4F13: Latent Dirichlet Allocation - Coursework 3

Raghavendra Narayan Rao

December 5, 2025

Abstract

This report investigates multinomial language models and Latent Dirichlet Allocation (LDA) for text modelling on the the Kos dataset ($W = 6906$ words, 2000 training docs, 1000 test docs). We compare Maximum Likelihood, Bayesian single model, Bayesian Mixture of Multinomials (BMM), and LDA models. LDA achieves perplexity of 1647, significantly outperforming Bayesian single model (2697) and uniform (6906) baselines, demonstrating the benefit of modelling latent topic structure.

1 Task (a) - Maximum Likelihood (ML) Multinomial

The ML estimate for word probabilities is the normalized word counts [1]:

$$\hat{p}_w^{ML} = \frac{n_w}{N} = \frac{n_w}{\sum_{w'} n_{w'}}$$

where n_w is the count of word w across all training documents and $N = 275,546$ is the total word count.

Test Set Log Probability

Highest: $\log(0.0141) \approx -4.26$. If the test set only contained "bush", which is the most frequent word in the training set, the test set probability will be maximised.

Lowest: $-\infty$ since ML assigns $p_w = 0$ to unseen words in the training set. Of $W = 6906$ vocabulary words, 14 have zero training count. If any of these appear in the test set, the probability of test set $= 0$.

Implications: ML should not be used in situations where the test set is slightly different from the training set. Any new data in test set (not seen in training set) willl be treated as impossible despite these data points being part of the vocabulary. This motivates a need for priors on the vocabulary.

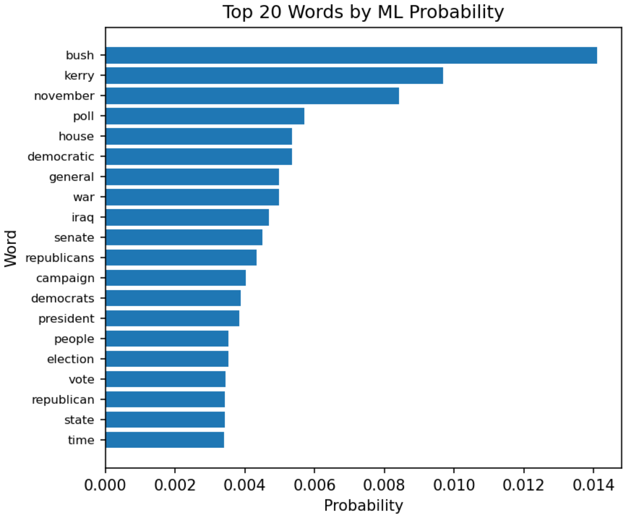


Figure 1: Top 20 words in the Kos training dataset. The most common word, "bush", has 1.4% probability.

2 Task (b) - Bayesian Inference with Dirichlet Prior

With a symmetric Dirichlet prior Dir($\alpha, ..., \alpha$), the posterior predictive is:

$$\hat{p}_w^{\text{Bayes}} = \frac{n_w + \alpha}{N + W\alpha}$$

The Dirichlet prior adds α *pseudo counts* to each word (Laplace smoothing when $\alpha = 1$).

	ML	**Bayesian**
Formula	n_w/N	$(n_w + \alpha)/(N + W\alpha)$
Unseen words	0	$\alpha/(N + W\alpha)$
Common words	n_w/N	$\sim n_w/(N + W\alpha)$

Conclusions from Figure 2: (Sparce prior) Small α (0.1) keeps common word probabilities near ML while giving unseen/rare words small but non zero probability. This solves the $-\infty$ test set log probability problem from part (a). (Uniform prior) Large α (100) over smooths the posterior distribution. The distribution shrinks towards uniform $1/W$, losing discriminative information. This impacts rare words much more than common words. The choice of α trades off between staying close to data (small α) versus hedging against unseen words in training set (large α).

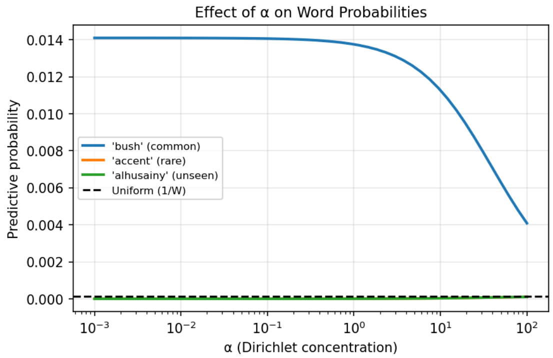


Figure 2: As α increases, all probabilities converge to uniform $1/W$. Small α preserves ML structure while avoiding zero probabilities.

