2η Εργασία Τεχνητή Νοημοσύνη

Δικονιμάκη Χρυσούλα p3170039

Κωνσταντόπουλος Κωνσταντίνος Ηλίας p3170086

Λαπάκης Γεράσιμος p3170089

<u>Σύνοψη Αρχείων ργ για Προεπεξεργασία Δεδομένων:</u>

split_enron: Διαχωρίζει τυχαία τα παραδείγματα στα 3, δίνοντας 80% στα training παραδείγματα, 10% στα development παραδείγματα και 10% στα test παραδείγματα και τα αποθηκεύει στον φάκελο examples_edit. Διατρέχοντας κανείς στα αρχεία training_path(training_types), dev_path(dev_types) και test_path(test_types) μπορεί να δει ποια μηνύματα χρησιμοποιήθηκαν για εκπαίδευση, ποια για επικύρωση και ποια για έλεγχο αντίστοιχα.

plit_pu: Διαχωρίζει τα δεδομένα κρατώντας part1-part8 για training, το part 9 για development και το part 10 για test και τα αποθηκεύει στο φάκελο examples_edit στα αντίστοιχα αρχεία.

<u>Αλγόριθμοι:</u>

Γενική δομή αρχείων: Για κάθε αλγόριθμο υπάρχουν 2 αρχεία, 1 αρχείο main και 1 αρχείο με τις συναρτήσεις για την εκπαίδευση και την παραγωγή προβλέψεων/αποτελεσμάτων του αλγορίθμου(και διάφορες άλλες βοηθητικές συναρτήσεις).

Όλες τις υπερ-παραμέτρους τις επιλέξαμε ελέγχοντας όλους τους δυνατούς συνδυασμούς υπερ-παραμέτρων από τα σύνολα ενδεικτικών τιμών που ορίσαμε γι' αυτές, δηλαδή κάνουμε grid search. Επιλέγαμε το συνδυασμό που έδινε το μεγαλύτερο accuracy.
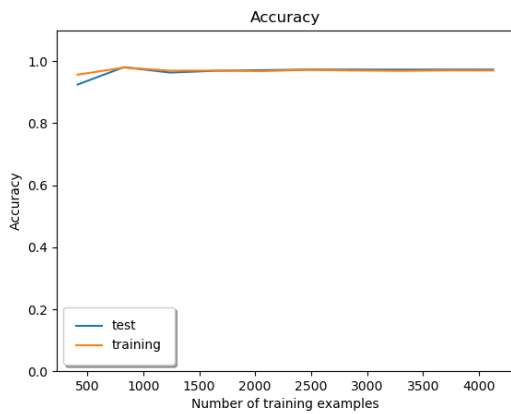
Σύνολο Δεδομένων: <u>Enron1</u>

Naive Bayes Bernoulli

Υπερ-παράμετροι που κληθήκαμε να επιλέξουμε:

1. Πλήθος Ιδιοτήτων

```
Max uperparametros m:  1500
```

Accuracy:

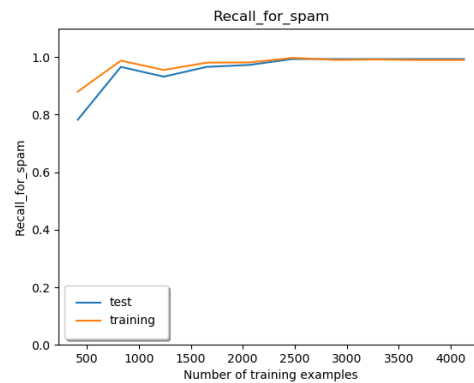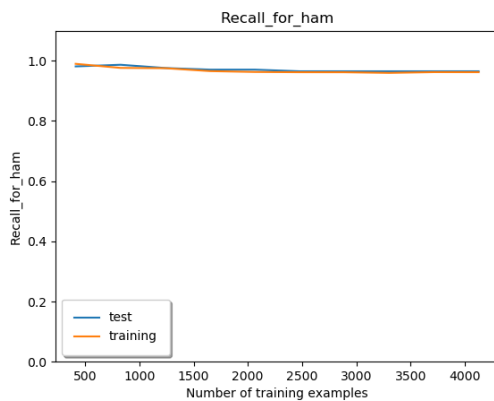| 1	x	test	training
2	413	0.9245647969052224	0.9564164648910412
3	827	0.9806576402321083	0.9794437726723095
4	1240	0.9632495164410058	0.9693548387096774
5	1653	0.9690522243713733	0.969751966122202
6	2063	0.9709864603481625	0.968007755695589
7	2475	0.9729206963249516	0.9721212121212122
8	2889	0.9729206963249516	0.9698857736240913
9	3301	0.9729206963249516	0.9687973341411693
10	3713	0.9729206963249516	0.9701050363587396
11	4127	0.9729206963249516	0.9699539617155318

