Ονόματα συμμετεχόντων:

ΒΑΣΙΛΕΙΟΣ-ΕΚΤΩΡ ΚΩΤΣΗΣ-ΠΑΝΑΚΑΚΗΣ: 3180094

ΑΛΕΞΙΟΣ - ΛΑΖΑΡΟΣ ΓΕΩΡΓΙΟΥ: 3180027

ΔΗΜΗΤΡΙΟΣ ΒΛΑΝΤΗΣ: 3180021

Γενικά στοιχεία εργασίας

Η γλώσσα που επιλέχθηκε για να γραφτεί ο κώδικας είναι η python. Πιο συγκεκριμένα δεν χρησιμοποιήθηκε καμία από τις γνώστες βιβλιοθήκες για ανάπτυξη προγραμμάτων machine learning, αντ΄αυτού ο απαραίτητος κώδικας δημιουργήθηκε από εμάς.

Εφόσον τα άτομα της ομάδας είναι 3, υλοποιήθηκαν 3 αλγόριθμοι. Πιο συγκεκριμένα υλοποιήθηκαν οι:

- Αφελής ταξινομητής Bayes
- ID3
- Logistic Regression (στοχαστική υλοποίηση)

Χρησιμοποιήθηκε δικό μας custom λεξικό το οποίο δημιουργήθηκε από τον γενικό κώδικα που παραθέτεται στο readdata.py.

Γενικά όλοι οι κώδικες πρέπει να τοποθετηθούν μέσα στο “/aclImdb/” directory.

Μπορούμε να αλλάξουμε τις υπερμεταβλητές N, M, DATA_COUNT στην αρχή του κάθε κώδικα και τα K, h μέσα στις αντίστοιχες συναρτήσεις.

Αφελής ταξινομητής Bayes

Η υλοποίηση του αλγορίθμου βασίστηκε αποκλειστικά στις διαφάνειες του μαθήματος.

Λεπτομέρειες για τον κώδικα καλύπτονται με τη μορφή σχολίων.

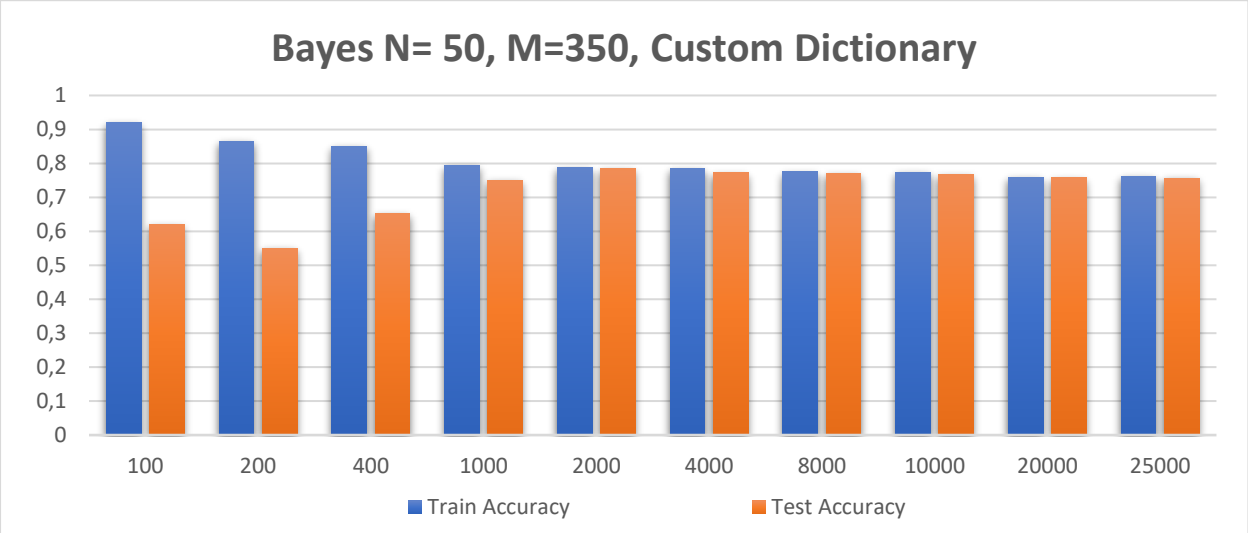
Οπότε παρακάτω παραδίδεται πίνακας ορθότητας των test data και train data σε συνάρτηση με το πλήθος των train data . Οι προκαθορισμένες τιμές των παραμέτρων είναι

