{\prod_{r=1}^{V} \Gamma(\beta_r)} \right]^K \left[\frac{\Gamma(\sum_{r=1}^{V} \hat{\beta}_r)}{\prod_{r1=1}^{V} \Gamma(\hat{\beta}_r)} \right]^K \left[\frac{\Gamma(\sum_{i=1}^{K} \alpha_i)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \right]^D \times \prod_{i=1}^{K} \left[\frac{\prod_{r=1}^{V} \Gamma(n^i_{(.),r} + \beta_r)}{\Gamma(\sum_{r=1}^{V} n^i_{(.),r} + \beta_r)} \right] \\
&\quad \times \prod_{i=1}^{K} \left[\frac{\prod_{r=1}^{V} \Gamma(\hat{n}^i_{(.),r} + \hat{\beta}_r)}{\Gamma(\sum_{r=1}^{V} \hat{n}^i_{(.),r} + \hat{\beta}_r)} \right] \times \prod_{j=1}^{D} \left[\frac{\prod_{i=1}^{K} \Gamma(n^i_{j,(.)} + \hat{n}^i_{j,(.)} + \alpha_i)}{\Gamma(\sum_{i=1}^{K} n^i_{j,(.)} + \hat{n}^i_{j,(.)} + \alpha_i)} \right]
\end{aligned}$$

By Bayes Theorem,

$$P(z_{m,n}=k \mid \mathbf{W},\hat{\mathbf{W}},\mathbf{Z}_{-(\mathbf{m},\mathbf{n})},\hat{\mathbf{Z}};\alpha,\beta) \propto P(z_{m,n}=k,\mathbf{W},\hat{\mathbf{W}},\mathbf{Z}_{-(\mathbf{m},\mathbf{n})},\hat{\mathbf{Z}};\alpha,\beta)$$
$$\propto \prod_{i=1}^{K}\left[\prod_{r\neq\nu}^{V}\Gamma(n^i_{(.),r}+\beta_r)\right]\times\prod_{i=1}^{K}\left[\frac{\prod_{r=1}^{V}\Gamma(\hat{n}^i_{(.),r}+\hat{\beta}_r)}{\Gamma(\sum_{r=1}^{V}\hat{n}^i_{(.),r}+\hat{\beta}_r)}\right]\times\prod_{j\neq m}^{D}\left[\frac{\prod_{i=1}^{K}\Gamma(n^i_{j,(.)}+\hat{n}^i_{j,(.)}+\alpha_i)}{\Gamma(\sum_{i=1}^{K}n^i_{j,(.)}+\hat{n}^i_{j,(.)}+\alpha_i)}\right]\times$$
$$\left[\frac{\prod_{i=1}^{K}\Gamma(n^i_{m,(.)}+\hat{n}^i_{j,(.)}+\alpha_i)}{\Gamma(\sum_{i=1}^{K}n^i_{j,(.)}\hat{n}^i_{j,(.)}+\alpha_i)}\right]\times\prod_{i=1}^{K}\left[\frac{\Gamma(n^i_{(.),\nu}+\beta_\nu)}{\Gamma(\sum_{r=1}^{V}n^i_{(.),r}+\beta_r)}\right]$$

Above follows because $\sum_{i=1}^{K}n^i_{j,(.)}+\sum_{i=1}^{K}\hat{n}^i_{j,(.)}=N+\hat{N}$ for a single document. Also note that $w_{(m,n)}=\nu$. We try to completely separate out effects of $z_{m,n}$ from the joint distribution. Thus,

$$P(z_{m,n}=k \mid \mathbf{W},\hat{\mathbf{W}},\mathbf{Z}_{-(\mathbf{m},\mathbf{n})},\hat{\mathbf{Z}};\alpha,\beta) \propto \prod_{i=1}^{K}\Gamma(n^i_{m,(.)}+\hat{n}^i_{m,(.)}+\alpha_i)\times\prod_{i=1}^{K}\frac{\Gamma(n^i_{(.),\nu}+\beta_\nu)}{\Gamma(\sum_{r=1}^{V}\beta_r+n^i_{(.),r})}$$

Because $z_{m,n}=k$, $n^i_{m,(.)}=n^{i,-(m,n)}_{m,(.)}+1$ and $n^k_{(.),\nu}=n^{k,-(m,n)}_{(.),\nu}+1$. Thus,

$$P(z_{m,n}=k \mid \mathbf{W},\hat{\mathbf{W}},\mathbf{Z}_{-(\mathbf{m},\mathbf{n})},\hat{\mathbf{Z}};\alpha,\beta) \propto \prod_{i\neq k}\Gamma(n^{i,-(m,n)}_{m,(.)}+\hat{n}^i_{m,(.)}+\alpha_i)\prod_{i\neq k}\left[\frac{\Gamma(n^{i,-(m,n)}_{(.),\nu}+\beta_\nu)}{\Gamma(\beta_r+\sum_{r=1}^{V}n^{k,-(m,n)}_{(.),r})}\right]$$
$$\times\Gamma(\alpha_k+n^{k,-(m,n)}_{m,(.)}+1+\hat{n}^k_{m,(.)})\times\frac{\Gamma(n^{k,-(m,n)}_{(.),\nu}+\beta_\nu+1)}{\Gamma(\sum_{r=1}^{V}(n^{k,-(m,n)}_{(.),r}+\beta_r)+1)}$$

Using the property of Gamma function, $\Gamma(x+1)=x\Gamma(x)$, we have

$$P(z_{m,n}=k \mid \mathbf{W},\hat{\mathbf{W}},\mathbf{Z}_{-(\mathbf{m},\mathbf{n})},\hat{\mathbf{Z}};\alpha,\beta) \propto \prod_{i\neq k}\Gamma(n^{i,-(m,n)}_{m,(.)}+\hat{n}^i_{m,(.)}+\alpha_i)\prod_{i\neq k}\left[\frac{\Gamma(n^{i,-(m,n)}_{(.),\nu}+\beta_\nu)}{\Gamma(\beta_r+\sum_{r=1}^{V}n^{k,-(m,n)}_{(.),r})}\right]$$
$$\times\Gamma(\alpha_k+n^{k,-(m,n)}_{m,(.)}+\hat{n}^k_{m,(.)})\times\frac{\Gamma(n^{k,-(m,n)}_{(.),\nu}+\beta_\nu)}{\Gamma(\sum_{r=1}^{V}(n^{k,-(m,n)}_{(.),r}+\beta_r))}\times$$
$$\left(\alpha_k+n^{k,-(m,n)}_{m,(.)}+\hat{n}^k_{m,(.)}\right)\times\left(\frac{\beta_\nu+n^{k,-(m,n)}_{(.),\nu}}{\sum_{r=1}^{V}(\beta_r+n^{k,-(m,n)}_{(.),r})}\right)$$

Combining Gamma functions again, we get

$$P(z_{m,n}=k \mid \mathbf{W},\hat{\mathbf{W}},\mathbf{Z}_{-(\mathbf{m},\mathbf{n})},\hat{\mathbf{Z}};\alpha,\beta) \propto \prod_{i=1}^{K}\Gamma(n^{i,-(m,n)}_{m,(.)}+\hat{n}^i_{m,(.)}+\alpha_i)\times\prod_{i=1}^{K}\frac{\Gamma(n^{i,-(m,n)}_{(.),\nu}+\beta_\nu)}{\Gamma(\sum_{r=1}^{V}\beta_r+n^{i,-(m,n)}_{(.),r})}$$
$$\times\left(\alpha_k+n^{k,-(m,n)}_{m,(.)}+\hat{n}^k_{m,(.)}\right)\times\left(\frac{\beta_\nu+n^{k,-(m,n)}_{(.),\nu}}{\sum_{r=1}^{V}(\beta_r+n^{k,-(m,n)}_{(.),r})}\right)$$

