CS 635 Assignment 2 Report

~ Shubham Lohiya, 18D100020

BAG OF WORDS REPRESENTATIONS (BoW)

The following two variations have been considered:

- **Binary-BoW**: Only considers if a certain word in the vocabulary is present in the document or not. Word counts or repetitions are not considered.
- **Count-BoW**: This representation consists of all the words from the vocabulary that are contained in the document along with their word counts.

Note: The vocabulary is constructed from all the tokens present in the training corpus. Token text has directly been taken. For better results, vocabulary can also be constructed from word lemmas.

Hyperparameters legend → **k:** Laplace Smoothing Parameter, $\boldsymbol{\alpha}$**:** Trade-off parameter between simple average and weighted average. Used in estimation of poisson parameters.

MULTINOMIAL TOPIC MODEL FOR DOCUMENT GENERATION

Hyperparameter choices: **k = 0.1 for both B-BoW and C-BoW.** This hyperparameter was chosen based on tuning experiments performed using a validation set constructed from the training data.

Methodology: Same as detailed in the problem statement

Note: Multinomial model was evaluated on both B-BoW and C-BoW representations.

POISSON TOPIC MODEL FOR DOCUMENT GENERATION

Hyperparameter choices: $\boldsymbol{\alpha}$ **= 0.9, and k = 1e-5**. These hyperparameters were chosen based on tuning experiments performed using a validation set constructed from the training data.

Methodology: The method used for estimating parameters of the poisson distributions is similar to the one given in https://aclanthology.org/W03-1105.pdf. For estimating $P(X_v = i \mid c)$ [probability that word v occurs i times in document of class c], smoothed counts of word v were obtained for all documents belonging to class c. To account for document length differences, these smooth counts were scaled by the laplace smoothed document length. Then the length-scaled estimate of the poisson parameter was done by taking a convex combination (using trade-off α) of the simple average and the weighted average (using smoothed document lengths to calculate weights) of these scaled-smoothed-word-counts. Let's call this estimate $\lambda'_{v,c}$. Then the log likelihood of a document:

$$P(d \mid c) \propto \sum_{v \in d} x_v * \left(\log(\lambda'_{v,c}) - \lambda'_{v,c}\right)$$

The $(-\lambda'_{v,c})$ term is also inside the bracket in order to adjust the poisson document generator parameter to the document length. Thus, this formulation helps approximate the likelihood for the Naïve Bayes Classifier. Note that since document length adjustment is being done for $\lambda'_{v,c}$, we should have another term on the RHS above: $l_d \log(l_d)$ where $l_d = \sum_v x_v$. Since this term is class independent, we can drop it from our optimization objective.

Note: Poisson model was evaluated on only the C-BoW representation as word counts are modelled to be drawn from poisson distributions.

CLASS PRIORS: Priors were estimated as the proportion of class documents in the training data

CLASSIFICATION PERFORMANCE:

	Naïve Bayes - Multinomial		Naïve Bayes - Poisson
	Binary-BoW	Count-BoW	Count-BoW
Training Data	99.09%	98.66%	99.69%
Testing Data	82.46%	82.79%	**83.19%**