- N = 50, M = 350 (πρακτικά αυτό σημαίνει ότι σαν ιδιότητες χρησιμοποιήθηκαν 250 μεταβλητές, από την λέξη στη θέση 51 έως αυτή στη θέση 350)

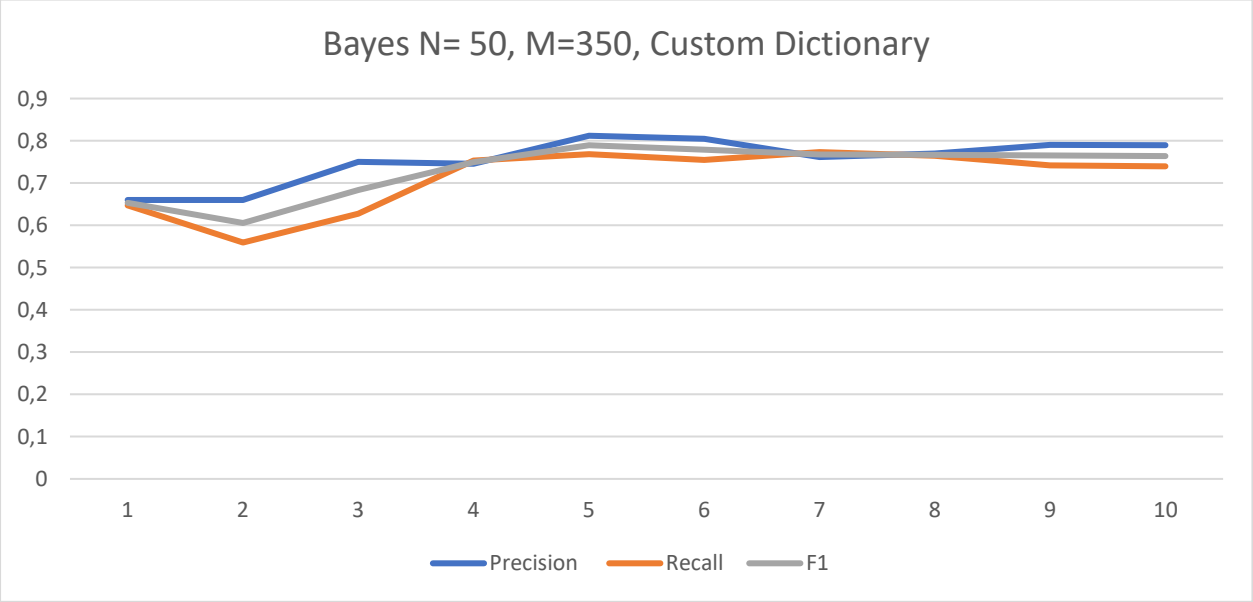
	Bayes N= 50, M=350, Custom Dictionary			Test	Test	Test	
DATACOUNT	datacount	AccuracyTrain	AccuracyTest	Precision	Recall	F1	datacount
50	100	0,92	0,62	0,66	0,647058824	0,653465347	100
100	200	0,865	0,55	0,66	0,559322034	0,605504587	200
200	400	0,85	0,6525	0,75	0,627615063	0,683371298	400
500	1000	0,793	0,75	0,746	0,753535354	0,749748744	1000
1000	2000	0,788	0,7835	0,812	0,768211921	0,789499271	2000
2000	4000	0,7845	0,77175	0,8045	0,75504458	0,778988138	4000
4000	8000	0,775625	0,769125	0,7615	0,773292714	0,767351052	8000
5000	10000	0,7743	0,7664	0,7698	0,764600715	0,767191549	10000
10000	20000	0,7593	0,75775			0,76545481	20000
12500	25000	0,76224	0,756			0,764004952	25000

Κύριες γραμμές πλέγματοςΚατακόρυφα (τιμών) άξονας

Παρακάτω παραθέτεται η καμπύλη ορθότητας:



Παρακάτω παραθέτεται οι καμπύλες ακριβείας, recall και f1 για τα test data:



Logistic Regression

Για την υλοποίηση του κώδικα προτιμήθηκε η στοχαστική υλοποίηση λόγω της πολυπλοκότητας του αρχικού κώδικα που υπολογίζει την πιθανοφάνεια σε όλα τα train data.