Finally,

$$\boxed{P(z_{m,n}=k \mid \mathbf{W},\hat{\mathbf{W}},\mathbf{Z}_{-(\mathbf{m},\mathbf{n})},\hat{\mathbf{Z}};\alpha,\beta) \propto \left(\alpha_k+n^{k,-(m,n)}_{m,(.)}+\hat{n}^k_{m,(.)}\right)\times\left(\frac{\beta_\nu+n^{k,-(m,n)}_{(.),\nu}}{\sum_{r=1}^{V}(\beta_r+n^{k,-(m,n)}_{(.),r})}\right)} \quad (2)$$

Similar calculation for headline word yields,

$$P(\hat{z}_{m,n} = k \mid \mathbf{W}, \hat{\mathbf{W}}, \mathbf{Z}, \hat{\mathbf{Z}}_{-(\mathbf{m},\mathbf{n})}; \alpha, \beta) \propto \left(\alpha_k + n^k_{m,(.)} + \hat{n}^{k,-(m,n)}_{m,(.)}\right) \times \left(\frac{\hat{\beta}_\nu + \hat{n}^{k,-(m,n)}_{(.),\nu}}{\sum_{r=1}^{V}(\hat{\beta}_r + \hat{n}^{k,-(m,n)}_{(.),r})}\right) \quad (3)$$

8.3 Data Processing

8.3.1 Preprocessing

Only body and headline were used for the purpose of training and testing the model. Body and headline were both preprocessed using the same rules. Following set of rules were used for preprocessing articles:

1. Remove stop words [website link of stop words and appendix]
2. Remove words with length less than 4 and greater than 21
3. Replace numbers with 'num'
4. Identify tag of words and lemmatize word to its base form using WordNet Lemmatizer
5. Remove 'LEAD :' from the body of article which is present in majority of articles denoting lead sentence.

8.3.2 Noise

Preprocessing, no matter how carefully it is done, can never be perfect. In our case, replacing numbers with 'num' does not help much as numbers can be found in articles belonging to different topics. For example, 'num' is present in articles about sports, finance, war etc. Lemmatizing seems to be a good idea as it brings down the size of unique words that helps in strengthning statistical relationship between words. The WordNet Lemmatizer, inbuilt in most of the natural language processing toolkits, uses dictionary of words and is therefore an ideal lemmatizing tool for converting words to their base form. It does require part of speech tag of the words to do this. While *part of speech* tagging in itself does not have an accurate solution we can not ensure that conversion of words to their base form will be perfect. This is the reason that several words like killed and killing, sales and sale, share and shares, etc. are present in final topic distribution.

8.3.3 Data Storage

With such huge datasets there are several issues that need to be addressed while handling the data in order to make the process computationally efficient. Working with huge dataset of 2 million articles requires skills of data engineering too. Across the whole corpus there were around 1.8 million unique words after preprocessing. For easy lookup of articles from a pool of 1.8 million articles a logical step was to have all those articles on a MongoDB instance. Once that was done article body and headline were preprocessed and taken down in a separate text file. Each line in the text file represents *headline words | body words*. Since LDA deals with word numbers instead of word, a hash table mapping each word to its unique number was created and a text file with all words replaced by their unique number was generated. It is this file that was used further for our model.

8.4 Stopwords

'd, 'll, 'm, 're, 's, 't, n't, 've, a, aboard, about, above, across, after, again, against, all, almost, alone, along, alongside, already, also, although, always, am, amid, amidst, among, amongst, an, and, another, anti, any, anybody, anyone, anything, anywhere, are, area, areas, aren't, around, as, ask, asked, asking, asks, astride, at, aught, away, back, backed, backing, backs, bar, barring, be, became, because, become, becomes, been, before, began, behind, being, beings, below, beneath, beside, besides, best, better, between, beyond, big, both, but, by, came, can, can't, cannot, case, cases, certain, certainly, circa, clear, clearly, come, concerning, considering, could, couldn't, daren't, despite, did, didn't, differ, different, differently, do, does, doesn't, doing, don't, done, down, down, downed, downing, downs, during, each, early, either, end, ended, ending, ends, enough, even, evenly, ever, every, everybody, everyone, everything, everywhere, except, excepting, excluding, face, faces, fact, facts, far, felt, few, fewer, find, finds, first, five, following, for, four, from, full, fully, further, furthered, furthering, furthers, gave, general, generally, get, gets, give, given, gives, go, goes, going, good, goods, got, great, greater, greatest, group, grouped, grouping, groups, had, hadn't, has, hasn't, have, haven't, having, he, he'd, he'll, he's, her, here, here's, hers, herself, high, high, high, higher, highest, him, himself, his, hisself, how, how's, however, i, i'd, i'll, i'm, i've, idem, if, ilk, important, in, including, inside, interest, interested, interesting, interests, into, is, isn't, it, it's, its, itself, just, keep, keeps, kind, knew, know, known, knows, large, largely, last, later, latest, least, less, let, let's, lets, like, likely, long, longer, longest, made, make, making, man, many, may, me, member, members, men, might, mightn't, mine, minus, more, most, mostly, mr, mrs, much, must, mustn't, my, myself, naught, near, necessary, need, needed, needing, needn't, needs, neither, never, new, new, newer, newest, next, no, nobody, non, none, noone, nor, not, nothing, notwithstanding, now, nowhere, number, numbers, of, off, often, old, older, oldest, on, once, one, oneself, only, onto, open, opened, opening, opens, opposite, or, order, ordered, ordering, orders, other, others, otherwise, ought, oughtn't, our, ours, ourself, ourselves, out, outside, over, own, part, parted, parting, parts, past, pending, per, perhaps, place, places, plus, point, pointed, pointing, points, possible, present, presented, presenting, presents, problem, problems, put, puts, quite, rather, really, regarding, right, right, room, rooms, round, said, same, save, saw, say, says, second, seconds, see, seem, seemed, seeming, seems, seen, sees, self, several, shall, shan't, she, she'd, she'll, she's, should, shouldn't, show, showed, showing, shows, side, sides, since, small, smaller, smallest, so, some, somebody, someone, something, somewhat, somewhere, state, states, still, still, such, suchlike, sundry, sure, take, taken, than, that, that's, the, thee, their, theirs, them, themselves, then, there, there's, therefore, these, they, they'd, they'll, they're, they've, thine, thing, things, think, thinks, this, those, thou, though, thought, thoughts, three, through, throughout, thus, thyself, till, to, today, together, too, took, tother, toward, towards, turn, turned, turning, turns, twain, two, under, underneath, unless, unlike, until, up, upon, us, use, used, uses, various, versus, very, via, vis-a-vis, want, wanted, wanting, wants, was, wasn't, way, ways, we, we'd, we'll, we're, we've, well, wells, went, were, weren't, what, what's, whatall, whatever, whatsoever, when, when's, where, where's, whereas, wherewith, wherewithal, whether, which, whichever, whichsoever, while, who, who's, whoever, whole, whom, whomever, whomso, whomsoever, whose, whosoever, why, why's, will, with, within, without, won't, work, worked, working, works, worth, would, wouldn't, ye, year, years, yet, yon, yonder, you, you'd, you'll, you're, you've, you-all, young, younger, youngest, your, yours, yourself, yourselves

8.5 Results

Following are the results of running BHLDA model on 100,000 articles from the corpus. First section shows output of BHLDA model. Second section shows output of LDA on body only. Third section shows output of headline only.