Precision:
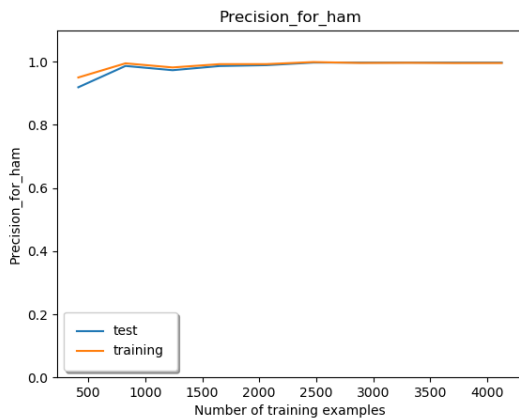
1	x	test	training
2	413	0.9189873417721519	0.95
3	827	0.9864864864864865	0.9947735191637631
4	1240	0.9730458221024259	0.9817559863169898
5	1653	0.9862637362637363	0.9922010398613518
6	2063	0.9889807162534435	0.992296918767507
7	2475	0.9972067039106145	0.998814463544754
8	2889	0.9972067039106145	0.9959758551307847
9	3301	0.9972067039106145	0.9964570416297609
10	3713	0.9972067039106145	0.9956777996070727
11	4127	0.9972067039106145	0.995768688293371
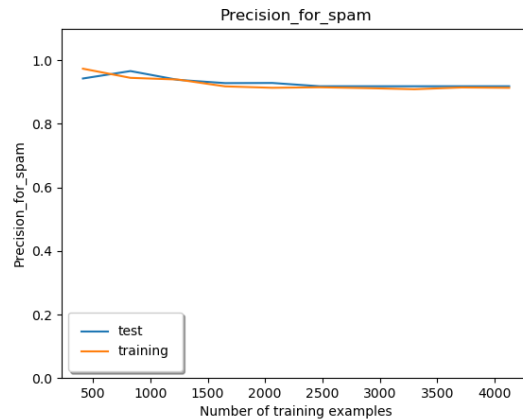
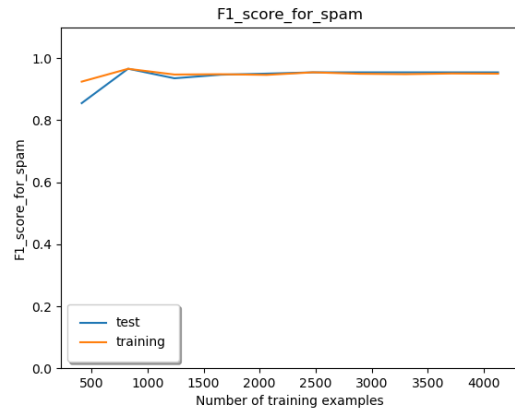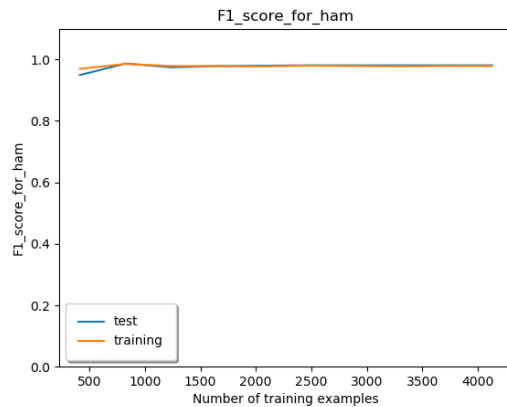| 1	x	test	training
2	413	0.9426229508196722	0.9734513274336283
3	827	0.9659863945578231	0.9446640316205533
4	1240	0.9383561643835616	0.9393939393939394
5	1653	0.9281045751633987	0.9178356713426854
6	2063	0.9285714285714286	0.9133858267716536
7	2475	0.9182389937106918	0.9149746192893401
8	2889	0.9182389937106918	0.9123196448390677
9	3301	0.9182389937106918	0.9089165867689357
10	3713	0.9182389937106918	0.9143835616438356
11	4127	0.9182389937106918	0.9132455460883037

