UNIVERSITY OF
CAMBRIDGE

Latent Dirichlet Allocation

5568D

6th December 2019

1 Task A

In this task, a naive implementation of finding **topics** of the word documents is investigated. It consists of looking at the frequency(c_m) of each unique word(w_m) from the total words(N_d) in each document(d) from all the documents(D) and based on them find a distribution ($\boldsymbol{\beta}$) of the topics.

It is required to find $\boldsymbol{\beta}$ such that the *Log Likelihood*($\log p(\mathbf{w}|\boldsymbol{\beta})$) of the words in the training dataset is maximised. Therefore the problem can be formulated as:

$$\begin{aligned} \max \quad & \log p(\mathbf{w}|\boldsymbol{\beta}) = \sum_{m=1}^{M} c_m \log \beta_m \\ \text{subject to} \quad & \sum_{m=1}^{M} \beta_m = 1 \end{aligned} \tag{1}$$

The expression above is a constrained optimisation which can be transformed into unconstrained optimisation using a *Lagrange Method*. The solution of this unconstrained optimisation can be found analytically:

$$\beta_m = \frac{c_m}{N} \tag{2}$$

where N is the total number of words in the training dataset(A). Using Equation 2 above, the $\boldsymbol{\beta}$ values for A can be found. The $\boldsymbol{\beta}$ found by this method reflects the aggregation of all the documents.In *Figure 1* below, the top 20 values from the $\boldsymbol{\beta}$ distribution are displayed:

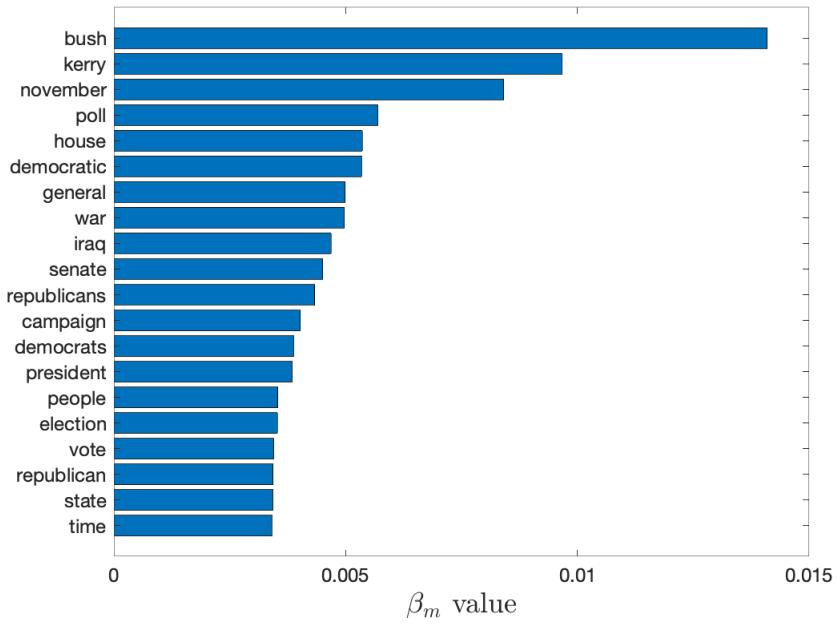


Figure 1: Words with the highest probability in A

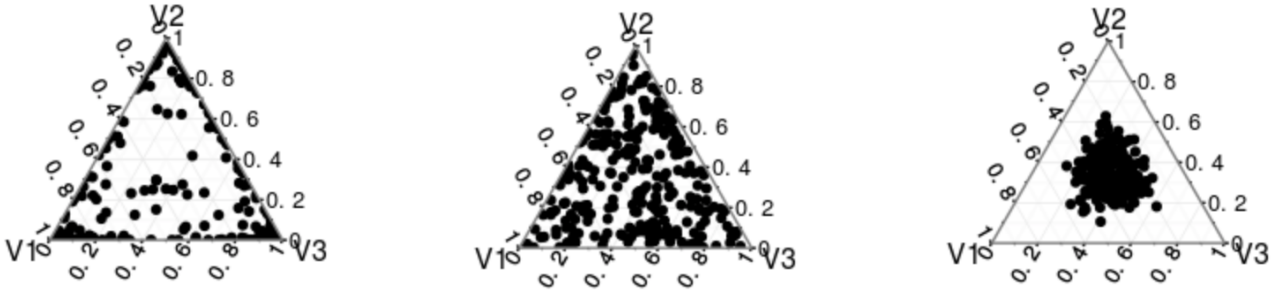
Using the $\boldsymbol{\beta}$ values calculated the *Log Likelihood* of the test data can be found by using the formula in 1 but with the words from the test data($\tilde{\mathbf{w}}$). The likelihood of the test data will depend very much on what words are used in the testing documents. If a new word($\tilde{w}_j$) is encountered, β_j will be zero. Therefore the log probability will be equal to $\log p(\tilde{\mathbf{w}}|\boldsymbol{\beta}) = -\infty$. Intuitively, the highest log probability will be attained if the test document contains only the highest frequency word found in the A.