8.5.1 BHLDA

8.5.1.1 Word Distribution

Results

```
                                    Topic 0 : "Entertainment: Art/Music/Travel/others"
----------------------------------------------------------------------------------------------------
                      BODY                                          HEADLINE
----------------------------------------------------------------------------------------------------
        Word                    Prob            ¡       Word                    Prob
----------------------------------------------------------------------------------------------------
        film                    0.0234137       ¡       review/film             0.0300011
        television              0.011184        ¡       mail                    0.0287416
        movie                   0.00985363      ¡       answering               0.0243902
        race                    0.00623159      ¡       review/television        0.021642
        num                     0.00584287      ¡       home                    0.0198099
        time                    0.00445975      ¡       film                    0.0153441
        video                   0.00441605      ¡       movies                  0.0127104
        network                 0.00434976      ¡       star                    0.0108783
        star                    0.00378174      ¡       television              0.00927516
        camera                  0.00351958      ¡       camera                  0.00870262
        horse                   0.00344877      ¡       racing                  0.00858811
        producer                0.0032966       ¡       horse                   0.0083591
        screen                  0.00325893      ¡       video                   0.00813008
```

wall	0.00299827	¡	world	0.00755754
series	0.00298321	¡	movie	0.00721402
hollywood	0.00291842	¡	films	0.00641246
produce	0.00288075	¡	hollywood	0.00641246
home	0.0028687	¡	improvement	0.00629795
tape	0.00277981	¡	coping	0.00618344
track	0.00275269	¡	consumer	0.00606893

Topic 1 : "Crime"

BODY			HEADLINE	
Word	Prob	¡	Word	Prob
police	0.0184455	¡	case	0.0248299
charge	0.0111801	¡	police	0.0205996
officer	0.00913811	¡	trial	0.01476
court	0.00801631	¡	death	0.0129207
lawyer	0.00772961	¡	drug	0.0112194
num	0.00732547	¡	fire	0.0107596
drug	0.00709889	¡	killing	0.00965606
yesterday	0.00660874	¡	bronx	0.00951812
judge	0.00640805	¡	brooklyn	0.0088284
trial	0.00612136	¡	killed	0.00864447
official	0.00544532	¡	crime	0.00864447
crime	0.00537873	¡	inquiry	0.00818466
federal	0.00533989	¡	held	0.00786279
arrest	0.00524926	¡	charges	0.00767887
kill	0.0051466	¡	guilty	0.00740298
investigation	0.00472766	¡	shot	0.00717307
city	0.00471009	¡	judge	0.00717307
jury	0.00456027	¡	says	0.00703513
report	0.00409139	¡	jury	0.00694317
department	0.0040359	¡	charged	0.00671326

Topic 2 : "Life"

BODY			HEADLINE	
Word	Prob	¡	Word	Prob
people	0.0107835	¡	life	0.0147475
time	0.00883833	¡	journal	0.0145362
life	0.00736629	¡	home	0.0117896
child	0.00726842	¡	children	0.011536
woman	0.00620914	¡	quotation	0.00781745
look	0.00585216	¡	just	0.0077752
tell	0.00568939	¡	family	0.00756391
family	0.00545968	¡	time	0.00739489
home	0.00452971	¡	still	0.00714135
little	0.00416309	¡	back	0.00693007
live	0.00412405	¡	little	0.00629622
friend	0.00387964	¡	notebook	0.00621171
leave	0.00370723	¡	times	0.00600042
call	0.00352671	¡	world	0.00566237
mother	0.00348412	¡	love	0.00519755
love	0.00319559	¡	again	0.00490175
talk	0.00318748	¡	good	0.00481724
feel	0.00300087	¡	child	0.00464821
father	0.00295169	¡	towns	0.00460596
night	0.00271539	¡	find	0.00447919

Topic 3 : "Real Estate"

BODY			HEADLINE	
Word	Prob	¡	Word	Prob
num	0.0289247	¡	region	0.0174451
building	0.0147058	¡	sales	0.0161245

city	0.0141093	¡	postings	0.0134835
space	0.00844753	¡	housing	0.013275
project	0.00762176	¡	york	0.0116069
street	0.00700158	¡	jersey	0.0115374
housing	0.00676435	¡	recent	0.0115374
york	0.0065339	¡	num	0.0111899
house	0.0058403	¡	space	0.00973033
build	0.00573863	¡	connecticut	0.00910481
apartment	0.00541668	¡	westchester	0.0087573
plan	0.00527095	¡	g.m.	0.0087573
cost	0.00497272	¡	building	0.00861829
property	0.00491398	¡	ford	0.00854879
car	0.00489817	¡	home	0.00834028
development	0.0044384	¡	city	0.00785377
construction	0.00428025	¡	shuttle	0.00778426
office	0.00427686	¡	real	0.00743675
home	0.00406674	¡	plant	0.00736725
tax	0.0040306	¡	census	0.00729775

Topic 4 : "Markets/Finance"

BODY			HEADLINE	
Word	Prob	¡	Word	Prob
share	0.0974261	¡	report	0.213522
company	0.0653879	¡	earnings	0.211454
earn	0.0612959	¡	march	0.0505409
num	0.0553811	¡	inc.	0.0462507
reports	0.0489477	¡	corp	0.0451396
loss	0.035445	¡	sept	0.0331332
shares	0.029023	¡	june	0.0289665
outst	0.028459	¡	year	0.0126854
revenue	0.0282319	¡	industries	0.00902792
corp	0.0227865	¡	bancorp	0.00800938
sale	0.01842	¡	international	0.00628096
sales	0.0179279	¡	group	0.00617293
inc.	0.0157115	¡	financial	0.00611121
quarter	0.0148276	¡	bank	0.00558651
march	0.014767	¡	systems	0.00501551
nyse	0.0132207	¡	first	0.00490748
income	0.0125147	¡	american	0.00479946
cent	0.0115929	¡	april	0.00395068
june	0.00854186	¡	national	0.00391981
operation	0.00782074	¡	savings	0.00388895

Topic 5 : "Sports"

BODY			HEADLINE	
Word	Prob	¡	Word	Prob
num	0.0143816	¡	plus	0.0283488
game	0.0141354	¡	results	0.0277551
play	0.0098829	¡	bridge	0.027013
player	0.00935543	¡	mets	0.0192208
team	0.00862918	¡	baseball	0.0178108
season	0.00726929	¡	wins	0.0127644
baseball	0.00720153	¡	num	0.0106865
league	0.00705582	¡	chess	0.00972171
time	0.00676103	¡	yanks	0.00920223
club	0.00648431	¡	question	0.00890538
inning	0.00579646	¡	yankees	0.00846011
run	0.005522	¡	sports	0.00801484
pitch	0.00539663	¡	game	0.00794063
victory	0.00528368	¡	pirates	0.00749536
lead	0.00517186	¡	steinbrenner	0.00697588
mets	0.00511878	¡	week	0.00667904
third	0.00494145	¡	pastimes	0.00615955
home	0.0048285	¡	victory	0.00564007
start	0.00470991	¡	johnson	0.00556586

yankees	0.00463762	¡	title	0.00549165

Topic 6 : "Entertainment: Dance/Music/Theater"

BODY			HEADLINE	
Word	Prob	¡	Word	Prob
music	0.0153532	¡	review/music	0.0292917
play	0.010969	¡	chronicle	0.0204783
theater	0.0104732	¡	review/dance	0.0191174
dance	0.00660826	¡	review/theater	0.0178213
performance	0.00626698	¡	music	0.0156827
p.m.	0.00577906	¡	reviews/music	0.0111464
num	0.00558616	¡	festival	0.00952628
song	0.0051995	¡	stage	0.00952628
opera	0.00455447	¡	theater	0.00900784
concert	0.00452916	¡	town	0.00823019
musical	0.00440172	¡	opera	0.00803577
production	0.0039173	¡	ballet	0.00764694
festival	0.00391468	¡	sounds	0.00764694
director	0.00390159	¡	guide	0.00758214
stage	0.00384311	¡	jazz	0.0069989
york	0.00380034	¡	dance	0.005638
program	0.00376979	¡	rock	0.00537878
perform	0.00376455	¡	works	0.00511956
ballet	0.00367116	¡	review/pop	0.00460113
sing	0.00349746	¡	concert	0.00460113

Topic 7 : "International Affairs"

BODY			HEADLINE	
Word	Prob	¡	Word	Prob
soviet	0.0149149	¡	u.s.	0.0417916
united	0.0131222	¡	europe	0.0233992
states	0.009882	¡	east	0.0231267
american	0.00872309	¡	gulf	0.0220368
official	0.00819741	¡	soviet	0.0219346
president	0.00794829	¡	iraq	0.0137943
iraq	0.0073414	¡	u.n.	0.0123638
country	0.00726458	¡	bush	0.0118869
union	0.00725909	¡	gorbachev	0.0112738
military	0.00659897	¡	talks	0.0112398
east	0.00644752	¡	evolution	0.0108651
germany	0.00643874	¡	upheaval	0.00946866
force	0.00639869	¡	moscow	0.00858311
bush	0.00605464	¡	german	0.00810627
europe	0.00563651	¡	confrontation	0.00797003
gorbachev	0.00529575	¡	soviets	0.00776567
german	0.00527051	¡	world	0.00773161
world	0.00524088	¡	trade	0.00766349
government	0.00522167	¡	japan	0.00759537
foreign	0.00522003	¡	germany	0.00749319