3 Task (c) - Test Document Log Probability and Perplexity

Multinomial vs Categorical: We use the categorical distribution $P(\mathbf{n}) = \prod_w p_w^{n_w}$, not the multinomial. The multinomial coefficient $N!/\prod_w n_w!$ includes the ordering of words, which should not be included when considering the probability of a specific sequence. This is not included when calculating perplexities either. The log probability of test document with ID 2001 is -3691.

Perplexity: Perplexity $= \exp\left(-\frac{1}{N_{\text{test}}} \sum_w n_w \log p_w\right)$

Perplexity is a measure of how uncertain the model is (the effective vocabulary size the model is "uncertain" between). Higher perplexity means the model is more uncertain.

Document/Model	**Perplexity**
Document 2001	4399
Overall test set (all B)	2697
Uniform multinomial	6906 (= W)

Why uniform perplexity = W: $P(w_1, \ldots, w_n) = \frac{1}{W^n}$ so $\exp\left(\frac{1}{n}P(w_1, ..., w_n)\right) = \exp(\log W) = W$. This is the *maximum possible* perplexity as you cannot have more than W plausible choices.

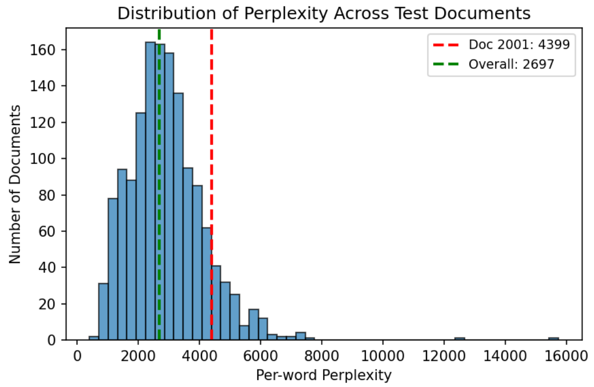


Figure 3: Perplexity varies widely across documents. Doc 2001 has unusually high perplexity, suggesting atypical vocabulary.

Conclusions from Figure 3: Perplexity varies from ∼1500 to ∼6000 across test documents. Documents using common words ("bush", "kerry") score well. Those with rare/unusual vocabulary score poorly. Doc 2001's high perplexity (4399 versus 2697 average) indicates it contains atypical content for this corpus. The Bayesian model (2697) substantially beats uniform (6906), confirming it captures meaningful word frequency structure.

4 Task (d) - BMM Gibbs Sampling

The Bayesian Mixture of Multinomials assigns each *document* to one of K clusters via Gibbs sampling. Here K is the number of mixture components, α is the Dirichlet prior parameter over mixture proportions, and γ is the Dirichlet prior parameter over word distributions within each cluster. Mixing proportions: $\pi_k = (n_k + \alpha)/(D + K\alpha)$.

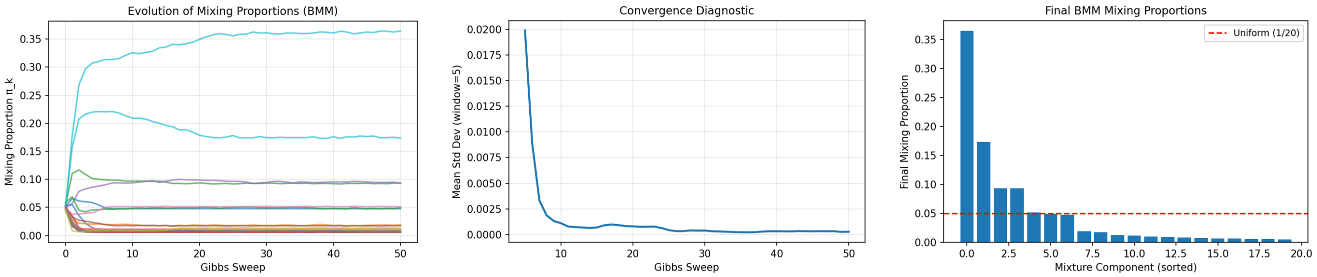


Figure 4: BMM Gibbs sampling ($K = 20$, $\alpha = 10$, $\gamma = 0.1$). Left: Mixing proportions over sweeps. Middle: Rolling std dev (convergence). Right: Final proportions.

Convergence Conclusions (Figure 4): The sampler converges by sweep 15. The std dev plot also agrees with this. Mixing proportions stabilize after initial burn-in (sweeps 0 to 5). The final distribution is highly non-uniform as one cluster dominates ($\sim$36%) while several have <1% of documents. This suggests the effective number of clusters is much less than $K = 20$ as many clusters are essentially empty.

5 Task (e) - Latent Dirichlet Allocation

LDA assigns each word (not document) to a topic, allowing documents to exhibit multiple topics [2]. We use $K = 20$, $\alpha = 0.1$, $\gamma = 0.1$.

