**Homework 3: Language Models**

Name: Prajwala Mugajji Shambulingappa

**Aim:**

Work with Language Models to analyse a dataset using NLTK package.

**Observation Criteria and Analysis:**

1. Dataset: As part of the assignment, we are using dataset provided by Kaggle which has a large number of Wikipedia comments that have been labelled by toxic behaviour by human reader. To load dataset to train and test model “pandas” package, generally used for data manipulation, is used.
2. Training dataset, used for training the language model has to be normalized. As part of normalization, we define a normalization function, which will be used once data is ready, where lowering upper case characters to lower case characters, replacing special characters, numbers and punctions and tokenizing the string is performed on “comment\_text” column’s every row.
3. Initially, using the nltk’s features ‘Stopword’ list, excluded the stopwords such as "a", "an", "the", "and", "but", "in", "is", and "of". But this normalization step resulted in model’s poor performance, as the text context of the comments was lost.
4. Moving on, “Stopword” exclusion was removed and additional spaces were removed.
5. For training model, columns such as “severe\_toxic”, “obscene”, “threat”, “insult” and “identity\_hate” was removed. Additionally, three different datasets were created. First set of data had all comments labeled by toxic behavior, where; 0 denotes non-toxic behavior and 1 denotes toxic behavior. After which we normalize the datasets created.
6. Training the model:
   1. For training the model, a “nltk” function called “padded\_everygram\_pipeline” is used. This feature is used to build a padded 2-gram/bigram language model that is used to create a pipeline of text preprocessing steps that prepare a corpus for training using the n-gram language models. Two parameters are passed to the feature specifying the order of the n-gram and dataset for training the model.
   2. As second step, smoothing is added for the training data using Laplace smoothing. Smoothing is a technique used to smooth categorical data by adding a constant value, this method is used to adjust the probability estimates of words that do not appear in a training dataset.
   3. As third step, “fit” method is used to train the model with the processed dataset. The first argument passes to the model is the preprocessed training dataset that the model will learn from, and the second argument is the vocabulary that is a set of tokens that the model will use during model training.
   4. As fourth step, now that a trained language model is ready, with the help of a “pickle” package, the model is saved using the “dump” feature of pickle.

All of these steps are followed to train 3 language models, using the three-dataset created using the toxic label. First model is trained using all the datasets, second model is trained using dataset that has only toxic label ‘0’ and third model is trained using dataset that has only toxic label ‘1’.

1. After which models trained are tested, and log scores are calculated.
2. As step one, datasets are divided into two subsets using the “text\_label.csv” using the column “toxic”. Dataset 1 consists of all the datasets with “toxic” labeled as ‘0’ and ‘1’. Dataset 2 consists of all the datasets with “toxic” labeled as ‘0’ and Dataset 3 consists of all the datasets with “toxic” labeled as ‘1’.
3. Step 2: Load the trained models using pickle.
4. Step 3: Calculate the log score.

**Summary:**

1. Language Models was successfully implemented to analyse a dataset which has a large number of Wikipedia comments that have been labelled by toxic behaviour using NLTK package.
2. Below is the comparative table between the models for the datasets created:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Data\_all | Data\_toxic | Data\_non\_toxic |
| Lang Model 1 | -12.419897233980555 | -14.12960198684052 | -14.56757603408250 |
| Lang Model 2 | -12.640925646188009 | -14.273360474276005 | -14.54651800123752 |
| Lang Model 3 | -12.070496448630424 | -13.242600599310313 | -14.028768118515595 |

1. In conclusion, it can be said that all models have similar performance.