Topic 8 : "Education"

BODY			HEADLINE	
Word	Prob	¡	Word	Prob
school	0.0212504	¡	york	0.0248148
city	0.0166143	¡	life	0.0235164
student	0.0142545	¡	campus	0.0217851
program	0.0112034	¡	school	0.0172165
child	0.00869296	¡	city	0.0170722
university	0.00869219	¡	dinkins	0.011638
people	0.00851684	¡	schools	0.0113494
black	0.00803716	¡	black	0.0112052

york	0.007739	¡	students	0.0107723
education	0.00689087	¡	more	0.00880062
college	0.00612462	¡	education	0.00812734
board	0.00550204	¡	college	0.00759835
public	0.0050857	¡	social	0.00706935
teacher	0.00479063	¡	help	0.00673271
community	0.00465622	¡	poor	0.00658844
help	0.00463537	¡	state	0.00649226
mayor	0.00458902	¡	board	0.00644417
percent	0.0044971	¡	plan	0.00610753
service	0.0040545	¡	homeless	0.0057228
director	0.0037378	¡	women	0.0052419

Topic 9 : "Markets/Finance"

BODY			HEADLINE	
Word	Prob	¡	Word	Prob
percent	0.0278079	¡	prices	0.0214248
price	0.0160027	¡	market	0.0176193
market	0.0153254	¡	rates	0.017391
rate	0.0125998	¡	rise	0.015983
bank	0.0124625	¡	bank	0.0155644
rise	0.00964166	¡	u.s.	0.01252
stock	0.00869362	¡	dollar	0.0116067
bond	0.00816455	¡	digest	0.0107695
increase	0.00731753	¡	place	0.0100464
loan	0.00650095	¡	price	0.0093995
fund	0.00611565	¡	economic	0.00875257
yesterday	0.00570897	¡	data	0.00799148
dollar	0.00553801	¡	stocks	0.00772509
week	0.0055283	¡	fall	0.00753482
report	0.00547002	¡	drop	0.00749677
economy	0.00542275	¡	money	0.00707816
company	0.00523171	¡	trading	0.00700206
month	0.00480043	¡	bond	0.00700206
money	0.00473114	¡	debt	0.006964
decline	0.00468387	¡	mixed	0.00681178

Topic 10 : "International Affairs"

BODY			HEADLINE	
Word	Prob	¡	Word	Prob
government	0.014751	¡	news	0.0183981
party	0.0122893	¡	summary	0.0170171
leader	0.00697705	¡	south	0.0130969
people	0.00692922	¡	party	0.0109141
president	0.00667387	¡	leader	0.00944405
country	0.0065975	¡	israel	0.00944405
political	0.00636683	¡	mandela	0.00913222
south	0.00569412	¡	u.s.	0.00890948
official	0.00528679	¡	india	0.0086422
national	0.00518342	¡	africa	0.00841946
minister	0.00469354	¡	east	0.00810763
election	0.00414812	¡	upheaval	0.00792944
united	0.00409412	¡	china	0.0059248
army	0.00405169	¡	panama	0.0059248
military	0.00393983	¡	mexico	0.00574661
force	0.00381408	¡	rebels	0.00561297
african	0.00371071	¡	says	0.00543478
communist	0.0034515	¡	israeli	0.00521205
israel	0.00341524	¡	army	0.00512295
africa	0.00340135	¡	journal	0.00490021

Topic 11 : "Exhibitions"

BODY			HEADLINE	

Word	Prob	¡	Word	Prob
museum	0.011049	¡	fashion	0.0186517
num	0.00793491	¡	style	0.0184008
street	0.00689787	¡	pastimes	0.0122951
artist	0.00569978	¡	review/art	0.0109568
painting	0.00566148	¡	design	0.0106223
p.m.	0.00557408	¡	makers	0.0105386
design	0.00539535	¡	museum	0.00978588
plant	0.00448696	¡	garden	0.00911676
exhibition	0.00428858	¡	street	0.00911676
collection	0.00419627	¡	designer	0.00861492
house	0.00417565	¡	show	0.008364
look	0.00415601	¡	guide	0.00828036
garden	0.00408137	¡	currents	0.00777852
gallery	0.00378676	¡	spring	0.006273
a.m.	0.00361883	¡	paris	0.0058548
avenue	0.00350982	¡	artist	0.00577116
color	0.00324664	¡	events	0.00518568
white	0.00308558	¡	works	0.00510204
designer	0.00298247	¡	stamps	0.00485112
black	0.0028823	¡	auctions	0.0046002

Topic 12 : "Creative Writing"

BODY			HEADLINE	
Word	Prob	¡	Word	Prob
book	0.0130651	¡	corrections	0.084565
write	0.00939551	¡	times	0.0332116
editor	0.00596609	¡	books	0.0304295
life	0.00577702	¡	correction	0.0143164
world	0.00529524	¡	best	0.0132151
american	0.00472367	¡	book	0.0104909
writer	0.00451782	¡	history	0.00724512
story	0.00435211	¡	sellers	0.00689735
novel	0.00414042	¡	america	0.00678143
author	0.00400829	¡	american	0.00660755
history	0.00389296	¡	words	0.0063757
article	0.00384551	¡	notebook	0.00620182
time	0.00328854	¡	short	0.0060859
publish	0.00321116	¡	fiction	0.00602794
woman	0.00310604	¡	topics	0.00573813
word	0.00288924	¡	editorial	0.00550629
york	0.00277755	¡	notes	0.00544833
page	0.00267025	¡	nonfiction	0.00539037
reader	0.00240235	¡	language	0.00515852
america	0.0023184	¡	does	0.00510056

Topic 13 : "??"

BODY			HEADLINE	
Word	Prob	¡	Word	Prob
mrs.	0.0220598	¡	weds	0.0273395
york	0.0186195	¡	chronicle	0.022801
university	0.0179235	¡	dies	0.0209999
graduate	0.0113756	¡	paid	0.0175059
die	0.010827	¡	notice	0.0164253
daughter	0.0107663	¡	miss	0.015921
father	0.00967006	¡	deaths	0.0154888
president	0.00887744	¡	married	0.0143722
church	0.00787456	¡	executive	0.0121749
yesterday	0.0077835	¡	john	0.0108422
college	0.00732816	¡	marry	0.0100137
marry	0.00704821	¡	dead	0.00943736
school	0.00699761	¡	lawyer	0.00904114
n.j.	0.00628706	¡	marries	0.00796052

manhattan	0.0056181	¡	bride	0.00709603
n.y.	0.00539324	¡	robert	0.00706001
john	0.00519987	¡	professor	0.00698797
retire	0.00501211	¡	david	0.00691593
director	0.00484009	¡	engaged	0.0059794
home	0.00479624	¡	becomes	0.00572725

Topic 14 : "National Budget"

BODY			HEADLINE	
Word	Prob	¡	Word	Prob
court	0.00811334	¡	court	0.0152333
house	0.00749361	¡	bush	0.0145586
president	0.00723223	¡	budget	0.0142746
bush	0.00672285	¡	house	0.0113983
federal	0.00660749	¡	bill	0.0109367
vote	0.00602175	¡	plan	0.010049
budget	0.00593871	¡	washington	0.0088417
committee	0.00578322	¡	panel	0.00859314
bill	0.00576929	¡	congress	0.00859314
republican	0.00568737	¡	rights	0.00756338
campaign	0.00551739	¡	senate	0.00710177
issue	0.00546165	¡	cuomo	0.00703075
congress	0.00543435	¡	vote	0.00685321
senate	0.00526827	¡	abortion	0.00653363
senator	0.00489153	¡	state	0.00649812
political	0.004867	¡	judge	0.00649812
government	0.00462792	¡	rules	0.00646261
party	0.00434703	¡	u.s.	0.00632057
governor	0.00432084	¡	says	0.00571692
public	0.00431582	¡	race	0.00536184

