Selle ülesanne eesmärgiks oli realiseerida Naiivset Bayesi klassifikaatorit selleks, et välja selgitada, kas antud kiri on spam või pole.

Programm koosneb 2-st osast. Esimene osa on treening. Me parsime andmeid tekstifailidest ja leiame väärtused järgmiseks valemiks:

$$P(w|c) = \frac{N_{w,c} + 1}{N_c + |V|}$$

Joonis 1 Valem 1

Kus $N_{w,c}$ sõna esinemise sagedus ham-is või spam-is (meetod count_word_frequency()). N_c näitab kui palju sõnu ham-is või spam-is (meetod count_words()). V on ham ja spam sõnad ilma kordusteta (meetod get_unique_words()).

Teine programmi osa on kontrollimine, kas kiri on spam või pole. Kirjad näidest asuvad direktooriumides test/message1 ja test/message2. Me parsime neid, viskame ära duplikaate ja sõnu, mis ei osalenud treeningus. Me arvutame hami ja spami tõenäosust Valemi 1 abil FOR tsüklis (kus on kommentaar #count ham for message 1 jne main() meetodis). Enne seda me arvutame ka P(ham) ja P(spam) meetodis count_probabilities().

Tulemused

```
Message #1, HAM: -676.7008365761612
Message #1, SPAM: -705.0415555424338
Message #2, HAM: -1052.0471277625898
Message #2, SPAM: -984.0077680194872
```

Esimene kiri ei ole spam, teine (kus on Nigeeria bank) on spam.