Recall:

x	test	training
413	0.981081081081081	0.9895833333333334
827	0.9864864864864865	0.976068376068376
1240	0.9756756756756757	0.9750849377123443
1653	0.9702702702702702	0.9654300168634065
2063	0.9702702702702702	0.9626358695652174
2475	0.9648648648648649	0.96175799086758
2889	0.9648648648648649	0.9616318601262749
3301	0.9648648648648649	0.9594882729211087
3713	0.9648648648648649	0.9620349278663629
4127	0.9648648648648649	0.9618528610354223

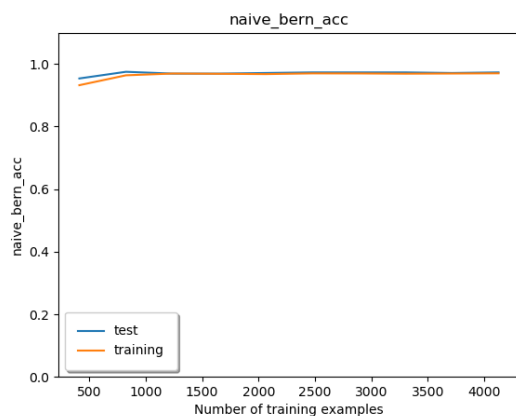| x	test	training
413	0.782312925170068	0.88
827	0.9659863945578231	0.987603305785124
1240	0.9319727891156463	0.9551820728291317
1653	0.9659863945578231	0.9807280513918629
2063	0.9727891156462585	0.9813874788494078
2475	0.9931972789115646	0.9972337482710927
2889	0.9931972789115646	0.9903614457831326
3301	0.9931972789115646	0.9916317991631799
3713	0.9931972789115646	0.989805375347544
4127	0.9931972789115646	0.9899244332493703

F1:

F1_score_for_ham

F1_score_for_spam

x	test	training
413	0.9490196078431373	0.9693877551020408
827	0.9864864864864865	0.9853321829163072
1240	0.9743589743589743	0.978409090909091
1653	0.9782016348773842	0.9786324786324786
2063	0.9795361527967258	0.9772413793103449
2475	0.9807692307692307	0.9799360279150916
2889	0.9807692307692307	0.9785025945144551
3301	0.9807692307692307	0.9776232891592439
3713	0.9807692307692307	0.9785672909828153
4127	0.9807692307692307	0.9785169785169786

x	test	training
413	0.8550185873605949	0.9243697478991596
827	0.9659863945578231	0.9656565656565657
1240	0.9351535836177475	0.9472222222222223
1653	0.9466666666666667	0.94824016563147
2063	0.9501661129568106	0.9461663947797716
2475	0.954248366013072	0.9543348775645268
2889	0.954248366013072	0.9497400346620452
3301	0.954248366013072	0.9484742371185593
3713	0.954248366013072	0.9506008010680908
4127	0.954248366013072	0.9500402900886382

scikit(BernoulliNB):