Topic 15 : "Sports"

BODY			HEADLINE	
Word	Prob	¡	Word	Prob
game	0.0225624	¡	week	0.0197484
team	0.0148472	¡	question	0.0189308
num	0.0142277	¡	football	0.0126415
play	0.0141811	¡	knicks	0.0114465
season	0.0112314	¡	coach	0.0113208
coach	0.0100116	¡	giants	0.0110692
score	0.00889259	¡	college	0.0101887
player	0.00840575	¡	jets	0.0101887
goal	0.00625791	¡	num	0.01
football	0.00608926	¡	rangers	0.00937107
league	0.00574561	¡	sports	0.00893082
leave	0.00561833	¡	nets	0.0081761
time	0.00545393	¡	devils	0.00761006
pass	0.00534468	¡	basketball	0.00691824
basketball	0.00503496	¡	game	0.00628931
shot	0.00486738	¡	islanders	0.00597484
victory	0.00483132	¡	team	0.00559748
lead	0.00467646	¡	victory	0.00515723
national	0.00456085	¡	people	0.00503145
yard	0.00427765	¡	next	0.0045283

Topic 16 : "Food"

BODY			HEADLINE	
Word	Prob	¡	Word	Prob
food	0.0115535	¡	food	0.0426385
wine	0.00960924	¡	sunday	0.0197593
num	0.00924595	¡	lifestyle	0.0166394

minute	0.00886395	¡	wine	0.0141138
restaurant	0.00853808	¡	num	0.0127767
serve	0.0079565	¡	menu	0.0120339
cook	0.00634431	¡	restaurants	0.0114396
time	0.00610264	¡	talk	0.0102511
pepper	0.00557563	¡	eating	0.00950825
sauce	0.00547117	¡	journal	0.00935968
tablespoon	0.00530122	¡	notes	0.00906255
salt	0.00524509	¡	dinner	0.00906255
chicken	0.00518272	¡	diner	0.00861685
fresh	0.00515933	¡	gourmet	0.00802258
dish	0.00507358	¡	taste	0.00757688
water	0.00476954	¡	summer	0.00683405
taste	0.00444835	¡	table	0.00594265
cooking	0.00415678	¡	fish	0.00579409
fish	0.0041178	¡	orange	0.00564552
heat	0.00408038	¡	cooking	0.00534839

Topic 17 : "Health"

BODY			HEADLINE	
Word	Prob	¡	Word	Prob
health	0.00894664	¡	aids	0.024764
drug	0.0084477	¡	health	0.018573
study	0.00796545	¡	drug	0.0166728
patient	0.00665225	¡	study	0.0140983
aids	0.00657156	¡	patents	0.0109722
medical	0.00588622	¡	care	0.0106657
disease	0.0054457	¡	u.s.	0.00937845
test	0.00541417	¡	test	0.00766213
hospital	0.00540953	¡	found	0.00698786
people	0.00516562	¡	hospital	0.00674267
report	0.00504599	¡	nuclear	0.00668138
research	0.00482434	¡	drugs	0.0062523
find	0.00444689	¡	cancer	0.00606841
percent	0.00442649	¡	science	0.00582322
doctor	0.00434487	¡	medical	0.00576192
cause	0.00420947	¡	research	0.00545544
plant	0.00402121	¡	says	0.00533284
environmental	0.00387283	¡	tests	0.00508765
scientist	0.00380698	¡	environment	0.00502636
treatment	0.00363912	¡	risk	0.00459728

Topic 18 : "BUsiness"

BODY			HEADLINE	
Word	Prob	¡	Word	Prob
company	0.0407828	¡	business	0.0425145
business	0.0125272	¡	media	0.0190595
executive	0.0103585	¡	briefs	0.0184912
corporation	0.00849508	¡	advertising	0.0170892
president	0.00835844	¡	executive	0.0152325
computer	0.00825666	¡	deal	0.0147778
inc.	0.00782654	¡	unit	0.0133758
sell	0.00629358	¡	company	0.0126937
industry	0.00567942	¡	changes	0.0109507
sale	0.00550723	¡	chief	0.010155
advertising	0.00531692	¡	computer	0.00943503
news	0.00510987	¡	sale	0.00916979
american	0.00507014	¡	news	0.0091319
product	0.00467348	¡	people	0.00780569
chief	0.0046442	¡	plans	0.00735099
percent	0.00460377	¡	stake	0.00712364
vice	0.00457658	¡	accounts	0.00700997
service	0.00428518	¡	president	0.00617635
chairman	0.00426009	¡	sell	0.00613846
market	0.0039917	¡	plan	0.005949

Topic 19 : "Travel"

BODY			HEADLINE	
Word	Prob	¡	Word	Prob
island	0.00808076	¡	island	0.018196
water	0.0068362	¡	long	0.0115147
park	0.0060742	¡	traffic	0.00938233
mile	0.00566	¡	west	0.00860047
num	0.00553713	¡	journal	0.00860047
hotel	0.00494854	¡	travel	0.00767645
city	0.00487918	¡	outdoors	0.00696567
river	0.00432824	¡	park	0.00675243
town	0.00410033	¡	alert	0.00639704
travel	0.00396755	¡	water	0.00618381
flight	0.00390611	¡	california	0.00611273
people	0.00383279	¡	spill	0.0055441
north	0.00381198	¡	beach	0.00547303
airport	0.00359596	¡	canada	0.00547303
beach	0.00357714	¡	land	0.0049044
land	0.00349687	¡	airlines	0.00483332
service	0.00344634	¡	town	0.00469116
road	0.00317979	¡	airport	0.00454901
ship	0.00313817	¡	river	0.00447793
airline	0.00310943	¡	hotel	0.00440685

8.5.1.2 Topic Distributions

Each table below represents proportion of topic allocations for each topic for headline and body in *BHLDA* model. Only 4 out of 100000 thousand documents are displayed here.

Document 0

Headline	Body
0	0
0	0.025641
0	0.0128205
0	0
0	0
0	0.0128205
1	0.025641
0	0
0	0.0512821
0	0
0	0.128205
0	0.0512821
0	0.435897
0	0.0512821
0	0
0	0.0384615
0	0
0	0
0	0.102564
0	0.0641026

Document 1

Headline	Body
0	0
0	0
0	0
0	0.0215054
0	0
1	0.569892
0	0.00537634
0	0
0	0.0806452
0	0
0	0
0	0.0268817
0	0
0	0.107527
0	0
0	0.016129
0	0.00537634
0	0.0215054
0	0
0	0.145161

Document 2

Headline	Body
0	0
0	0
0	0
0	0
0	0
0	0
0	0
0	0
0	0
0	0
0	0
0	0
0	0
0	0
0	0
1	1
0	0
0	0
0	0
0	0

Document 3

Headline	Body
0	0
0	0
0	0

0	0.666667
0	0
0	0
0	0
0	0
0	0
0	0
0	0
1	0.333333
0	0
0	0
0	0
0	0
0	0
0	0
0	0
0	0

8.5.2 Body LDA

8.5.2.1 Word Distribution

Topic 0

Body	
Word	Prob
budget	0.0116308
house	0.0114043
bush	0.0107077
bill	0.00812586
republican	0.00810293
president	0.00810197
senate	0.00787169
committee	0.00733946
congress	0.00691998
senator	0.00676518
administration	0.00614313
vote	0.006103
tax	0.00604375
governor	0.00575327
democrats	0.00570932
campaign	0.00567587
plan	0.00485985
political	0.00476907
increase	0.00476907
percent	0.00459134

Topic 1

Body	
Word	Prob

government	0.0130138
united	0.012869
states	0.010716
official	0.00868836
american	0.00856712
military	0.00841008
country	0.0078039
president	0.0069764
force	0.00676973
army	0.00463222
people	0.00460374
china	0.00453702
foreign	0.00428234
political	0.00409357
japan	0.00403417
leader	0.00402603
minister	0.00373962
party	0.00372742
general	0.00326688
world	0.003262