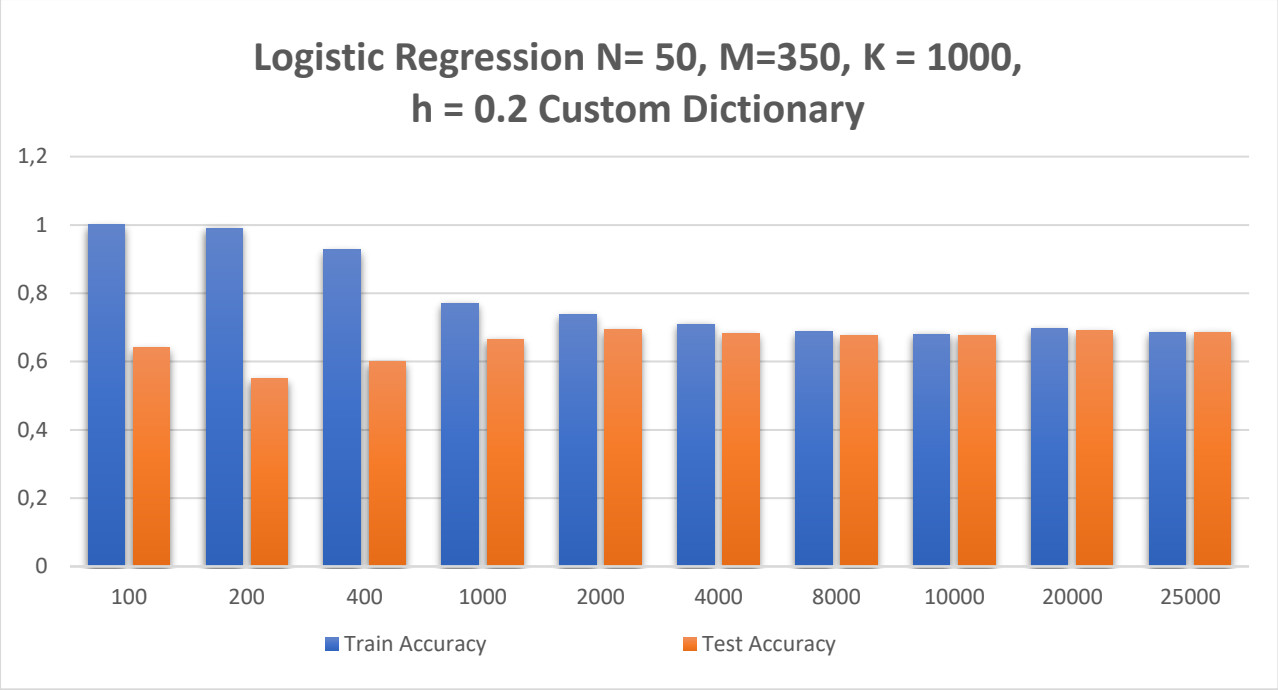
Λεπτομέρειες για τον κώδικα καλύπτονται με τη μορφή σχολίων.

Οπότε παρακάτω παραδίδεται πίνακας ορθότητας των test data και train data σε συνάρτηση με το πλήθος των train data . Οι προκαθορισμένες τιμές των παραμέτρων είναι