naive_bern_acc

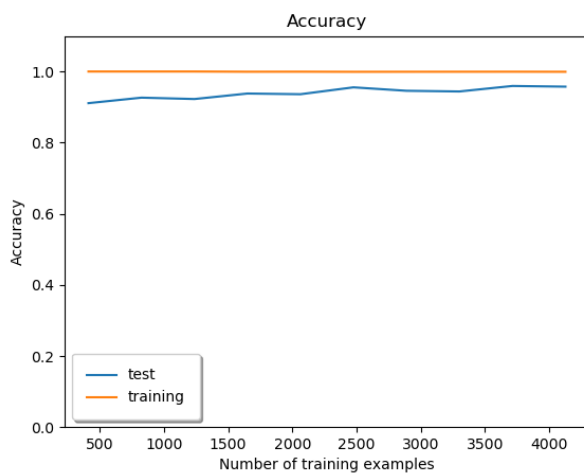| x	test	training
412	0.9535783365570599	0.9320388349514563
825	0.9748549323017408	0.9636363636363636
1238	0.9690522243713733	0.9693053311793215
1650	0.9690522243713733	0.9684848484848485
2063	0.9709864603481625	0.9675230247212797
2476	0.9729206963249516	0.9697092084006462
2888	0.9729206963249516	0.9695290858725761
3301	0.9729206963249516	0.9687973341411693
3714	0.9709864603481625	0.9693053311793215
4127	0.9729206963249516	0.9699539617155318

ID3 με πρόωρο τερματισμό της επέκτασης του δέντρου:

Υπερ-παράμετροι που κληθήκαμε να επιλέξουμε:

1. Πλήθος ιδιοτήτων
2. Ελάχιστο ποσοστό παραδειγμάτων που θα ανήκουν σε μια συγκεκριμένη κατηγορία σε κάθε κόμβο του δέντρου.

```
Max uperparametros m:   1500 Max uperparametros l:   1.0
```

Accuracy:

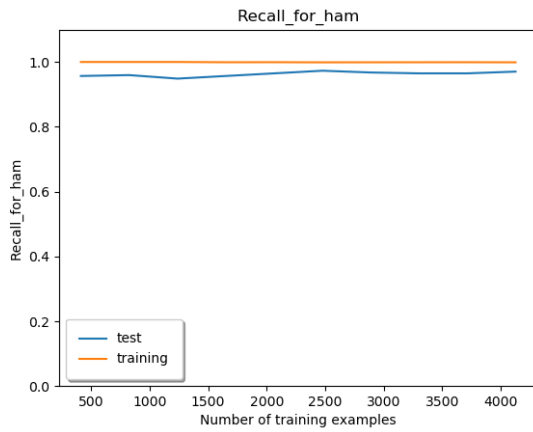| x	test	training
412	0.9110251450676983	1.0
825	0.9264990328820116	1.0
1238	0.9226305609284333	1.0
1650	0.9381044487427466	0.9993939393939394
2063	0.9361702127659575	0.9995152690256908
2476	0.9555125725338491	0.9991922455573505
2888	0.9458413926499033	0.9993074792243767
3301	0.9439071566731141	0.9993941229930324
3714	0.9593810444874274	0.9994614970382337
4127	0.9574468085106383	0.9992730797189242

Precision:
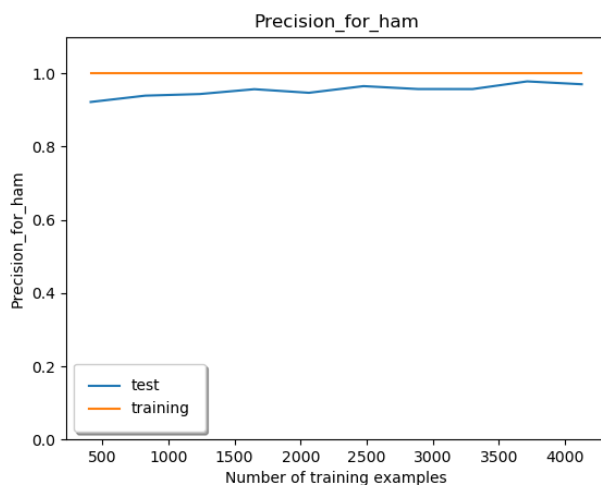
x	test	training
412	0.921875	1.0
825	0.9391534391534392	1.0
1238	0.9435483870967742	1.0
1650	0.9567567567567568	1.0
2063	0.946949602122016	1.0
2476	0.9651474530831099	1.0
2888	0.9572192513368984	1.0
3301	0.9571045576407506	1.0
3714	0.9780821917808219	1.0
4127	0.9702702702702702	1.0
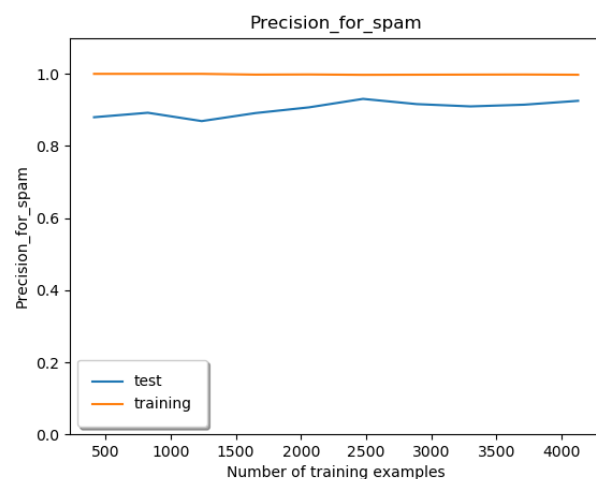
x	test	training
412	0.8796992481203008	1.0
825	0.8920863309352518	1.0
1238	0.8689655172413793	1.0
1650	0.891156462585034	0.9978586723768736
2063	0.9071428571428571	0.9983108108108109
2476	0.9305555555555556	0.9972413793103448
2888	0.916083916083916	0.9975961538461539
3301	0.9097222222222222	0.9979123173277662
3714	0.9144736842105263	0.9981498612395929
4127	0.9251700680272109	0.9974874371859297

