Classification of SMS messages into spam and non-spam using a naive Bayesian algorithm

The goal of the project
is to develop a model that automatically detects whether a text message is spam.

SETTING THE TASK

Project objective:

The text of an SMS message is sent to the input
At the output, the model should determine:

spam
ham (not spam)

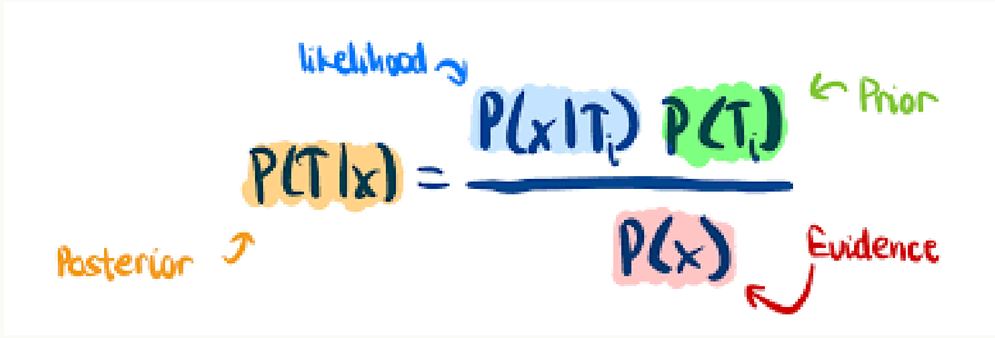
Why this is important:
Spam filtering is used in mail, messengers, banks
Text classification is a basic NLP task

TECHNOLOGIES USED

Programming language:
Python

Libraries:
pandas — loading data
re — clearing text
numpy — auxiliary operations

Algorithm:
Naive Bayes



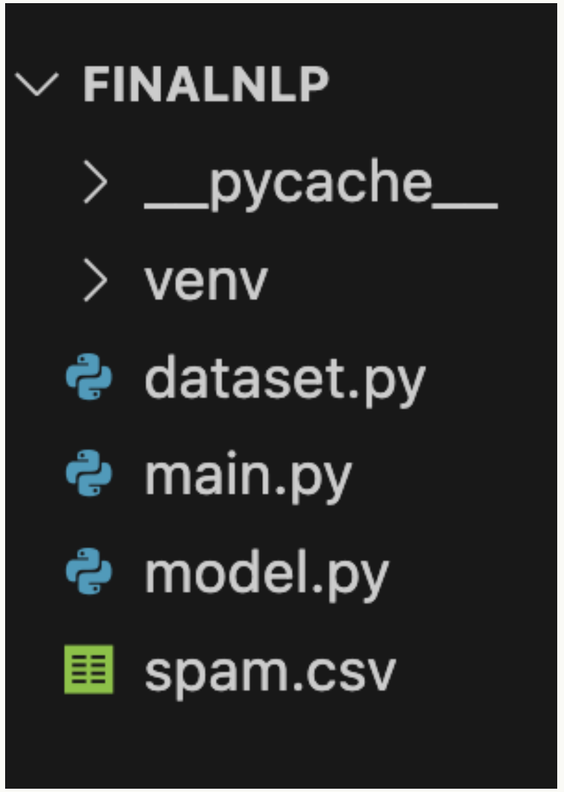
PROJECT STRUCTURE
THE PROJECT CONSISTS OF THREE FILES:

DATASET.PY

MODEL.PY

MAIN.PY

-data preparation
- training and model logic
- launch the project and check the result



DATASET — DATA PREPARATION

File: dataset.py
Class: Dataset

The main tasks of the class:

- Clearing the text
- Converting labels to numbers
- Splitting data into train / validation / test

Text cleaning (preprocessing)
--In the _transform method() is in progress:

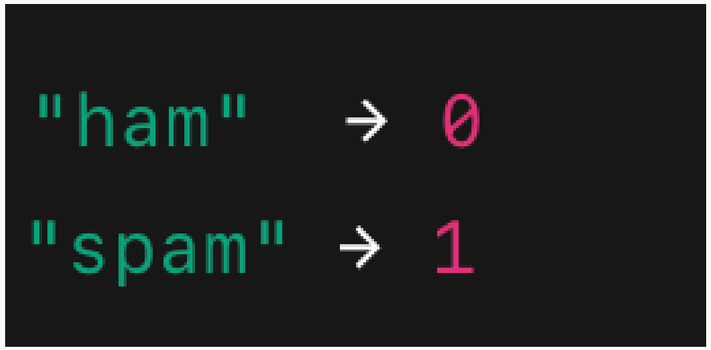
```
text = text.lower()
text = re.sub(r'[^a-z]', ' ', text)
text = ' '.join(text.split())
```

Why it is necessary:
to bring the text to a uniform format
, remove punctuation
marks, avoid duplicate words (Free and free)

This improves the quality of the model

LABEL CONVERSION

WHY IS IT SO:
MODELS ARE MORE CONVENIENT TO WORK
WITH NUMBERS
AND SIMPLIFY PROBABILITY CALCULATIONS.



DATA SPLITTING (SPLIT_DATASET)

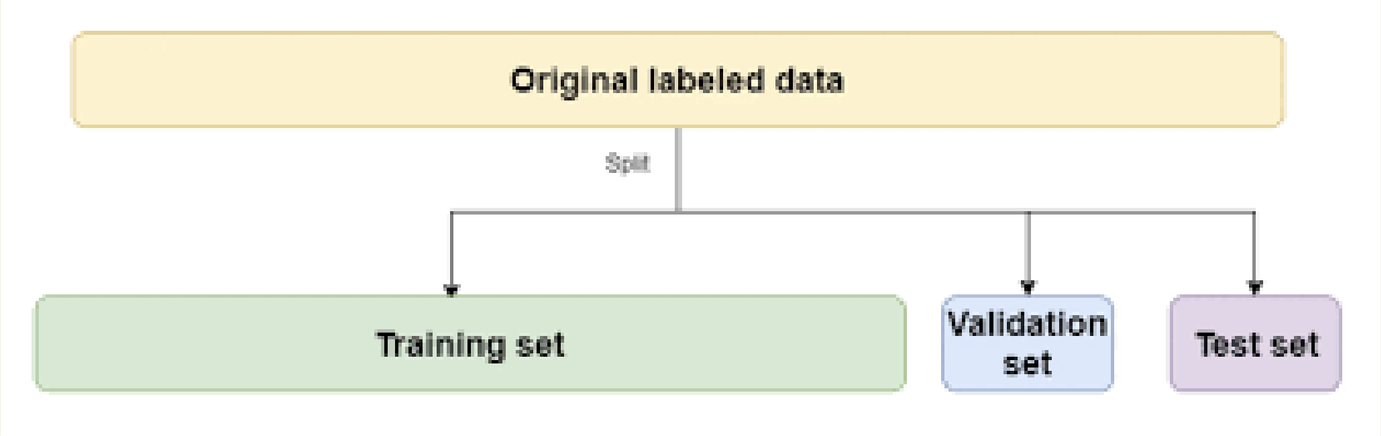
The data is divided in order, not randomly:

70% — TRAIN

15% — VALIDATION

15% — TEST

```
X_train = self._x[:n_train]
X_val   = self._x[n_train:n_train+n_val]
X_test  = self._x[n_train+n_val:]
```



Important:
the data is not mixed
deterministic partitioning
is used for a training project.