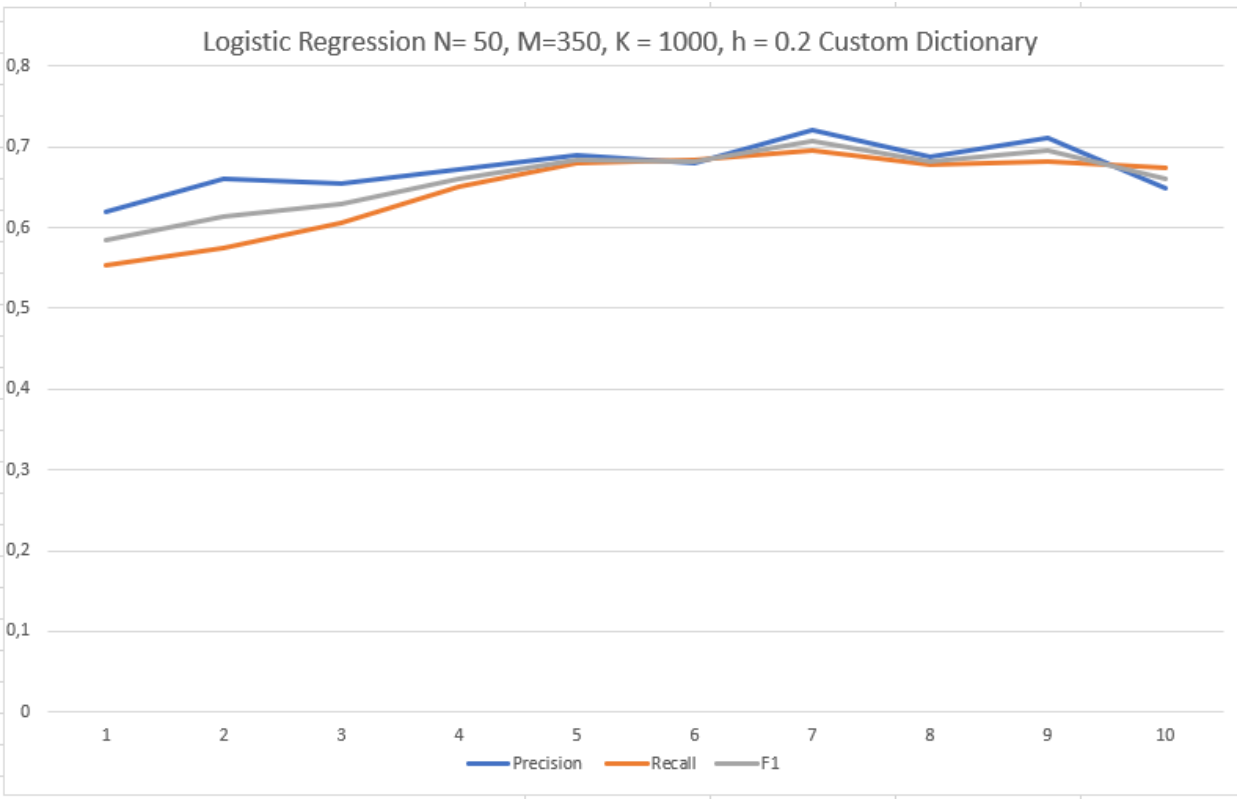
- N = 50, M = 350 (πρακτικά αυτό σημαίνει ότι σαν ιδιότητες χρησιμοποιήθηκαν 250 μεταβλητές, από την λέξη στη θέση 51 έως αυτή στη θέση 350)
- K = 1000 (πρακτικά αυτό σημαίνει ότι θα ανανεώσουμε 1000 φορές των πίνακα των βαρών)
- λ = 0.2 (πρακτικά αυτό είναι το μέγεθος του βήματος του διανύσματος των βαρών προς τη σωστή κατεύθυνση, προτιμάται μικρό)

	Logistic Regression N= 50, M=350, K = 1000, h = 0.2 Custom Dictionary		Test	Test	Test
datacount	AccuracyTrain	AccuracyTest	Precision	Recall	F1
100	1	0,64	0,62	0,553571429	0,58490566
200	0,99	0,55	0,66	0,573913043	0,613953488
400	0,9275	0,6	0,655	0,606481481	0,629807692
1000	0,77	0,665	0,672	0,649903288	0,660766962
2000	0,737	0,6935	0,69	0,679802956	0,684863524
4000	0,7095	0,68075	0,6795	0,684634761	0,682057716
8000	0,6875	0,676	0,7205	0,695799131	0,707934169
10000	0,6798	0,6776	0,687	0,678451511	0,682698996
20000	0,6953	0,6906	0,7102	0,681573896	0,695592556
25000	0,68568	0,68596	0,64864	0,673198273	0,660691004

Παρακάτω παραθέτεται η καμπύλη ορθότητας:



Παρακάτω παραθέτεται οι καμπύλες ακριβείας, recall και f1 για τα test data:



ID3

Για την υλοποίηση του κώδικα δεν χρησιμοποιήθηκε κλάδεμα για το δέντρο του ID3. Ο κώδικας βασίζεται κυρίως στον ψευδοκώδικα που παραδόθηκε στις διαλέξεις του μαθήματος.

