Ευάγγελος Κλειδόπουλος 3190084

Δεύτερο project Τεχνητής Νοημοσύνης

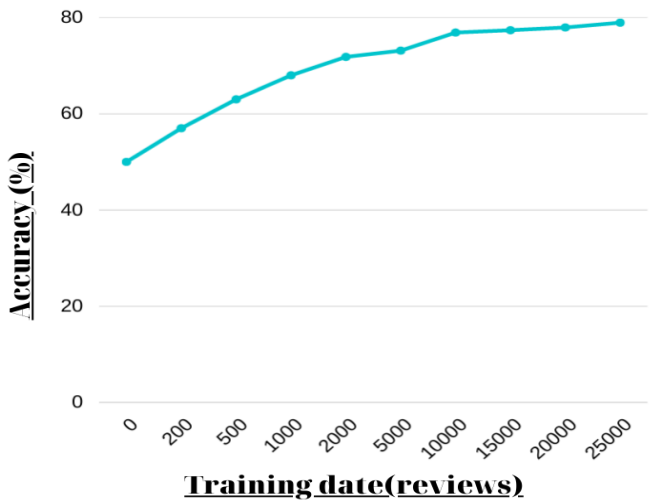
A. Υλοποίησα τον αλγόριθμο multinomial naive bayes χρησιμοποιώντας και information gain.

Χρησιμοποιώντας όλα τα training reviews και όλα τα test reviews, ο αλγόριθμος επιτυγχάνει ποσοστό 78.55%. Αφαιρώ από το λεξιλόγιο μου m=115782 λέξεις, k = 254 λέξεις και κρατάω n = 1922 λέξεις στο λεξιλόγιο. Σε αυτά τα νούμερα κατέληξα αφαιρώντας τις λέξεις που είχαν συχνότητα μικρότερη από 86 φορές και μεγαλύτερη από 1455 φορές.

```
Training started.
Most frequent removed: 254. Least frequent removed: 115782.
Words remaining in vocabulary: 1922.
Training finished, starting testing.
Total: 25000 total correct: 19637 total correct percentage: 78.548%.
Positives percentage: 81.264%.
Negatives percentage: 75.832%.
```

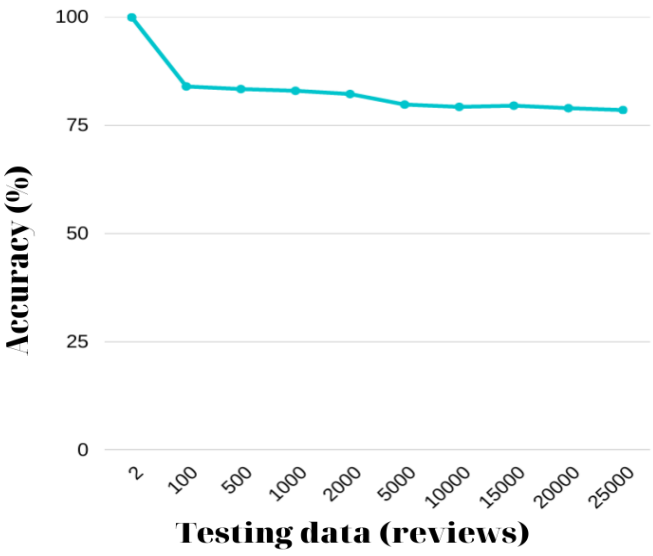
Training table και graph:

Training data (reviews)	Accuracy (%)
0	50
200	57
500	63
1000	68
2000	71.816
5000	73.12
10000	76.884
15000	77.37
20000	77.936
25000	78.548



Testing table και graph:

Testing data (reviews)	Accuracy (%)
2	100
100	84
500	83.4
1000	83
2000	82.25
5000	79.82
10000	79.28
15000	79.56
20000	78.985
25000	78.548

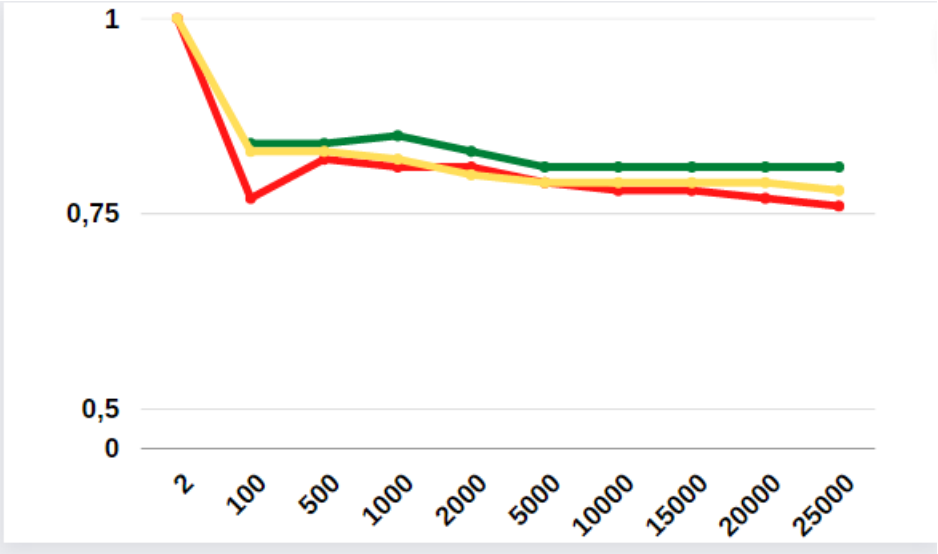


Precision, recall, F1 table and graph για τα testing data:

Χρησιμοποίησα τους τύπους:

- Precision = True Positives / (True Positives + False Positives)
- Recall = True Positives / (True Positives + False Negatives)
- F1 = 2 * Precision * Recall / (Precision + Recall)

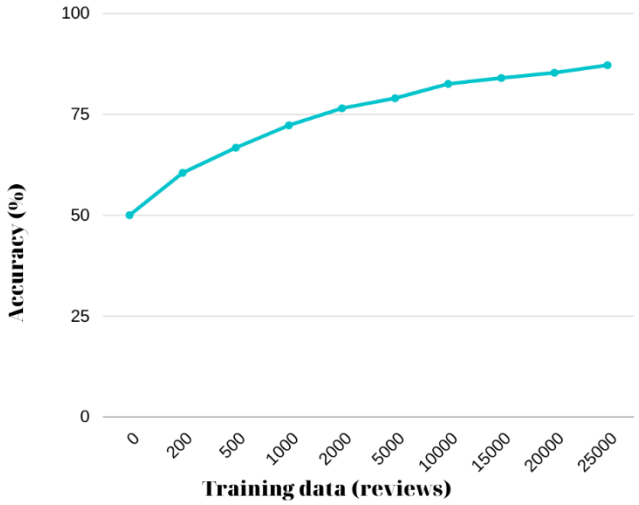
Testing data (reviews)	Precision	Recall	F1
2	1	1	1
100	0.92	0.94	0.93
500	0.92	0.92	0.92
1000	0.90	0.92	0.91
2000	0.89	0.91	0.90
5000	0.89	0.90	0.89
10000	0.88	0.90	0.88
15000	0.87	0.89	0.88
20000	0.87	0.88	0.87
25000	0.86	0.88	0.87



Β. Για το δεύτερο κομμάτι της εργασίας χρησιμοποίησα τον αντίστοιχο αλγόριθμο της Weka (weka.classifiers.bayes.NaiveBayesMultinomial).

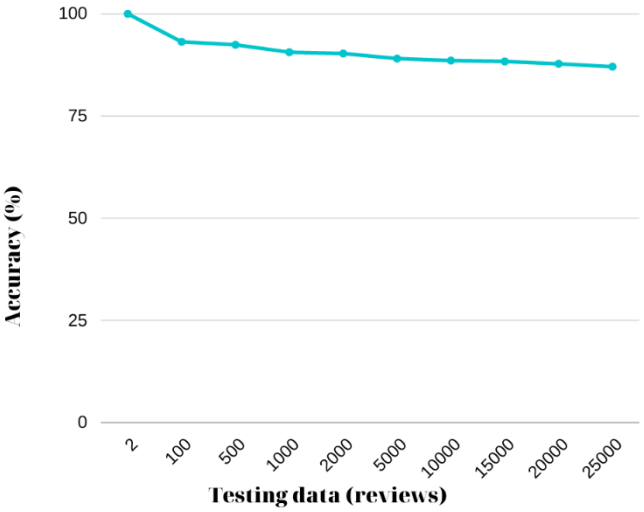
Training table και graph:

Training data (reviews)	Accuracy (%)
0	50
200	60.5
500	66.723
1000	72.28
2000	76.5
5000	79
10000	82.56
15000	84.02
20000	85.32
25000	87.02



Testing table και graph

Testing data (reviews)	Accuracy (%)
2	100
100	93.12
500	92.4
1000	90.57
2000	90.25
5000	89.01
10000	88.52
15000	88.3
20000	87.72
25000	87.02



Precision, recall, F1 table and graph για τα testing data:

Χρησιμοποίησα τους τύπους:

- Precision = True Positives / (True Positives + False Positives)
- Recall = True Positives / (True Positives + False Negatives)
- F1 = 2 * Precision * Recall / (Precision + Recall)

Testing data (reviews)	Precision	Recall	F1
2	1	1	1
100	0.77	0.84	0.83
500	0.82	0.84	0.83
1000	0.81	0.85	0.82
2000	0.81	0.83	0.8
5000	0.79	0.81	0.79
10000	0.78	0.81	0.79
15000	0.78	0.81	0.79
20000	0.77	0.81	0.79
25000	0.76	0.81	0.78