Topic 2

Body

Word	Prob
food	0.0083529
wine	0.00775865
minute	0.00734067
num	0.00683055
serve	0.00625098
restaurant	0.00592648
time	0.00567543
cook	0.00507984
water	0.00479006
sauce	0.0046859
pepper	0.00465252
tablespoon	0.00454034
salt	0.00423988
dish	0.00422385
chicken	0.00421851
fresh	0.00417845
remove	0.00381389
fish	0.0036069
taste	0.00356818
heat	0.00346669

Topic 3

Body

Word	Prob
percent	0.0244879
company	0.0155721

price	0.0147078
market	0.0145751
stock	0.0109796
bank	0.0100649
rate	0.0100156
rise	0.0084092
bond	0.00724518
yesterday	0.00609052
sell	0.00580948
sale	0.00566554
increase	0.00514398
exchange	0.00502932
week	0.00502371
corporation	0.00487167
report	0.00479066
trading	0.00462179
fund	0.00458876
dollar	0.00458253

Topic 4

Body

Word	Prob
university	0.0204792
mrs.	0.0184984
york	0.0170356
president	0.0128655
graduate	0.0111963
college	0.00977416
school	0.00968087
daughter	0.00915589
die	0.00883932
father	0.00850786
yesterday	0.00719888
vice	0.0062293
marry	0.00593654
student	0.00579165
director	0.00564081
company	0.00552073
n.j.	0.00540164
name	0.00513766
manhattan	0.00458688
john	0.00451245

Topic 5

Body

Word	Prob
people	0.0119037
city	0.00824251
num	0.00818789
police	0.00648786

street	0.00593402
time	0.00539278
home	0.00525484
night	0.00460788
child	0.00453576
family	0.0043089
leave	0.00404563
live	0.00380547
life	0.00379217
fire	0.00369695
town	0.00361152
tell	0.00358352
hour	0.003503
call	0.00330975
woman	0.00320192
look	0.00318581

Topic 6

Body

Word	Prob
plant	0.0124587
water	0.00808522
environmental	0.00631755
company	0.00391835
energy	0.00369026
chemical	0.00338363
tree	0.00314371
department	0.00296081
nuclear	0.00295759
power	0.00288012
mile	0.00285107
time	0.0028188
grow	0.00269507
official	0.00264451
island	0.00262084
cause	0.00259286
produce	0.00258533
waste	0.0025423
fuel	0.00251002
industry	0.00245838

Topic 7

Body

Word	Prob
museum	0.00550921
look	0.00514954
artist	0.00471549
painting	0.00410967
time	0.00354908
world	0.0033681

design	0.00335506
woman	0.00311427
black	0.00295782
collection	0.00291718
style	0.00279141
num	0.00274233
white	0.00270476
color	0.00270246
wear	0.00269632
exhibition	0.00258819
image	0.00252838
century	0.00240108
gallery	0.00236273
life	0.00234816

Topic 8

Body

Word	Prob
num	0.037614
street	0.0165849
p.m.	0.014643
building	0.0131044
avenue	0.00941306
house	0.00912038
city	0.00822917
park	0.00753044
a.m.	0.0070532
york	0.00704529
center	0.00639139
hotel	0.00584163
west	0.00583504
museum	0.00561883
east	0.00496097
manhattan	0.00481727
sunday	0.00412513
square	0.00409613
include	0.00388915
build	0.00385355

Topic 9

Body

Word	Prob
child	0.0121692
school	0.00888737
people	0.00773239
student	0.00746216
study	0.00742635
woman	0.00626486
drug	0.00587173
health	0.00568941

university	0.00566418
aids	0.00552907
patient	0.00540535
program	0.00495687
medical	0.0044156
disease	0.00436595
percent	0.00400049
test	0.00378236
time	0.00371155
research	0.00370992
doctor	0.00362771
parent	0.00357806

Topic 10

Body

Word	Prob
police	0.0123347
charge	0.012171
court	0.0108778
lawyer	0.00930189
drug	0.00871472
judge	0.00861565
federal	0.00743609
officer	0.00697303
trial	0.00691045
yesterday	0.00551501
investigation	0.0055025
official	0.00534606
arrest	0.00523134
crime	0.00513539
jury	0.00499876
former	0.00466503
num	0.00447938
prison	0.0044554
prosecutor	0.00441577
attorney	0.00436571

Topic 11

Body

Word	Prob
city	0.0120619
york	0.0068392
board	0.00617863
program	0.00573534
school	0.00561109
percent	0.00545188
people	0.00543627
plan	0.00532139
federal	0.00518902
company	0.00499859

money	0.00495676
cost	0.004845
official	0.0047039
service	0.00429557
business	0.00427809
public	0.0040989
government	0.00408204
housing	0.00384978
agency	0.00380982
department	0.00366684

Topic 12

Body

Word	Prob
black	0.0115722
court	0.0101663
white	0.0058748
issue	0.00578114
political	0.00565459
people	0.00526998
abortion	0.00486145
national	0.00484452
right	0.00477975
judge	0.00475883
campaign	0.00475384
president	0.00474388
party	0.00447485
public	0.00441606
vote	0.0043513
decision	0.00425066
woman	0.00417095
candidate	0.00392085
support	0.00343859
supreme	0.00336087

Topic 13

Body

Word	Prob
soviet	0.0136997
president	0.00701941
government	0.00692169
iraq	0.00682186
united	0.00659895
party	0.00650863
union	0.00639876
country	0.00631741
east	0.00617532
germany	0.00585998
official	0.00575592
gorbachev	0.00506925

europe	0.00485163
german	0.00484581
west	0.00478613
leader	0.00456111
states	0.00455636
kuwait	0.00447343
force	0.00428802
american	0.00427587

Topic 14

Body

Word	Prob
music	0.0173674
play	0.00914244
theater	0.00904943
dance	0.00739838
performance	0.00660525
song	0.00591244
opera	0.00518409
concert	0.00517155
musical	0.00495942
program	0.00470863
num	0.00467937
ballet	0.00430318
perform	0.00426661
sound	0.00422168
orchestra	0.00414853
sing	0.00406075
band	0.00401059
festival	0.00381727
york	0.00356126
company	0.00347766

Topic 15

Body

Word	Prob
company	0.0225368
business	0.00868179
computer	0.00815759
executive	0.00666883
advertising	0.00559852
system	0.00550108
news	0.00504637
industry	0.00500105
corporation	0.00491041
president	0.0047178
television	0.00470647
american	0.00458638
network	0.00455918
sell	0.00441114

agency	0.00425856
program	0.00415433
product	0.00403649
service	0.00387863
time	0.00380385
num	0.00374494

Topic 16

Body

Word	Prob
num	0.0128526
game	0.0124881
play	0.00888818
player	0.00838206
team	0.00788103
race	0.00699608
time	0.00685148
season	0.00650218
baseball	0.00613761
league	0.00611623
inning	0.00522619
club	0.00506834
run	0.00485551
pitch	0.00467628
world	0.00465693
start	0.00462332
mets	0.00460295
victory	0.00457444
lead	0.00436771
home	0.00436364

Topic 17

Body

Word	Prob
game	0.0214491
num	0.0138355
team	0.013444
play	0.0134293
season	0.010747
coach	0.00960476
score	0.0084364
player	0.00782664
goal	0.00583762
football	0.00575931
time	0.00565699
league	0.0054419
leave	0.00543772
pass	0.00504514
basketball	0.0049804
victory	0.00459199

shot	0.00456902
lead	0.00452621
national	0.00434036
giants	0.00411797

Topic 18

Body	
Word	Prob
share	0.0914699
company	0.0671279
earn	0.0578368
num	0.0486523
reports	0.0479255
loss	0.0329922
shares	0.0285831
outst	0.0280917
revenue	0.0268624
corp	0.0233911
inc.	0.0189202
sale	0.0178927
sales	0.0170613
quarter	0.0136685
march	0.0133789
nyse	0.01305
income	0.011746
cent	0.0109407
operation	0.00793092
june	0.00777398