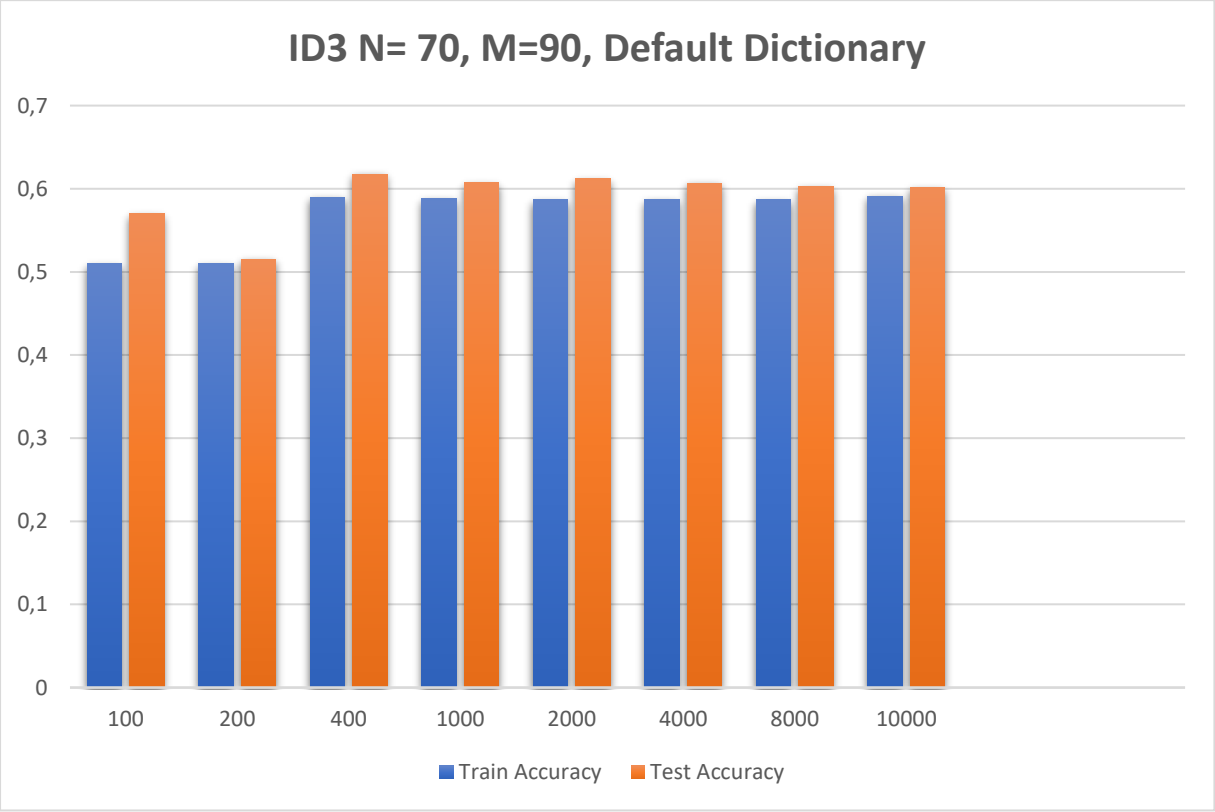
Λεπτομέρειες για τον κώδικα καλύπτονται με τη μορφή σχολίων.