2 Task B

In this task, a *Bayesian Inference* using a symmetric *Dirichlet* prior is investigated. Firstly, the symmetric *Dirichlet* is further explored. The probability density function of a symmetric *Dirichlet* distribution is given as:

$$Dir(\boldsymbol{\beta}|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{m=1}^{M} \beta_m^{\alpha-1} \tag{3}$$

In the expression above the *Dirichlet* distribution is characterised by a single scalar value α called the concentration parameter. In the figure bellow displays the behaviour of *Dirichlet* distribution for different values of α is given.



(a) $\alpha = 0.2$ (b) $\alpha = 1$ (c) $\alpha = 10$

Figure 2: Dirichlet Distribution for different values of concentration parameter [1].It can be observed that as α increases(2c) the output distribution will be close to a uniform distribution. From 2a as α gets smaller the output distribution will result in most of the mass being 0 but not all of it [2].

From the figures above it can be visualised what the Dirichlet prior does is to output a distribution over distribution. Now that the prior of $\boldsymbol{\beta}$ has been studied it is now time to look at how the posterior is being evaluated using this prior:

$$\begin{aligned} p(\boldsymbol{\beta}|\mathbf{w}) &\propto p(\mathbf{w}|\boldsymbol{\beta})p(\boldsymbol{\beta}) \\ &\propto \prod_{m=1}^{M} \beta_m^{c_m} \prod_{m=1}^{M} \beta_m^{\alpha-1} \\ &\propto \prod_{m=1}^{M} \beta_m^{c_m+\alpha-1} \\ &\propto Dir(\boldsymbol{\beta}|\mathbf{c}+\boldsymbol{\alpha}) \end{aligned} \tag{4}$$

Therefore the predictive probability for that the next word is $\tilde{w}_j$ will be found by *Bayesian Inference*:

$$p(\tilde{w}_j|\mathbf{w}) = \int p(\tilde{w}_j|\boldsymbol{\beta})p(\boldsymbol{\beta}|\mathbf{w}) = \frac{\alpha + c_j}{M\alpha + n} \tag{5}$$

It is observed why a *Bayesian model* might be preferred instead of Max Likelihood model. The reason is that even if a new word is encountered where the count will be zero the probability will not be zero due to α parameter. It is observed that if α is zero the same expression for likelihood as in previous task is obtained. As mentioned above is observed why a high value of α the probability of a new word regardless weather is a common or a rare word will have a uniform probability equal to $\frac{1}{M}$. A small value of α it will ensure that common words have roughly the same probability while rare words it would extra weight on them.

3 Task C

In this task the *Bayesian Inference* model defined in **Task B** is applied to the test data. It is now important to discuss whether a Multinomial or Categorical distribution to compute the log probability of a test document. In a Multinomial Distribution it is looked at the counts of the words in the document while in a Categorical Distribution it is observed the sequence of words. Ideally it is desired to model a document such that the dependencies between the words would be taken into account. However such a model would be complex and it would work well only for a small vocabulary where the sequence of words is comparable size but for large vocabulary by using the *Bayesian Inference* model it will fail to generalise and therefore it will not be accurate [3]. As a result, Multinomial distribution is preferred for the current task. The log probability for test document with ID 2001 can be found by:

$$\log p(\tilde{\mathbf{w}}_{2001}|\mathbf{w}, \boldsymbol{\beta}) = \sum_{m=1}^{M} c_m^{\tilde{\mathbf{w}}_{2001}} \log(\boldsymbol{\beta}_m^{\mathbf{w}}) = \sum_{m=1}^{M} c_m^{\tilde{\mathbf{w}}_{2001}} \log(\frac{\alpha + c_m^{\mathbf{w}}}{M\alpha + N}) \tag{6}$$

The equation above can be generalised to find all probabilities in test documents B and the total probability. However to understand how well a probability distribution or predicts a sample it is required to look at the *perplexity* of the document. A low perplexity indicates the probability distribution is good at predicting the sample. The *perplexity* is defined as:

$$\text{perp} = \exp(\frac{-\log p(\tilde{\mathbf{w}}|\mathbf{w}\boldsymbol{\beta})}{N}) \tag{7}$$