Topic 19

Body	
Word	Prob
book	0.00885158
film	0.00870925
write	0.00674822
life	0.00610715
story	0.00498235
time	0.00498013
play	0.00454311
movie	0.00402436
woman	0.00351561
character	0.00339051
world	0.00328543
people	0.003172
novel	0.00311918
love	0.00301966
writer	0.00296628
television	0.00277724
tell	0.00265548
york	0.00261044

author	0.00260265
family	0.00236969

Topic 19

Body	
Word	Prob
book	0.00885158
film	0.00870925
write	0.00674822
life	0.00610715
story	0.00498235
time	0.00498013
play	0.00454311
movie	0.00402436
woman	0.00351561
character	0.00339051
world	0.00328543
people	0.003172
novel	0.00311918
love	0.00301966
writer	0.00296628
television	0.00277724
tell	0.00265548
york	0.00261044
author	0.00260265
family	0.00236969

8.5.2.2 Topic Distributions

Each line represents body topic distribution. Number at i^{th} position represents proportion of time topic 'i' was allocated to words in the body.

```
0,0,0,0,0,0,0,0.0128205,0.24359,0,0,0,0.410256,0,0,0.0897436,0,0,0,0.24359
0,0,0,0,0.134409,0.241935,0,0,0,0.0806452,0,0,0.016129,0,0,0,0.526882,0,0,0
0.0408163,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0.959184,0,0
0,0,0,0,0,0.666667,0,0,0,0.333333,0,0,0,0,0,0,0,0,0,0
0,0,0,0,0,0,0,0,0,0,0,0.00166667,0,0.983333,0,0.005,0.00666667,0,0.00333333,0

0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0
0,0.160237,0,0,0.0207715,0.513353,0,0,0,0,0.181009,0,0.121662,0,0,0,0.00296736,0,0,0

0,0,0,0.995238,0,0.0047619,0,0,0,0,0,0,0,0,0,0,0,0,0,0
0,0,0,0.0163934,0,0,0,0.052459,0,0,0,0,0,0.0262295,0,0,0.904918,0,0,0
0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0
0,0,0,0.417112,0,0,0.208556,0,0,0,0,0,0,0.374332,0,0,0,0,0,0
```

8.5.3 Headline LDA

Each line represents headline topic distribution. Number at i^{th} position represents proportion of time topic i was allocated to words in the headline.

8.5.3.1 Word Distribution

Topic 0

Headline	
Word	Prob
notes	0.0102721
times	0.00856823
review/dance	0.00832481
world	0.00817876
review/music	0.00735115
children	0.00603671
food	0.00554988
traffic	0.00545251
city	0.00545251
num	0.00481963
books	0.00481963
theater	0.00457621
york	0.00452753
works	0.00447885
music	0.00447885
travel	0.0043328
alert	0.0043328
book	0.00428411
life	0.00428411
star	0.00423543

Topic 1

Headline	
Word	Prob
report	0.20151
earnings	0.199088
march	0.0489313
corp	0.0475634
inc.	0.0465945
sept	0.0333143
june	0.0291821
year	0.0112283
bancorp	0.0111998
industries	0.00948988
financial	0.00846395
bank	0.00698205
international	0.00683956
first	0.00644058
group	0.00635509
national	0.00550014
american	0.00478769
savings	0.00475919
april	0.00407524
federal	0.00370476

Topic 2

Headline

Word	Prob
report	0.168001
earnings	0.161769
march	0.0367358
corp	0.0261086
inc.	0.0251246
june	0.0215823
sept	0.0196143
year	0.0118735
systems	0.00774075
american	0.00701916
group	0.00688796
l.p.	0.00518237
restaurants	0.00491997
diary	0.00452637
stores	0.00452637
metropolitan	0.00452637
ltd.	0.00426397
bancorp	0.00419837
partners	0.00413277
electronics	0.00406717

Topic 3

Headline

Word	Prob
news	0.047635
briefs	0.0325818
summary	0.025557
company	0.0187329
chief	0.0140496
makers	0.01037
style	0.00970094
head	0.00863049
president	0.00849669
finance	0.00762695
health	0.00756005
strike	0.00608818
dies	0.00528534
named	0.00508463
executive	0.00501773
personal	0.00454941
u.s.	0.0044156
post	0.0042818
designer	0.0042818
daily	0.00421489

Topic 4

Headline

Word	Prob

corrections	0.11959
bridge	0.0244262
correction	0.0203279
quotation	0.0172951
miss	0.0164754
executive	0.0105738
weds	0.00991803
dies	0.00868852
plans	0.00729508
transactions	0.00663934
profits	0.00639344
headline	0.00622951
marry	0.00581967
wedding	0.00581967
computer	0.00565574
noted	0.00508197
pleasure	0.00483607
john	0.00459016
noteworthy	0.00442623
mark	0.00393443

Topic 5

Headline

Word	Prob
u.s.	0.0115534
bush	0.0115098
budget	0.0101583
york	0.00797838
washington	0.00719362
race	0.00697563
plan	0.00684484
house	0.00614727
taxes	0.00540611
bill	0.00531892
dinkins	0.00505733
senate	0.00483934
democrats	0.00466495
cuomo	0.00457776
vote	0.00449056
congress	0.00440337
campaign	0.00414178
state	0.00396739
g.o.p.	0.00388019
city	0.003793

Topic 6

Headline

Word	Prob
times	0.0285745

books	0.0227528
home	0.00801154
business	0.00795813
life	0.0077979
children	0.00710356
review/theater	0.00683651
short	0.00608877
fiction	0.00544784
review/film	0.00518079
mind	0.00512738
nonfiction	0.00496715
keep	0.00491374
sports	0.00486033
topics	0.00475351
time	0.00437964
people	0.00437964
media	0.00427282
critic	0.00427282
festival	0.00427282

Topic 7

Headline

Word	Prob
u.s.	0.0255762
gulf	0.0188875
u.n.	0.0114711
iraq	0.0108473
talks	0.00793623
says	0.00790158
south	0.00762433
confrontation	0.00755502
east	0.0068619
soviet	0.00651534
bush	0.00613412
today	0.00568359
iraqi	0.00547565
israel	0.00540634
mandela	0.00526772
troops	0.00512909
africa	0.00509444
gorbachev	0.00492116
leader	0.0048865
peace	0.00460925

Topic 8

Headline

Word	Prob
u.s.	0.0125827
court	0.0102022
plan	0.00884187

drug	0.0087663
aids	0.00846401
bush	0.00759494
says	0.00680144
house	0.00638579
panel	0.00623465
study	0.00600793
bill	0.00574343
health	0.00563008
budget	0.00555451
judge	0.00479879
rights	0.00438315
abortion	0.00430758
rules	0.00408086
more	0.00377858
care	0.00374079
case	0.00362743

Topic 9

Headline

Word	Prob
home	0.0108139
child	0.00742019
journal	0.00580961
court	0.00511936
world	0.00437158
still	0.0041415
fire	0.00391142
residential	0.00373886
parent	0.00368133
close	0.00368133
resales	0.00362381
nation	0.00345125
ideas	0.00339373
york	0.00322117
case	0.00316365
improvement	0.00304861
spill	0.00304861
have	0.00293356
review/film	0.00287604
yorkers	0.00281852

Topic 10

Headline

Word	Prob
plus	0.0222902
results	0.0221726
fund	0.010057
money	0.00799859
week	0.00617538

question	0.00594013
funds	0.00588132
more	0.00570488
neediest	0.00564606
yields	0.00558725
world	0.00552844
social	0.00482268
place	0.00476387
market	0.00470505
cases	0.00429336
still	0.00417573
mixed	0.00411692
york	0.00411692
baseball	0.00370523
assets	0.00364642

Topic 11

Headline

Word	Prob
num	0.0142404
week	0.0129876
question	0.0126952
mets	0.00956318
football	0.0082686
giants	0.00822684
college	0.00822684
baseball	0.00726635
knicks	0.00726635
coach	0.00705755
wins	0.0066817
jets	0.00651466
people	0.00643114
rangers	0.00634762
victory	0.00609705
game	0.00580473
nets	0.00542888
back	0.00542888
sports	0.0050948
devils	0.00501128

