Resit Assignment

Due on February 5, 2024 (23:59:59)

Instructions. The goal of this problem set is to make you understand and familiarize with Naive Bayes algorithm and K-Mean clustering.

Book Genre Classification with Naive Bayes and Clustering

In this part of the assignment, For this assignment, you will implement Naive Bayes classifiers and K-Means clustering to classify the examples on the Book Genre Dataset (Figure 1) mentioned below.

Title	**Genre**	**Summary**
The Elves of Cintra	fantasy	Beginning where Armageddon's Children ended, Knight of the Word Logan Tom races to save the gypsy morph Hawk...
Long Bright River	thriller	Two sisters travel the same streets, though their lives couldn't be more different. Then, one of them goes missing...
American Psycho	horror	Set in Manhattan during the Wall Street boom of the late 1980s, American Psycho is about the daily life of wealthy young investment banker Patrick Bateman...
The Stone Carvers	history	In the mid-19th century, Father Gstir is sent from Bavaria to Canada to minister to German-Catholic communities...
The Man Who Fell to Earth	science	Thomas Jerome Newton is a humanoid alien who comes to Earth seeking to construct a spaceship to ferry others from his home planet, Anthea, to Earth...
Under a Monsoon Cloud	crime	Inspector Ghote is temporarily assigned to a badly run hill station at Vigatpour. Additional Deputy Inspector General ”Tiger” Kelkar...

Figure 1: Some examples from different genres in the dataset

Dataset

- You can download the dataset from given link.
- Dataset consists of 3000 samples with 6 discrete (”genre” attribute) ground-truth class types.
- Book Genre Dataset is a dataset provided to determine the genre of a book from the summary of the book. It includes the following features:
  - **Title:** Title of the book.
  - **Genre:** Ground-truth genre of the book (crime, thriller, fantasy, horror, history, science)
  - **Summary:** Summary of the book.
- You should divide your data as 80% training and 20% testing.

Approach

Feature Extraction

You will represent your data with features and use them with the Naive Bayes and K-Means algorithm.

You will use the Bag of Words (BoW) model which learns a vocabulary from all of the documents, and then models each document by counting the number of times each word appears. You will use BoW with two options:

- Unigram: The occurrences of words in a document(frequency of the word).
- Bigram: The occurrences of two adjacent words in a document.

You will use the Term Frequency-Inverse Document Frequency (TF-IDF) model to represent the words in your dataset.

Also, you will remove stopwords from the dataset. Stopwords is the common words like ’a’, ’to’, and ’the’ in English.

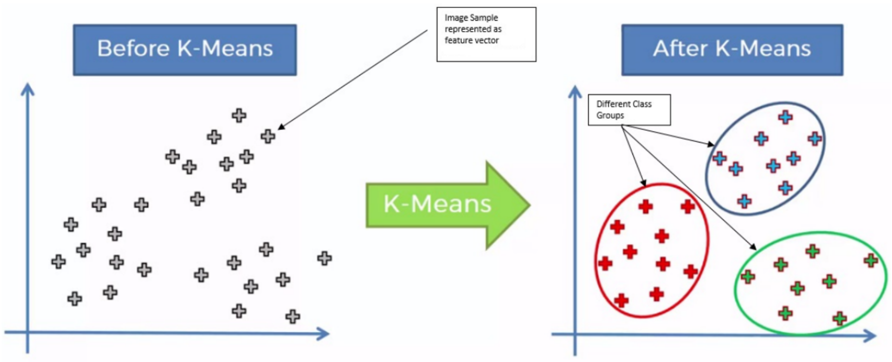
For implementing BoW and TF-IDF models you can use Scikit Learn machine learning library.

Naive Bayes

For this assignment you will implement your own Multinomial Naive Bayes for classifying genres of the book summaries. For implementing Naive Bayes you can use NumPy and other mathematical libraries.

K-Means Clustering

For this assignment you will implement your own K-Means Clustering method for classification. For implementing Naive Bayes you can use NumPy and other mathematical libraries.



Classification Performance Metric

You will compute "Accuracy", "Precision" and "Recall" of your model to measure the success of your classification and clustering method:

$$\textbf{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \tag{1}$$

$$\textbf{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\textbf{Recall} = \frac{TP}{TP + FN} \tag{3}$$

You will report accuracy, precision and recall.

Analyze results

In this part, you need to analyze your results. Before analyzing your results make sure that getting results of the given experiments are in Figure 2.

1. Compare feature extraction methods. Explain which feature extraction method will be useful for better classification. Also, you may compare the training time of each feature extraction method.

2. Explain whether removing the stopwords helps the classification or not.

3. Compare Naive Bayes and K-Mean clustering. Explain which method will be useful for better classification of book genre.

#	Feature	Stopwords	Algorithm	Accuracy	Precision	Recall
1	BoW (Unigram)		NB			
2	BoW (Unigram)	Removed	NB			
3	BoW (Bigram)		NB			
4	BoW (Bigram)	Removed	NB			
5	TF-IDF		NB			
6	TF-IDF	Removed	NB			
7	BoW (Unigram)		K-Means			
8	BoW (Unigram)	Removed	K-Means			
9	BoW (Bigram)		K-Means			
10	BoW (Bigram)	Removed	K-Means			
11	TF-IDF		K-Means			
12	TF-IDF	Removed	K-Means			

Figure 2: Expected experiments for assignment.

Implementation Details

- **You can't use ready-made libraries for your Naive Bayes classification and K-Means clustering methods implementations. You must implement these on your own.**
- **You can't use ready-made libraries for computing "Accuracy", "Precision" and "Recall" metrics. You must implement these on your own.**
- You may use Numpy array functions for your intermediate implementation steps for your implementations.
- You may use "Pandas" library for reading and writing/creating .csv files.

Submit

You are required to submit all your code in a Jupyter notebook, along with a report in ipynb format, which should also be prepared using Jupyter notebook. The code you submit should be thoroughly commented. Your report should be self-contained and include a concise overview of the problem and the details of your implemented solution. Feel free to include pseudocode or figures to highlight or clarify specific aspects of your solution. Finally, prepare a ZIP file named name-surname-resit.zip containing:

- name_surname_resit.ipynb (including your report and code)
- name_surname_resit.py (py file version of your ipynb file)
- Do not send the dataset.

The ZIP file will be submitted via Google Classroom. Click here to accept your Resit.

Grading

- Code (60): NB: 20 points, k-Means: 15 points, BoW: 10 points, TF-IDF: 5 points, Stopwords: 5 points, Performance metric: 5 points
- Report (40): Analysis of the results for prediction: 40 points.

  Notes for the report: Preparing a good report is important as well as your solutions!

Academic Integrity

All work on assignments must be done individually unless stated otherwise. You are encouraged to discuss with your classmates about the given assignments, but these discussions should be carried out in an abstract way. That is, discussions related to a particular solution to a specific problem (either in actual code or in the pseudocode) will not be tolerated. In short, turning in someone else's work, in whole or in part, as your own will be considered as a violation of academic integrity. Please note that the former condition also holds for the material found on the web as everything on the web has been written by someone else.