Recall:

Recall_for_ham

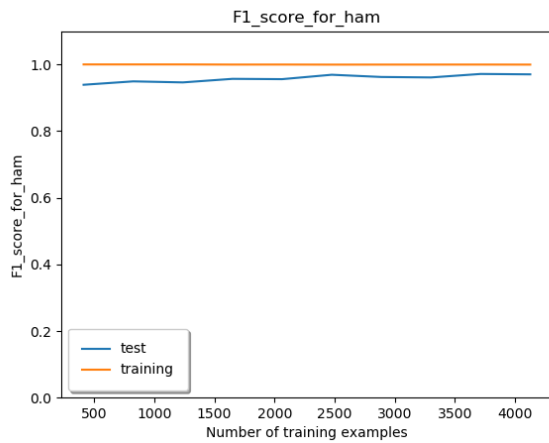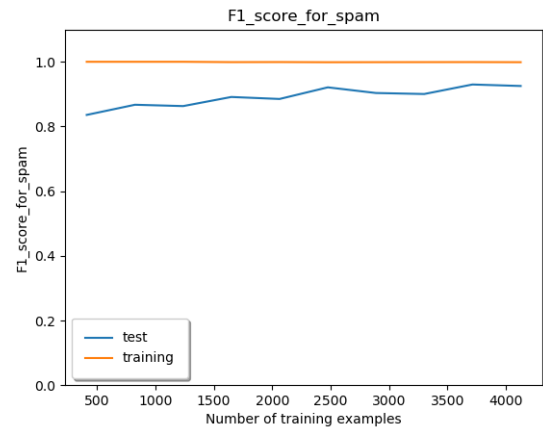|	x	test	training
1			
2	412	0.9567567567567568	1.0
3	825	0.9594594594594594	1.0
4	1238	0.9486486486486486	1.0
5	1650	0.9567567567567568	0.9991554054054054
6	2063	0.9648648648648649	0.9993206521739131
7	2476	0.972972972972973	0.9988590986879635
8	2888	0.9675675675675676	0.9990281827016521
9	3301	0.9648648648648649	0.9991471215351813
10	3714	0.9648648648648649	0.9992409867172676
11	4127	0.9702702702702702	0.9989782016348774