Topic 12

Headline

Word	Prob
report	0.189953
earnings	0.186616
march	0.0447879
inc.	0.0430754
corp	0.0376745
sept	0.0298147
june	0.0245455

year	0.0145341
mail	0.0104066
data	0.0104066
answering	0.00935277
industries	0.00851849
bank	0.006323
international	0.00562044
group	0.00513744
savings	0.00491789
systems	0.00474225
general	0.00456661
american	0.0045227
financial	0.00430315

Topic 13

Headline

Word	Prob
unit	0.0112412
plan	0.00961538
deal	0.00887217
plans	0.00873281
sale	0.00822185
bank	0.00789669
stake	0.00678187
real	0.00552768
sell	0.00524898
debt	0.00510962
cuts	0.00506317
york	0.00506317
pact	0.00497027
u.s.	0.00492382
deals	0.00487737
offer	0.00483092
estate	0.00450576
more	0.00450576
life	0.00413415
group	0.00394835

Topic 14

Headline

Word	Prob
life	0.0238987
campus	0.0166437
island	0.0150315
journal	0.0140358
long	0.0131822
region	0.0105268
york	0.0101475
sales	0.00972071
guide	0.00938878

recent	0.00905685
connecticut	0.00896202
westchester	0.00772915
city	0.00692304
students	0.00640144
jersey	0.0060221
space	0.00521599
home	0.00507374
notebook	0.0049789
shuttle	0.00426763
more	0.00412537

Topic 15

Headline

Word	Prob
chronicle	0.0601027
town	0.0118997
fashion	0.0114165
lifestyle	0.0111749
sunday	0.0108728
street	0.0106312
quotation	0.00906071
sounds	0.00712776
wall	0.00561764
menu	0.00483238
review/music	0.00483238
evening	0.0038659
style	0.0038659
review/art	0.00374509
dinner	0.00368469
guide	0.00356388
life	0.00356388
design	0.00344307
music	0.00338266
hours	0.00338266

Topic 16

Headline

Word	Prob
east	0.0334803
europe	0.0283587
evolution	0.0166927
upheaval	0.0155072
soviet	0.0141794
economic	0.0106227
german	0.00896287
u.s.	0.00834637
germany	0.00834637
gorbachev	0.00749277
best	0.00744535

party	0.00663916
west	0.00654432
moscow	0.00602267
sellers	0.00573813
scene	0.0048371
eastern	0.00474226
union	0.00464741
talk	0.00459999
bush	0.0044103

Topic 17

Headline

Word	Prob
weds	0.0265424
paid	0.0202188
executive	0.019262
notice	0.0189708
deaths	0.0180139
dies	0.0163914
married	0.0150601
miss	0.0114407
changes	0.00940217
marry	0.00861172
bride	0.0081957
marries	0.00802929
lawyer	0.00790448
john	0.00753006
becomes	0.0066564
dead	0.0066148
david	0.00619878
robert	0.00607397
engaged	0.00603237
professor	0.00594916

Topic 18

Headline

Word	Prob
prices	0.0259818
rates	0.0191056
market	0.0153214
rise	0.0143984
dollar	0.0140754
u.s.	0.0121372
place	0.0103835
pastimes	0.0102912
stocks	0.00936822
sales	0.00904518
fall	0.00766071
trading	0.00747612
gold	0.00738382

japan	0.00729152
price	0.00719922
treasury	0.00706078
drop	0.00687618
decline	0.00673774
sharply	0.00627625
profits	0.00618395

Topic 19

Headline

Word	Prob
business	0.0368501
case	0.0201001
media	0.0177254
advertising	0.0159868
digest	0.0142906
trial	0.0104741
death	0.00886269
drug	0.00835383
police	0.00767535
u.s.	0.00759054
killing	0.00708167
guilty	0.00682724
judge	0.00678484
accounts	0.00636078
charged	0.00614876
suspect	0.00606395
held	0.00597914
charges	0.00551268
brooklyn	0.00547027
killed	0.00542787

Topic 19

Headline

Word	Prob
business	0.0368501
case	0.0201001
media	0.0177254
advertising	0.0159868
digest	0.0142906
trial	0.0104741
death	0.00886269
drug	0.00835383
police	0.00767535
u.s.	0.00759054
killing	0.00708167
guilty	0.00682724
judge	0.00678484
accounts	0.00636078
charged	0.00614876

suspect	|	0.00606395
held	|	0.00597914
charges	|	0.00551268
brooklyn	|	0.00547027
killed	|	0.00542787

8.5.3.2 Topic Distributions

```
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0
0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
0,0,0,0,0.5,0,0,0,0,0,0,0.5,0,0,0,0,0,0,0,0
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0
0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0
0,0,0,0,0,0,0,0,0.166667,0,0,0,0,0,0,0,0,0,0,0.833333
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0
0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0
0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0
```

8.6 Plots

8.6.1 Informed priors

Following plot shows that it is necessary to have different priors for word distribution over body and headline. Number of headline words is far less than number of body words. For the purpose of bringing curves on the same scale number of headline words are scaled linearly to bring it on same scale.

Acknowledgement

I extend my sincerest gratitude to Prof. David M. Blei, Prof. Garud Iyengar, Prof. Martin Haugh, Francois Fagan and Kui Tang for collaboration and guidance through the project. I also like to thank Michael Dewar and Evan Sandhaus for letting us use *The New York Times* corpus for our research.

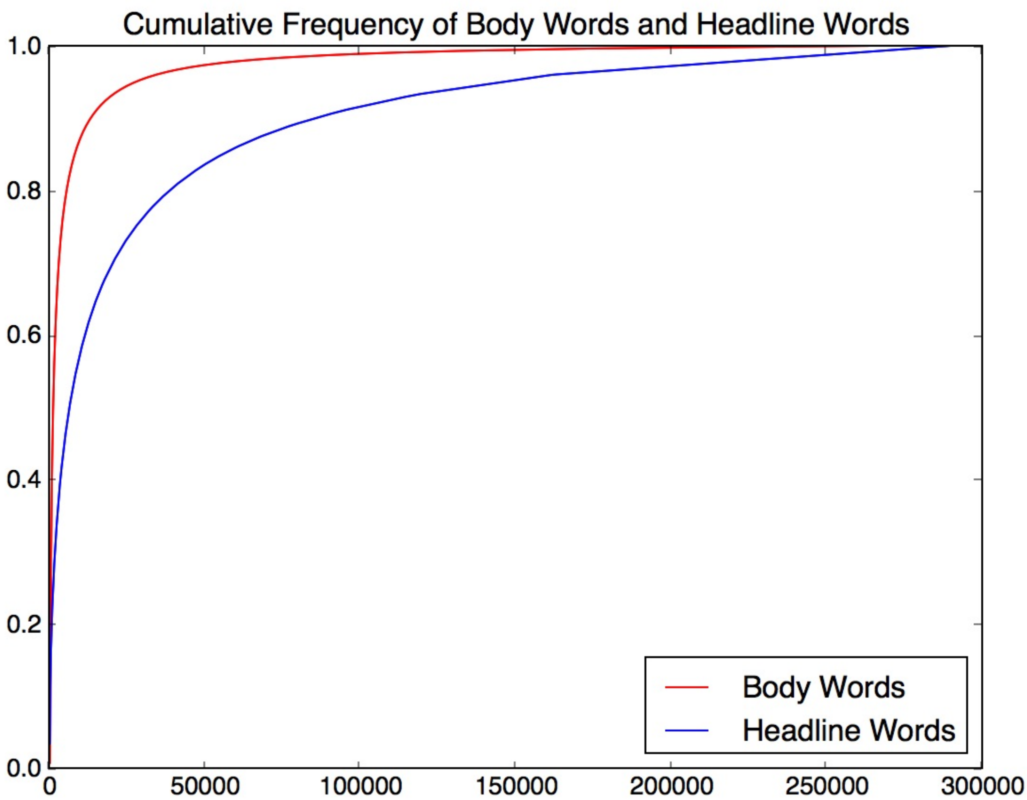

Figure 3: Informed priors β and $\hat{\beta}$