Artificial Intelligence (CS571)

Department of CSE, IIT Patna

Assignment - 4: News Headlines classification using Naive Bayes classifier

(Read all the instructions carefully and adhere to them.)

Date : 30-November-2020 Deadline :- 05-November-2020

Instructions:

1. Markings will be based on the correctness and soundness of the outputs.
2. Marks will be deducted in case of plagiarism.
3. Proper indentation and appropriate comments (if necessary) are mandatory.
4. You should zip all the required files and name the zip file as:
roll_no_of_all_group_members .zip , eg. 1501cs11_1201cs03_1621cs05.zip.
5. Upload your assignment (the zip file) in the following link:
https://www.dropbox.com/request/hDmHExQKm0wvQHpRSP74

For any queries regarding this assignment contact:

Ramakrishna Appicharla (ramakrishnaappicharla@gmail.com)

---

1. **Problem statement:** Given the headline of news, the objective is to find the category of the news. (Note: Use only headline as input to find the category)
   For example:
   Short description: …
   Headline: Why Keeping a Food Journal Is Better Than Going on a Diet
   Date: …
   Link: …
   Authors: …
   Category: HEALTHY LIVING

   Consider the following categories only: Business, Comedy, Sports, Crime, Religion, Healthy Living, Politics

2. **Dataset: news_category_dataset.json**
3. **Classification Algorithm:**
   Naive Bayes

4. **Features:**

   Train the classifier using the following features.

   a. **Bag-of-words**
   b. **TF-IDF**
   c. **Create your own custom feature vectors.**

      For example, feature vector can contains following features:

      1. Current word (Unigram)
      2. POS tag of current word
      3. Position of the word
      4. Length of the news instance

   Here, a total of **3 models** needs to be trained i.e., one model using **Bag-of-words** features, one model using **TF-IDF** features and one model using **Custom feature vectors**.

   For more information on feature selection, refer the following paper:
   Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan*: Thumbs up? Sentiment Classification using Machine Learning Techniques*. In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002).

5. **Evaluation:**

   Perform 3-fold cross-validation for each model and report

   a. Overall precision, recall and F1-score
   b. Category-wise precision, recall and F1-score

Implementation notes:

1. You **can use** existing libraries to implement Naive Bayes algorithm and other tasks such as feature extraction, POS tagging, cross-validation and calculation of metrics.