Recall_for_spam
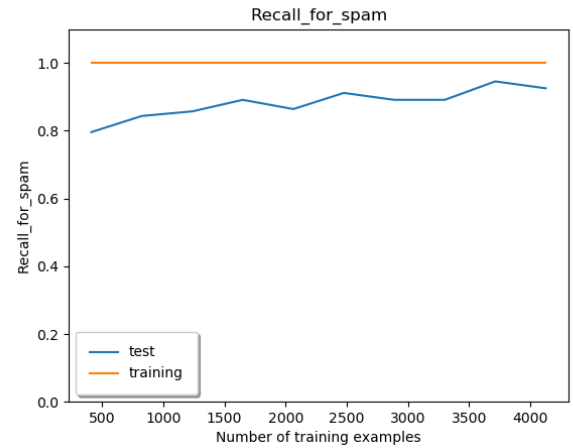
	x	test	training
1			
2	412	0.7959183673469388	1.0
3	825	0.8435374149659864	1.0
4	1238	0.8571428571428571	1.0
5	1650	0.891156462585034	1.0
6	2063	0.8639455782312925	1.0
7	2476	0.9115646258503401	1.0
8	2888	0.891156462585034	1.0
9	3301	0.891156462585034	1.0
10	3714	0.9455782312925171	1.0
11	4127	0.9251700680272109	1.0

F1:

F1_score_for_ham

	x	test	training
1			
2	412	0.9389920424403183	1.0
3	825	0.9491978609625668	1.0
4	1238	0.9460916442048517	1.0
5	1650	0.9567567567567568	0.9995775242923532
6	2063	0.9558232931726909	0.999660210669385
7	2476	0.9690444145356663	0.9994292237442922
8	2888	0.9623655913978495	0.9995138551288284
9	3301	0.9609690444145357	0.9995733788395904
10	3714	0.9714285714285715	0.9996203492786636
11	4127	0.9702702702702702	0.999488839666042

F1_score_for_spam

	x	test	training
1			
2	412	0.8357142857142857	1.0
3	825	0.8671328671328672	1.0
4	1238	0.8630136986301369	1.0
5	1650	0.891156462585034	0.9989281886387995
6	2063	0.8850174216027874	0.9991546914623838
7	2476	0.9209621993127148	0.9986187845303868
8	2888	0.9034482758620689	0.9987966305655837
9	3301	0.9003436426116839	0.9989550679205852
10	3714	0.9297658862876254	0.999074074074074
11	4127	0.9251700680272109	0.9987421383647799

scikit(tree.DecisionTreeClassifier):

dec_tree_acc
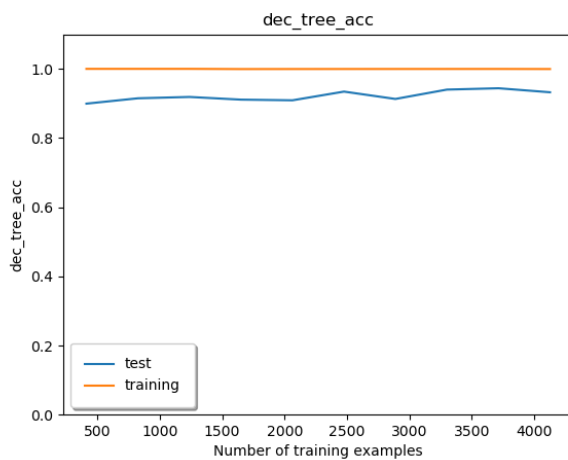
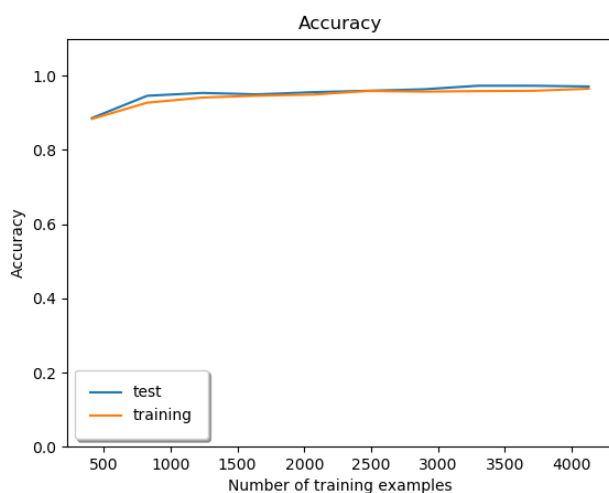| 1	x	test	training
2	412	0.8994197292069632	1.0
3	825	0.9148936170212766	1.0
4	1238	0.9187620889748549	1.0
5	1650	0.9110251450676983	0.9993939393939394
6	2063	0.9090909090909091	0.9995152690256908
7	2476	0.9342359767891683	0.9995961227786753
8	2888	0.9129593810444874	0.9996537396121884
9	3301	0.9400386847195358	0.9996970614965162
10	3714	0.9439071566731141	0.9997307485191168
11	4127	0.9323017408123792	0.9995153864792827

