**Classification of small-size documents Using Multinomial Naïve Bayes**

Hamid Hooshmandi

CSCI 4141

Final Project

Hamid42643@gmail.com

# INTRODUCTION/ Hypothesis

While Multinomial Naïve Bayes has been used before and found and effective at classification of documents (Andrew McCallum), I examined application MNB for classification of small documents (4 to 110 words) and tried to find out if it is an effective tool for this purpose. The result of the experience shows that TF normalization, words stemming and document length normalization improve the accuracy. While applying IDF significantly decrease the accuracy.

## Multinomial naïve Bayes

Multinomial naïve Bayes is special form of Naïve Bayes algorithm. While Simple Naïve Bayes only takes into account the present or absent of a words. Multinomial Naïve Bayes also takes into account word frequencies of a document. NBM is simple to implement, computationally inexpensive and gives good accuracy on task of text classification. Although it is still outperformed by Support Vector Machines. (Ashraf M. Kibriya)

# EXPERIMENT SETUP

The training set includes two data sets. Data set 1 which includes 7408 instances from 24 different categories. The data set 2 which includes 11903 instances from 24 unique categories. Both the datasets are quotes from famous authors the quotes are gathered from http://www.brainyquote.com/ Website using a crawler. Each category in Data set 1 includes different number of documents Figure 1. While Data set 2 contains uniform number of documents Figure 2. `

|  |
| --- |
| **Figure 1: Data Set 1 has different number of documents for each categories** |
|  |
|  |
| **Figure 2: Data Set 2 has more uniform number of documents for each category** |
|  |

# FEATURE GENERATION

Each document converted to a high dimension vector. Indicating the present or absent of a word in the document. All the words have been turned to lower case. Alphabetic string tokenize with a list of stop words (Weka 3-6) has been used. Naïve Bayes Multinomial has been used for all the experiences with 5 fold cross validation for evaluation. Different stemmer and IDF TF has been used. The results are shown below.

|  |
| --- |
| **Table 1: setup for different experiences** |
| |  |  |  | | --- | --- | --- | | **Option** | **Top** | **Description** | | IDF | **variable** | First | | lowerCase terms | Yes | whether or not word terms are converted to lower case | | Stopwords list | Yes | stemming algorithm used | | tokenizer | Alphabetic | tokenization method used | | TF | **variable** | word frequencies transformed to: log(1+fij) | | normalizeDoc Length | **variable** | Whether or not word frequencies for a document should be normalized . | | stemmer | **variable** | The stemming algorithm used | |

# EVALUATION

## Results

Table 2 shows accuracy of classification using different tokenization techniques. the results shows, besides size of training set, most important factors in accuracy of small documents classification using NBM algorithm are TF normalization, stemming and document length normalization. Furthermore Inverse document frequency significantly decreases results accuracy.

# REFERENCES

Andrew McCallum, K. N. (n.d.). A Comparison of Event Models for Naive Bayes Text Classi. Pittsburgh, PA.

|  |
| --- |
| **Table 2: Classification accuracy using different tokenization techniques** |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | | **Experience no.** | **Data set** | **Classification and**  **Evaluation methods** | **Tokenization** | **Accuracy** | | 1 | Data Set 1 | NB multinomial  5% cross validation |  | 74% | | 2 | Data Set 1 | NB multinomial  5% cross validation | TF normalization | 77% | | 3 | Data Set 1 | NB multinomial  5% cross validation | TF normalization  without stemming | 78% | | 4 | Data Set 2 | NB multinomial  5% cross validation | TF normalization  Lovin stemming | 79% | | 5 | Data Set 2 | NB multinomial  5% cross validation | TF normalization  Porter stemmer  stemming | 83% | | 6 | Data Set 2 | NB multinomial  5% cross validation | TF normalization  Porter stemmer  document length normalization | 84% | | 7 | Data Set 2 | NB multinomial  5% cross validation | TF normalization  Porter stemmer  document length normalization  IDF | 51% | |
|  |
| **Table 3: confusion matrix for experience no. 6** |
|  |
|  |
| **Table 4: Accuracy by class** |
| |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | |  | **TP** | **FP** | **Precision** | **Recall** | **F-Measure** | **Class** | |  | **0.904** | **0.01** | **0.83** | **0.904** | **0.866** | **Education** | |  | **0.932** | **0.012** | **0.803** | **0.932** | **0.863** | **Health** | |  | **0.976** | **0.017** | **0.76** | **0.976** | **0.855** | **Morning** | |  | **0.928** | **0.015** | **0.766** | **0.928** | **0.839** | **Money** | |  | **0.807** | **0.009** | **0.827** | **0.807** | **0.817** | **Business** | |  | **0.927** | **0.012** | **0.795** | **0.927** | **0.856** | **Family** | |  | **0.873** | **0.012** | **0.783** | **0.873** | **0.825** | **Peace** | |  | **0.941** | **0.005** | **0.891** | **0.941** | **0.915** | **Dreams** | |  | **0.841** | **0.008** | **0.832** | **0.841** | **0.836** | **War** | |  | **0.845** | **0.005** | **0.882** | **0.845** | **0.863** | **Art** | |  | **0.858** | **0.006** | **0.864** | **0.858** | **0.861** | **Home** | |  | **0.895** | **0.007** | **0.845** | **0.895** | **0.869** | **Happiness** | |  | **0.872** | **0.004** | **0.9** | **0.872** | **0.886** | **Success** | |  | **0.857** | **0.005** | **0.878** | **0.857** | **0.867** | **Change** | |  | **0.735** | **0.005** | **0.865** | **0.735** | **0.795** | **Nature** | |  | **0.896** | **0.004** | **0.899** | **0.896** | **0.898** | **Truth** | |  | **0.825** | **0.004** | **0.898** | **0.825** | **0.86** | **Power** | |  | **0.638** | **0.005** | **0.828** | **0.638** | **0.721** | **Time** | |  | **0.732** | **0.006** | **0.816** | **0.732** | **0.772** | **Love** | |  | **0.856** | **0.006** | **0.855** | **0.856** | **0.856** | **God** | |  | **0.563** | **0.005** | **0.804** | **0.563** | **0.662** | **Life** | |  | **0.79** | **0.001** | **0.978** | **0.79** | **0.874** | **Intelligence** | |  | **0.849** | **0.002** | **0.947** | **0.849** | **0.895** | **Courage** | |  | **0.554** | **0.001** | **0.941** | **0.554** | **0.698** | **Friendship** | | **Weighted Avg.** | **0.842** | **0.007** | **0.847** | **0.842** | **0.84** |  | |
|  |
|  |