The results for the log probability and perplexity for document 2001 and all other documents in B can be seen below:

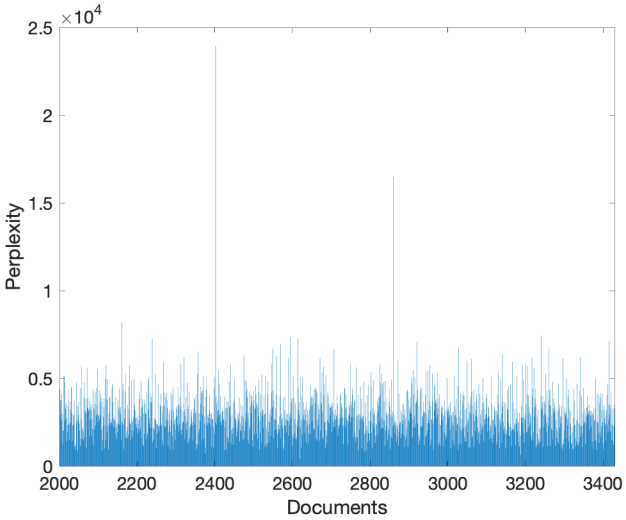


Figure 3: Perplexity of all the documents in B.

Documents	Log Probability	Perplexity
Document 2001	-3691.5	4401.9
All documents in B	-1547.5×10^3	4188.9×10^3
All documents in B with large α	-1731×10^3	9875.6×10^3

Table 1: Perplexity and Log Probability

It is observed that documents have different values of perplexity. The reason for this behaviour is the fact that if there are more common words ,therefore, the β value will be higher than zero which would make the log closer to zero than to $-\infty$which would cause the perplexity to be lower. If the *alpha* parameter increases to large value, essentially every β term will have probability $1/M$ which judging from the value of perplexity in *Table 1* it performs worse.

4 Task D

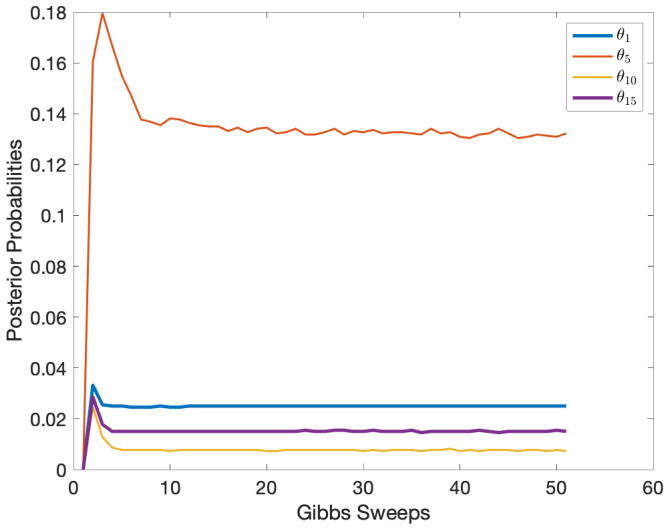
In this task, a *Bayesian Mixture of Multinomials* model is applied to the KOS dataset. Using this model the latent topic assignments($\mathbf{z}_d$) can be sampled using a collapsed Gibbs Sampler which will be more efficient as it samples in a lower dimensional space[4]:

$$\begin{aligned} p(z_d k|\mathbf{w}_d, \mathbf{z}_{-d}, \boldsymbol{\beta}, \alpha) \propto & \underbrace{p(\mathbf{w}_d|\beta_k)}_{\text{Likelihood given topic}} \times \underbrace{p(z_d = k|\mathbf{z}_{-d}, \alpha)}_{\substack{\text{Prior based on} \\ \text{counts of other documents}}} \\ &= p(\mathbf{w}_d|\beta_k) \frac{c_{-d,k} + \alpha}{\sum_{i=1}^{K} c_{-d,i} + \alpha} \end{aligned} \tag{8}$$