Οπότε παρακάτω παραδίδεται πίνακας ορθότητας των test data και train data σε συνάρτηση με το πλήθος των train data . Οι προκαθορισμένες τιμές των παραμέτρων είναι

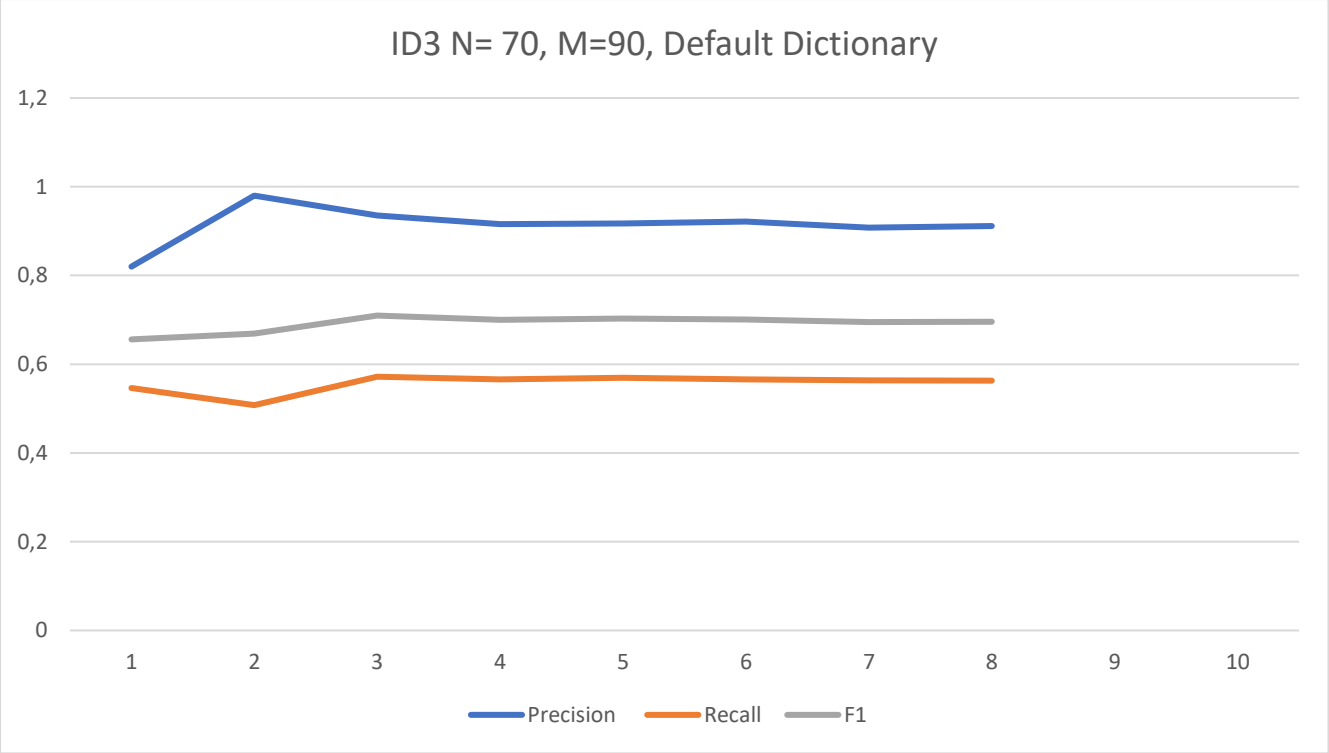
- N = 50, M = 350 (πρακτικά αυτό σημαίνει ότι σαν ιδιότητες χρησιμοποιήθηκαν 250 μεταβλητές, από την λέξη στη θέση 51 έως αυτή στη θέση 350)

	ID3 N= 70, M=90, Default Dictionary		Test	Test	Test
datacount	AccuracyTrain	AccuracyTest	Precision	Recall	F1
100	0.51	0.57	0.82	0.546667	0.656
200	0.51	0.515	0.98	0.507772	0.668942
400	0.59	0.6175	0.935	0.571865	0.709677
1000	0.588	0.607	0.916	0.566131	0.699771
2000	0.5865	0.612	0.917	0.569565	0.702682
4000	0.58675	0.60675	0.9215	0.565511	0.700894
8000	0.586625	0.602125	0.90775	0.563382	0.695261
10000	0.591	0.6015	0.9112	0.562678	0.695732

Παρακάτω παραθέτεται η καμπύλη ορθότητας:



Παρακάτω παραθέτεται οι καμπύλες ακριβείας, recall και f1 για τα test data:


ID3 N= 70, M=90, Default Dictionary
1,2
1
0,8
0,6
0,4
0,2
0
1
2
3
4
5
6
7
8
9
10
Precision
Recall
F1