MODEL — THE GENERAL IDEA

The file: model.py
Class: Model
Model: learns on train, remembers word frequencies, uses a Naive Bayesi algorithm, makes predictions for new messages

Why naive Bayes is used
Naive Bayes is well suited for text because:
It works with word frequencies
and is effective for spam filtering
and does not require complex calculations.
The basic idea:
The probability of a class depends on the words in the message

What the model remembers during training
During fit(), the model saves:
the dictionary of all words (vocab)
the frequency of words in spam (self.spam)
the frequency of words in ham (self.ham)
total number of words
dictionary size

```
self.Nvoc = len(self.vocab)
self.Nspam = spam_words_count
self.Nham = ham_words_count
```

WHERE EXACTLY IS BAYES USED

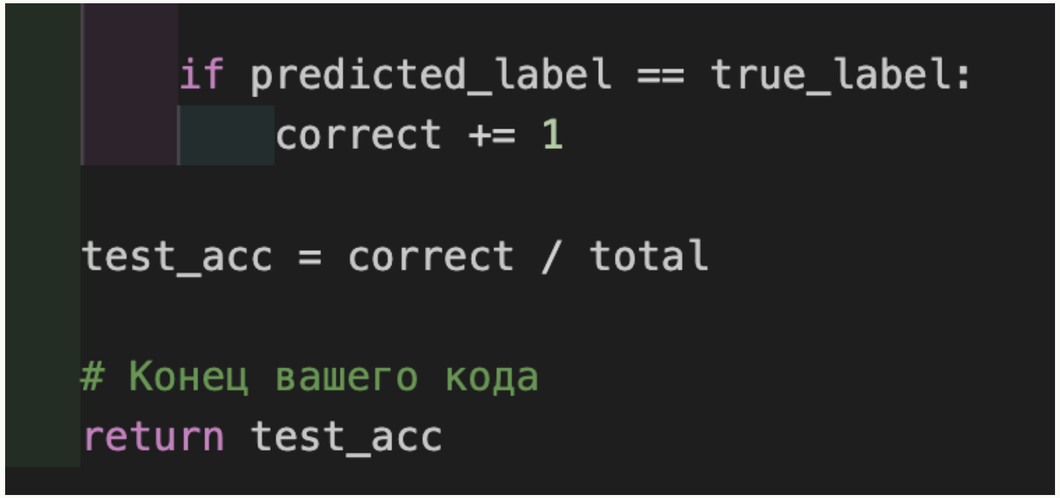
In the inference() method:
A priori probabilities are considered:
P(spam), P(ham)
It counts for each word:
P(word | spam)
P(word | ham)
All probabilities are multiplied
The class is more likely to be selected

```
if pspam > pham:
    return "spam"
return "ham"
```

Why do I need the alpha
alpha = 1 parameter?
This is Laplace smoothing
Solves the problem:
if the word did not occur in the training
, the probability does not become equal to 0
Without alpha:
, the model is breaking down

MODEL QUALITY CONTROL

MODEL QUALITY CONTROL
TWO METHODS ARE USED:
VALIDATION()
TEST()
THEY:
CALL INFERENCE()
COMPARE PREDICTIONS WITH
CORRECT LABELS
AND CALCULATE ACCURACY



```
    if predicted_label == true_label:
        correct += 1

test_acc = correct / total

# Конец вашего кода
return test_acc
```

RESULTS

OUTPUT:

- ACCURACY ON VALIDATION
- ACCURACY ON TEST
- PREDICTIONS FOR NEW MESSAGES

```
Validation accuracy: 0.9832 (98.32%)
Test accuracy: 0.9880 (98.80%)
Message: Go until jurong point, crazy.. Available only in b... -> Prediction: ham
Message: Hi John, are we still meeting for lunch tomorrow?... -> Prediction: ham
Message: FREE entry into our £100 weekly draw. Just text WI... -> Prediction: spam
Message: Hey, don't forget to bring the documents for the m... -> Prediction: ham

Accuracy > 95% on both validation and test sets!
```

CONCLUSIONS

- ✓ IMPLEMENTED A NAIVE BAYESIAN CLASSIFIER
- ✓ TEXT CLEANING HAS BEEN USED
- ✓ THE DATA IS DIVIDED INTO TRAIN / VALIDATION / TEST
- ✓ THE MODEL SHOWS HIGH ACCURACY