IDM: Assignment #1
News Classification Based On Their Headlines

Sovila Srun, Chanpiseth Chap
srun.sovila@rupp.edu.kh,chap.chanpiseth@rupp.edu.kh

Faculty of IT Engineering, Royal University of Phnom Penh

—

November 27, 2021

Introduction

Text mining is a hot topic in natural language processing (NLP), and it has been gaining significant interests for the last few years. The available text from the sources commonly come in numerous forms and unstructured. Many research literature proposed different techniques to transform this scattered text into well-defined structured format for one of different purposes i.e., text classification. Using full and long length texts for classification obviously provides better classification accuracy as compared to the short length text, however, the computing time is extraordinarily expensive. In this project, we want to explore three techniques to perform text classification using short texts [1] i.e., the headlines or titles of the text. The shortness of sentences may lead to the lack of context, which degrades the performance of the classifiers. Nevertheless, news titles may provide rich information of the semantic content in a concise way.

1 Objective

The objective of this assignment is to perform news classification based on their headlines using three different machine learning techniques i.e., Decision Tree, Multinomial Naive Bayes and Artificial Neural Network. Then outputs of the three models will be analyzed and presented elaborately.

2 Performance Metrics

The final stage is to evaluate how well the classifier models perform for classifying the category of news based on their headlines. Based on this evaluation metrics, we deeply understand how accurate our results are and how efficiently each news headline is classified into its pre-defined class. Many Researchers in literature have applied numerous measures for this purpose i.e. accuracy, precision/recall [2, 3], fallout [3], error, and much more. Few measures are briefly enlisted below.

- Precision is defined as a fraction of news headlines that is relevant.
- Recall is defined as fraction of relevant news headlines that is retrieved.
- Accuracy of a news headline is defined as the sum of true negative and true positive.
- True Positive means that news headline is classified to its correct class.
- False Negative means that news headline is classified to a wrong class.
- True Negative means that news headline does not belong to that class and is misclassified.

3 Requirements

- Python versions ^3.6
- Pandas ^0.25.5
- Sklearn ^0.23
- Numpy ^1.19.0
- Seaborn ^0.11.0 (Python data visualization library based on matplotlib)
- Jupyter Notebook to write and execute Python code

4 Dataset

Dataset [4] of references to news web pages collected from an online aggregator in the period from March 10 to August 10 of 2014. The resources are grouped into clusters that represent pages discussing the same news story. The dataset includes also references to web pages that point (has a link to) one of the news page in the collection.

4.1 Dataset and Attributes to be used in the assignment

For ease of use and reduction in file size, a new dataset **" newsCorpora_with_header.csv"** to be used in this assignment is extracted from the original dataset. This new CSV file contains only two attributes i.e. "TITLE" and "CATEGORY".

The attribute "CATEGORY" is the target class which we want our models to predict by using the independent attributes "TITLE".

4.2 Content

There are 422937 news pages. Each news headline has a corresponding category. Categories and the corresponding article counts are as follows:

- 152746 news of entertainment category

- 108465 news of science and technology category
- 115920 news of business category
- 45615 news of health category
- 2076 clusters of similar news for entertainment category
- 1789 clusters of similar news for science and technology category
- 2019 clusters of similar news for business category
- 1347 clusters of similar news for health category

4.3 The format of Dataset

The csv file uses UTF-8 encoding and Tab is used as delimiter.
FORMAT: ID \t TITLE \t URL \t PUBLISHER \t CATEGORY \t STORY \t HOSTNAME \t TIMESTAMP

- ID : Numeric ID
- TITLE : News title
- URL : Url
- PUBLISHER : Publisher name
- CATEGORY : News category (b = business, t = science and technology, e = entertainment, m = health)
- STORY : Alphanumeric ID of the cluster that includes news about the same story
- HOSTNAME : Url hostname
- TIMESTAMP : Approximate time the news was published, as the number of milliseconds since the epoch 00:00:00 GMT, January 1, 1970

5 Feature Extraction

When a huge number of features are given and each of the features is a well-known descriptive word for each class, and it may be possible that expected classification accuracy may not be achieved, and data may get over-trained. So, there is a need for feature extraction. The main objective of feature selection is to select a subset of input variables by removing features, which are irrelevant or of no predictive information. Feature selection has proven effective in enhancing learning efficiency, increasing predictive accuracy and reducing complexity of learned results. Feature selection has a main goal of finding a feature subset that produces higher classification accuracy.

Text data requires proper manipulation in order to be used as the input for the predictive modeling. So, the collection of text documents is converted to a matrix of *a vector of words or token counts* using *count vectorize* that produces a sparse representation of

the counts. Then, term frequency–inverse document frequency (TFIDF) is used to extract the most meaningful words in the corpus. TFIDF is the statistic that is intended to reflect how important a word is to a document in our corpus. This is generally used for feature selection but can also be used for removing stop words.

The token counts is an encoded vector containing a length of the entire vocabulary and an integer count for the number of times each word appeared in the document. To obtain the token counts, the CountVectorizer technique is employed to tokenize a collection of text documents and build a vocabulary of known words. This technique is also used to encode new documents using that vocabulary.

Equation for TF-IDF [5] is given below.

$$\text{TF-IDF}_{wk} = f_{wk} * \log(N/n_w) \tag{1}$$

Where

- f_{wk}: the frequency of word w in document k
- N: the number of documents in the collection
- n_w: the total number of times the word w occurs in the whole collection

What is tf-idf? Typically, the tf-idf weight is composed by two terms i.e. Term Frequency (TF) and the Inverse Document Frequency (IDF).

- TF stands for Term Frequency. It measures how frequently a term or word w occurs in a document. Since every document is different in length, it is possible that a term would appear much more times in long documents than shorter ones. Thus, the term frequency is often divided by the document length (aka. the total number of terms in the document) as a way of normalization.

$$f_{wk} = \frac{\textit{Number of times word } w \textit{ appears in a document } k}{\textit{Total number of words in the document } k} \tag{2}$$

- IDF stands for **Inverse Document Frequency**. It measures how important a term/-word is. While computing TF, all terms are considered equally important. However it is known that certain terms, such as "is", "of", and "that", may appear a lot of times but have little importance. Thus we need to weigh down the frequent terms/words while scale up the rare ones, by computing the following:

$$IDF_w = \log_e \frac{\textit{Total number of documents}}{\textit{Number of documents that contain the word } w} \tag{3}$$

6 Deadline

- 9/January/2021 before midnight.
- Submit to: chap.chanpiseth@rupp.edu.kh

References

[1] M. I. Rana, S. Khalid, and M. U. Akbar, "News classification based on their headlines: A review," in *17th IEEE International Multi Topic Conference 2014*. Karachi, Pakistan: IEEE, Dic. 2014, pp. 211–216. [Online]. Available: http://ieeexplore.ieee.org/document/7097339/

[2] I. Dilrukshi, K. De Zoysa, and A. Caldera, "Twitter news classification using SVM," in *2013 8th International Conference on Computer Science & Education*. Colombo, Sri Lanka: IEEE, Abr. 2013, pp. 287–291. [Online]. Available: https://ieeexplore.ieee.org/document/6553926

[3] G. Vinodhini and R. Chandrasekaran, "Sentiment Analysis and Opinion Mining: A Survey," vol. 2, num. 6, p. 11, 2012.

[4] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/News+Aggregator

[5] A. Krishnakumar, "Text categorization building a knn classifier for the reuters-21578 collection," *Department of Computer Science*, 2006.