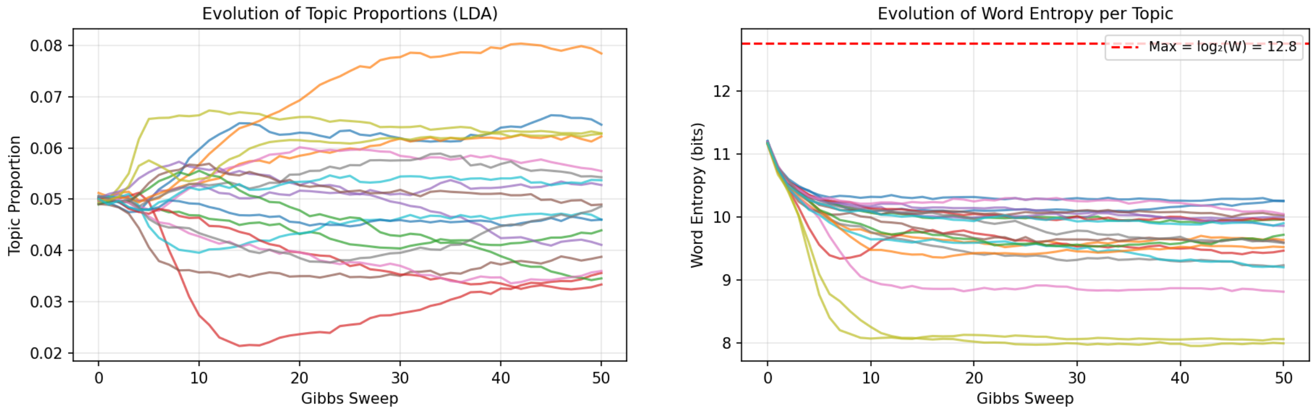
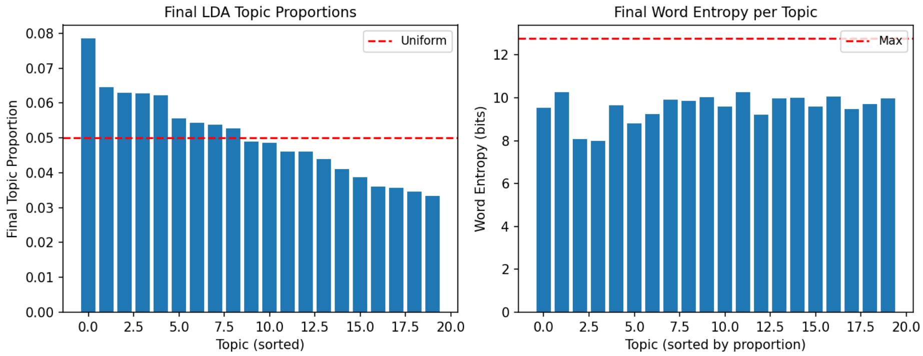


Figure 5: LDA ($K = 20$, 50 sweeps). Left: Topic proportions stabilize by sweep 30. Right: Word entropy (bits) decreases as topics become focused.

Conclusions from Figure 5: Topics differentiate quickly (within 15 sweeps) and stabilise by sweep 30. Word entropy drops from $\sim$11 bits (random) to 9 bits, indicating topics develop distinctive vocabularies. The 2 to 4 bit reduction from maximum ($\log_2 W = 12.75$) means each topic focuses on roughly $2^{8 \text{ to } 10} = 252$ to 1000 effective words rather than all 6906. Lower entropy topics (8 bits) are more specialized while higher entropy topics (10 bits) are more general topics.


Final LDA Topic Proportions

Final Word Entropy per Topic

Figure 6: Final state: topics have balanced proportions (0.03 to 0.08) and varying specialization (entropy 8 to 10 bits).

Perplexity Comparison:

Model	**Perplexity**
Uniform	6906
Bayesian unigram	2697
LDA (K=20)	1647

LDA's 39% improvement over unigram demonstrates that modelling latent topic structure substantially helps. Documents are better explained as mixtures of topics than draws from a single distribution.

Topics: The learned topics are semantically meaningful. This is shown using the top sampled words:

Topic	**Top Words**
Iraq/War	iraq, war, military, troops, soldiers, iraqi
Elections	kerry, bush, vote, campaign, election, polls
Senate	senate, race, republicans, democrats, house
Media	news, media, press, story, article

Are 50 sweeps adequate? Yes as both topic proportions and entropy stabilize by sweep 30. Additional sweeps would yield marginal improvement.

Is $K = 20$ appropriate? Yes as all 20 topics have non-negligible proportions. Moreover, the topics are semantic (see above). There is also significant perplexity improvement achieved. However, some topic overlap exists (multiple poll-related and Iraq-related topics), suggesting $K = 15$ might suffice.

References

[1] C. M. Bishop, *Pattern Recognition and Machine Learning.* Springer, 2006.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.