AAM IPL Week 6
Naïve Bayes Email Spam Classifier

B.Tech – CSE(AIML)
V Semester - ML and AIUP, Aug-Oct 2024
Department of Computer Science Engineering – AI and ML (CSM)
G.Pulla Reddy Engineering College (Autonomous), Kurnool, AP

Algorithm of Application
Naïve Bayes

Project Title
Email Spam Classifier

Project Objective
The objective of this project is to build an email spam filter using the **Naive Bayes (NB) algorithm**, which classifies emails as either spam or genuine (ham).

The project will utilize the **SpamAssassin dataset**, a popular dataset for spam filtering tasks from Apache SpamAssasin project.

You will preprocess the dataset, vectorize the email content, train a Naive Bayes model, and evaluate its performance using standard classification metrics.

Dataset
- Description
  - The **SpamAssassin** dataset is a collection of emails that have been manually labeled as either **spam** or **genuine (ham)**. It is widely used for research and building email filtering systems.
  - The dataset contains thousands of spam and genuine emails spread across multiple sub-folders under spam and genuine folders underneath mail-dataset folder provided.
- Dataset Details:
  - Number of Emails: 10,745 (Spam and Genuine)
  - Spam to Genuine Ratio: Includes a balanced or imbalanced mixture of spam and genuine emails
  - File Format: Emails are stored as plain text files (.txt) in two main directories: spam and genuine
- Features:
  - The emails are represented as raw text. The features for the Naive Bayes model will be derived from the email text content:

C-501, Salarpuria Serenity, 5th Main, Sector 7, HSR Layout, Bengaluru 560102 KA India
Mobile: +91 97012 22130, Email: vmelachervu@gmail.com, vmela23@iitk.ac.in, Website: www.linkedin.com/in/vmelachervu

  - Email Headers: Contain metadata such as subject, sender, and date.
  - Email Body: The main content of the email, which may be plain text or HTML.
  - Email Subject: Can be an important indicator of spam emails (e.g., "Free," "Win," etc.).
- Usage in Machine Learning:
  - The dataset will be used to train a Naive Bayes classifier for spam detection.
  - Preprocessing steps will include cleaning the text, converting it into numerical format using TF-IDF vectorization, and removing HTML tags and special characters.
- Data Source and Published By:
  - Source: The SpamAssassin Public Corpus
  - Published By: SpamAssassin.org, a well-known open-source spam filtering project.
- Data Download Link
  - SpamAssasin public corpus

Implementation Steps

1. Import necessary libraries for data processing, model training, evaluation, and visualization.
2. Define the path to the watermark image for plot customization.
3. Load, preprocess, and clean email data from spam and genuine directories:
   - Load email files.
   - Preprocess and clean text data.
   - Return processed data as a DataFrame with labels.
4. Vectorize data using TF-IDF:
   - Transform text data into numerical features using the TF-IDF vectorizer.
   - Return the transformed feature matrix.
5. Train the Naive Bayes model:
   - Instantiate and train the Naive Bayes classifier on the training dataset.
6. Evaluate the trained model:
   - Use the test set to make predictions.
   - Print the accuracy, classification report, and confusion matrix for model performance.
7. Plot confusion matrix with watermark:
   - Generate the confusion matrix.
   - Add watermark to the plot and display.
8. Plot classification report with watermark:
   - Generate a heatmap of the classification report.
   - Add watermark to the plot and display.
9. Calculate and save word frequencies:
   - Calculate word frequencies across the dataset.
   - Save frequencies in a text file (word_frequency.txt) for further analysis or reference.
10. Main function:
    - Define paths for spam and genuine email directories.
    - Prepare data by loading and preprocessing.
    - Split data into training and testing sets.
    - Vectorize the data.
    - Train the Naive Bayes model.
    - Evaluate and plot the results with watermarks.
    - Calculate and save word frequencies.
11. Run the main function if the script is executed directly.

C-501, Salarpuria Serenity, 5th Main, Sector 7, HSR Layout, Bengaluru 560102 KA India
Mobile: +91 97012 22130, Email: vmelachervu@gmail.com, vmela23@iitk.ac.in, Website: www.linkedin.com/in/vmelachervu

Project Files Provided

- Project shell code file - ***AAM-IPL-Wk-6-Naive-Bayes-Email-Spam-Classifier-Shell-Code.ipynb***
- Training Data – ***mail-dataset.7z***
- Watermark image for plots - ***AAM-IPL-Watermark-for-Plots.png***

Project Overview, Implementation and Submission Timeline

26-10-2024 – Saturday 10:30 AM – Project Announcement – Topic, Data Set, Shell Code etc. Announcement Channels – Google Class, Industry Projects WhatsApp Group.			
10	26-10-2024 - Saturday Duration: 1.5 Hrs	Naïve Bayes – Model Overview/Recap, Project Description, and Interactive Q&A	Online – Google Class
11	27-10-2024 - Sunday Duration: 1.5 Hrs	Naïve Bayes – Model Building, Output Demonstration, Q&A	Online – Google Class
31-10-2024 – Thursday 11:59 PM - Deadline to upload the project code submission by all students in Google Class.			

Guest Lecture Timings:
Saturdays: 10:30 AM IST – 12:00 Noon IST
Sundays: 10:30 AM IST – 12:00 Noon IST
Mondays: 6:30 PM IST – 8:00 PM IST

Development Environment

- Computing Language – Python
- IDE – Visual Studio Code with Jupyter Notebook

Instructor

Instructor	
Venkateswar Reddy Melachervu (alumnus of GPREC, ECE Class of '92) CTO, Brillium Technologies, Bengaluru Email: **venkat.reddy.gf@gprec.ac.in** Profile: LinkedIn	Visiting Faculty

Coordination

All the activities of this programme – lecture venues, weekly projects details announcements, general announcements, changes in lecture timings, etc. will be coordinated by CSM faculty member Sri V.Suresh.

Channels of Communication and Announcements

- Google Classroom
- Whatsapp group **- Applied AI & ML Industry Projects Lab**
- Emails (Strictly GPREC email addresses only)

Programme Coordinator	
Prof. V.Suresh	Faculty Member, CSM

© "...dare to dream; care to win..."

Email: **vsuresh.ecs@gprec.ac.in**	

Reference Books

- Pattern Recognition and Machine Learning by Chris Bishop, 2006 – PDF Link
- Machine Learning using Python by Manaranjan Pradhan and U Dinesh Kumar, Wiley 2019 – PDF Link

Policies

- **Attendance:** All sessions are expected to be attended by all the enrolled students. In case of inability to attend, prior information is expected to be provided by the student to the coordinator with a copy to the visiting faculty
- **Project Submissions:** Duly completed projects (Jupyter NB file and a PDF of the Jupyter NB file) are expected to be submitted through google class prior to the deadline. In case of inability to complete due various unforeseen circumstance, students are expected to seek extension for the submission deadline.
- **Academic Integrity:** Students are expected to uphold the highest standards of academic integrity in all assignments for the Applied AI & ML Industry Projects Lab. Each assignment must be the student's own work, and all sources and collaborators must be properly acknowledged. By submitting their completed project source code, students confirm that they have adhered to this integrity policy and completed their work in an honest and ethical manner.

Additional Information

For students interested in engaging with special projects in the field of Gen AI, please reach out to the visiting faculty at **venkat@brillium.in** for further details and opportunities.

Contact Information

For any questions or concerns or further details on this programme, please contact **Program Coordinator** during office hours or via email.

---------- End of the Document----------

"...dare to dream; care to win..."