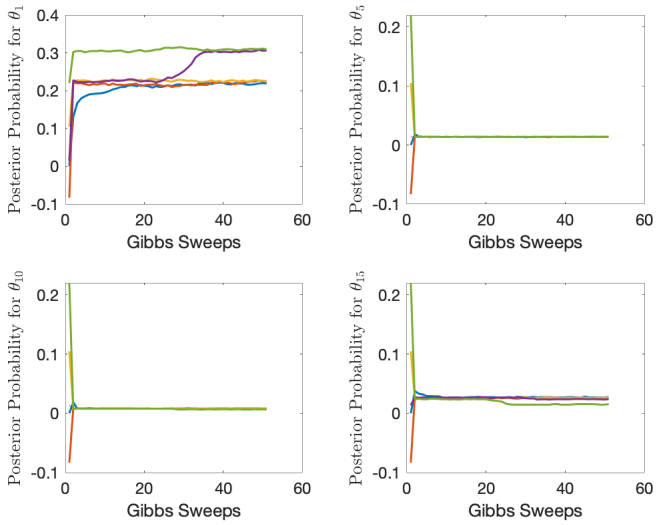
The posterior probabilities of each of the mixture components($\boldsymbol{\theta}$) can be computed iteratively at each Gibbs sweep until it will converge to the true distribution(stationary distribution):

$$\theta_k = \frac{c_k + \alpha}{\sum_{i=1}^{K} c_i + \alpha} \tag{9}$$

In *Figure 4a* below the mixture components $\theta_1, \theta_5, \theta_{10}$ and θ_{15} are displayed as a function of the Gibbs sweeps. As observed it would look like the components converge to a stationary distribution. However from *Figure 4b* it can be observed that there could be multiple stationary distributions they could converge to.



(a) Posterior Probabilities for θ_1,θ_5,θ_{10} and θ_{15}



(b) Convergence of θ_1,θ_5,θ_{10} and θ_{15} for different initialisations

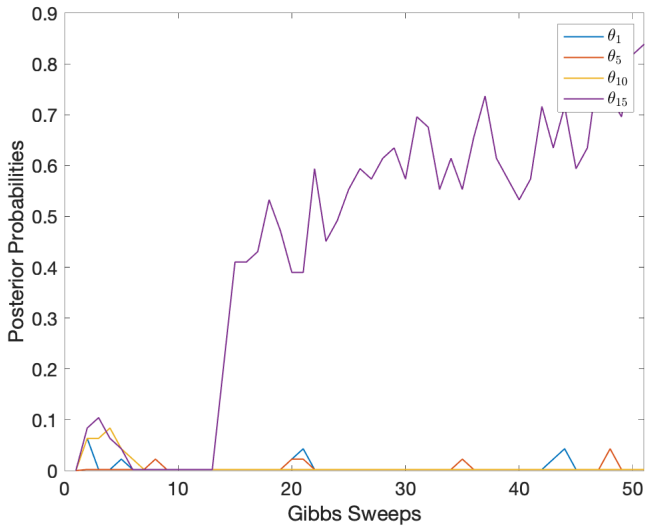
Therefore it is observed that the main problem with Gibbs sampling is that depending on where it is initialised it is likely that one parameter gets stuck into a single mode without actually visiting all the modes. This is a problem called *label switching*. A convergence of the Gibbs Sampling would occur if all the mixture components would visit all the $K!$ symmetric modes[5]. To address this problem one could incorporate a Metropolis–Hastings move that proposes a random permutation of the labels[5]. However, even if the model stays in this local modes it is observed that the value of the total perplexity for the test dataset B is 2122.6 which is significantly better than the *Bayesian Inference*.

5 Task E

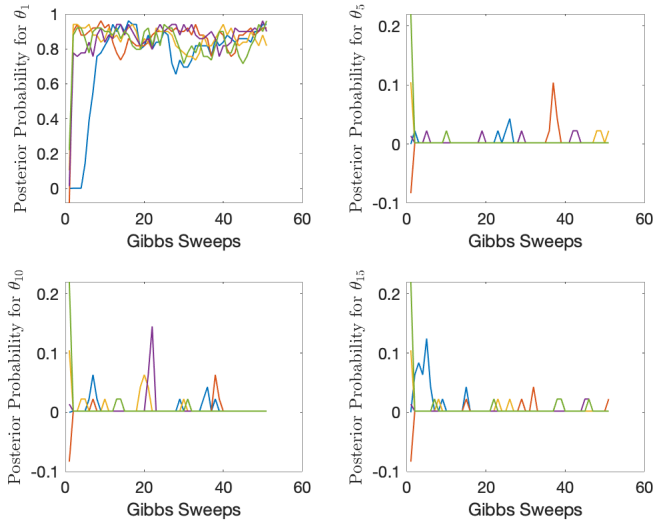
In this task a new model is being investigated, *Latent Dirichlet Allocation(LDA)*. This model allow for every word in a document to be drawn from a different topic and every document has its own distribution over topics. The predictive distribution for a single word (z_{nd}) can be expressed as:

$$p(z_{nd} = k|\mathbf{z}_{-nd}, w_{n,d}, \alpha, \gamma) \propto \frac{\alpha + c^k_{-nd}}{\sum_{i=1}^{K} \alpha + c^i_{-nd}} \frac{\gamma + \tilde{c}^k_{w_{-nd}}}{\sum_{m=1}^{M} \gamma + \tilde{c}^k_{-m}} \tag{10}$$

