Matthew Muranaka

Objectives:

1. Implement and evaluate a Naïve Bayes classifier algorithm.

Task:

My task is to implement, train, and test a Naïve Bayes classifier using a publicly available data set.

Data set:

A publicly available data set first and do an initial exploratory data analysis.

Deliverables:

- Python code file(s), named:

```
classifier.py
```

- Presentation slides in PPTX format, named:

```
Hotel_Review_Classifier_Presentation.pptx
```

- This document with observations and conclusions, named:

```
Hotel_Review_Classifier_WriteUp.pdf
```

Implementation:

My task is to implement (**from scratch without an out-of-the-box Python package classifier**), train, and test a Naïve Bayes classifier (as outlined in class) and apply it to classify sentences entered using user-input.

The program should:

- Accept one (1) command line argument, i.e. so your code could be executed with

```
python classifier.py TRAIN_SIZE
```

  where:
  - `classifier.py` is the python code file name,
  - `TRAIN_SIZE` is a number between 20 and 80 defining the size (in percentages) of the training set. For example: `60` would mean **FIRST** (as ordered in the dataset file) 60% of samples. **Note that your test set is always going to be the LAST (as ordered in the dataset file) 20% of samples.**

  Example

```
python classifier.py YES
```

If the number of arguments provided is NOT one (none, two or more) or the `TRAIN_SIZE` argument is out of the specified range, assume that the value for `TRAIN_SIZE` is `80`.

- Load and process input data set:
  - Apply any data clean-up / wrangling considered necessary first (mention and discuss choices in the Conclusions section below).
  - Text pre-processing:
    - treat every document in the data set as a single sentence, even if it is made of many (no segmentation needed),

- Train your classifier on the data set:
  - assume that vocabulary V is the set of ALL words in the data set,
  - divide the data set into:
    - training set: FIRST (as they appear in the data set) `TRAIN_SIZE` % of samples / documents,
    - test set: LAST 20 % of samples / documents,
  - use **binary** BAG OF WORDS with **"add-1" smoothing** representation for documents,
  - train the classifier,

- Test your classifier:
  - use the test set to test the classifier,
  - calculate (and display on screen) following metrics:
    - number of true positives,
    - number of true negatives,
    - number of false positives,
    - number of false negatives,
    - sensitivity (recall),
    - specificity,
    - precision,
    - negative predictive value,
    - accuracy,
    - F-score,

- Ask the user for keyboard input (a single sentence S):
  - use my Naïve Bayes classifier to decide which class S belongs to,

- display classifier decision along with P(CLASS_A |S) and P(CLASS_B | S) values on screen

Program output should look like this (if pre-processing step is NOT ignored, output `NONE`):

```
Training set size: 80 %

Training classifier…
Testing classifier…
Test results / metrics:

Number of true positives: xxxx
Number of true negatives: xxxx
Number of false positives: xxxx
Number of false negatives: xxxx
Sensitivity (recall): xxxx
Specificity: xxxx
Precision: xxxx
Negative predictive value: xxxx
Accuracy: xxxx
F-score: xxxx

Enter your sentence:

Sentence S:

<entered sentence here>

was classified as <CLASS_LABEL here>.
P(<CLASS_A> | S) = xxxx
P(<CLASS_B> | S) = xxxx

Do you want to enter another sentence [Y/N]?
```

If the user responds `Y`, classify a new sentence (without retraining the classifier).

where:

- `80` would be replaced by the value specified by `TRAIN_SIZE`,
- `xxxx` is an actual numerical result,
- `<entered sentence here>` is actual sentence entered by the user,
- `<CLASS_LABEL here>` is the class label decided by the classifier,
- `<CLASS_A>`, `<CLASS_B>` are available labels (SPAM/HAM, POSITIVE/NEGATIVE, etc.).

Classifier testing results:

Enter your classifier performance metrics below:
Size of Test Set: 4099

With `TRAIN_SIZE` set to 80:	With `TRAIN_SIZE` set to 20 (not 80)

`Number of true positives: 3380`
`Number of true negatives: 435`
`Number of false positives: 91`
`Number of false negatives: 193`
`Sensitivity (recall): 0.973`
`Specificity: 0.692`
`Precision: 0.945`
`Negative predictive value: 0.826`
`Accuracy: 0.930`
`F-score: 0.959`
	`Number of true positives: 3423`
	`Number of true negatives: 381`
	`Number of false positives: 145`
	`Number of false negatives: 150`
	`Sensitivity (recall): 0.959`
	`Specificity: 0.717`
	`Precision: 0.958`
	`Negative predictive value: 0.724`
	`Accuracy: 0.928`
	`F-score: 0.958`

What are your observations and conclusions? When did the algorithm perform better? a summary below

Summary / observations / conclusions

One takeaway from the dataset is the rarity of negative hotel reviews leading to a low chance for false positives. Overall, I thought my classifier performed exceptionally well. While it had quite a problem with falsely identifying negative review, it performed in every other category. I think, as a way to fix this, we could keep a threshold towards identifying positive reviews.

I am a bit surprised removing the stop words led to quite worse performance but after thinking about it, it kind of makes sense. My hypothesis is that removing the stop words evened the probability of positive/negative reviews which should not happen because positive reviews are much more apparent.

I think a big challenge I ran into was handling certain characters/strings in the document. I talked about this in the “Data Set Information” slide but some reviews weren’t in English or had errors in the translation which led to the appearance of unknown characters. In order to handle this, I just handled these words as an “other” category which often had only one appearance of that word. I think a way to fix this would be to further processing of these unknown words or totally dismissing them in the classifier.

Other than the ways I have already mentioned to improve the classifier, I think an implementation that I could add with my current understanding is to recognize negation words and include those as a 2-ary word.

The reviews processed from Tripadvisor to Kaggle had errors in processing which led to occasional unknown characters or characters that don’t make sense in the context. In order to account for these inconsistencies, I completely eliminated certain characters, replacing them with a space.