Applying Gibbs sampler to the *LDA* model has the same problems as in **Task D** as observed in the *Figures 5a and 5b* below:



(a) Posterior Probabilities for θ_1,θ_5,θ_{10} and θ_{15}

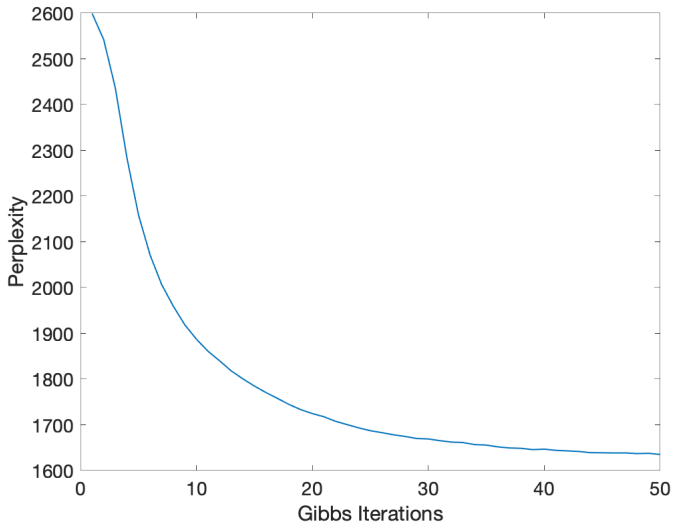


(b) Convergence of θ_1,θ_5,θ_{10} and θ_{15} for different initialisations

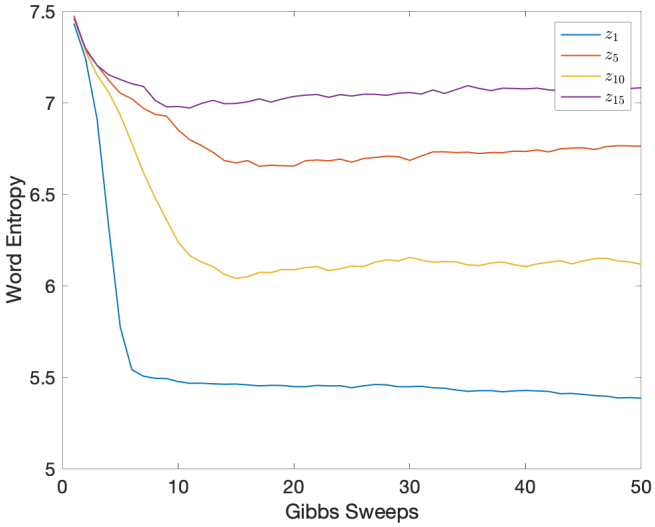
It is observed that using an *LDA* model most of the words could be explained by a single topic. To further explain this behaviour it is looked at the word entropy for each topics. A large entropy corresponds to very specific topic while a small entropy corresponds to a broad topic. The entropy is given by the following expression:

$$H(k) = -\sum_{m=1}^{M} (\beta_{km}) \ln(\beta_{km}) \tag{11}$$

The results can be observed in *Figure 6b* below. As expected in the begging the entropies are very high due to the fact that all topics have the same probability. Together with *Figure 6a* it is observed that even if the perplexity will continue to decrease, the computational complexity generated by increasing the Gibbs sweeps would not be a good trade-off. It is also observed that compared with other model the *LDA* model has the lowest perplexity of 1635.3.



(a) Total perplexity of all the documents in B as a function of the Gibbs sweeps



(b) Word entropy for topics 1,5,10 and 15

WORD COUNT= 1050

References

[1] Stack Exchange user. α value meaning. https://stats.stackexchange.com/questions/244917/what-exactly-is-the-alpha-in-the-dirichlet-distribution, 13 April 2017. [Online; accessed 3-December-2019].

[2] Wikipedia contributors. Concentration parameter. https://en.wikipedia.org/wiki/Concentration$_p$*arameter*, 20 *December* − 2019].

[3] Andrew McCallum, Kamal Nigam, et al. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, pages 41–48. Citeseer, 1998.

[4] K.P. Murphy. *Machine Learning: A Probabilistic Perspective*. Adaptive Computation and Machine Learning series. MIT Press, 2012.

[5] Carlos E. Rodríguez and Stephen G. Walker. Label switching in bayesian mixture models: Deterministic relabeling strategies. *Journal of Computational and Graphical Statistics*, 23(1):25–45, 2014.