Λογιστική παλινδρόμηση

Υπερ-παράμετροι που κληθήκαμε να επιλέξουμε:

1. Πλήθος ιδιοτήτων
2. Learning Rate (η)
3. Παράμετρος Κανονικοποίησης (λ)
   Θα τρέξουμε τον αλγόριθμο με πλήθος εποχών=10.

```
Max m:  500
Max lambda:  0.0001
Max eta:  0.01
```

Accuracy:

Accuracy

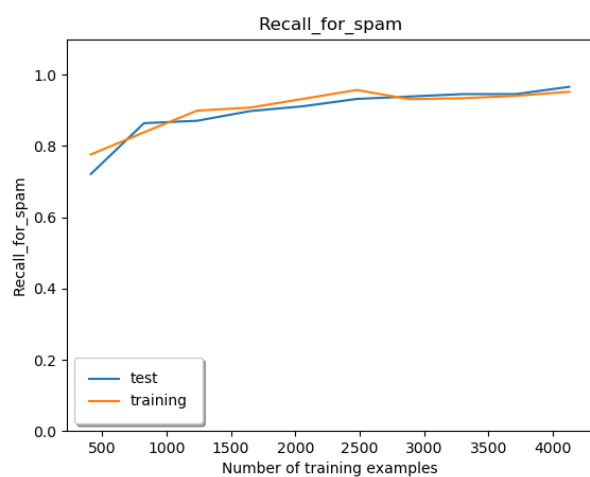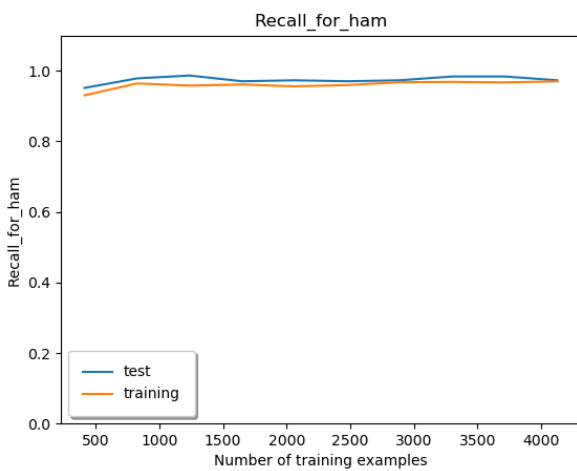| 1	x	test	training
2	412	0.8858800773694391	0.883495145631068
3	825	0.9458413926499033	0.9272727272727272
4	1238	0.9535783365570599	0.9410339256865913
5	1650	0.9497098646034816	0.946060606060606
6	2063	0.9555125725338491	0.9491032476975279
7	2476	0.9593810444874274	0.9588045234248789
8	2888	0.9632495164410058	0.9570637119113573
9	3301	0.9729206963249516	0.9584974250227204
10	3714	0.9729206963249516	0.9593430263866451
11	4127	0.9709864603481625	0.9648655197480009

Precision:
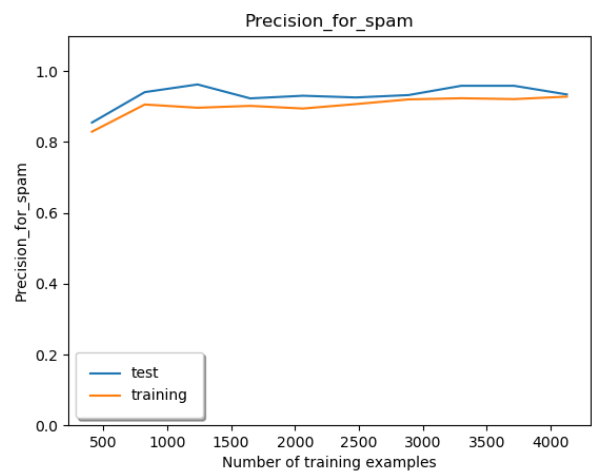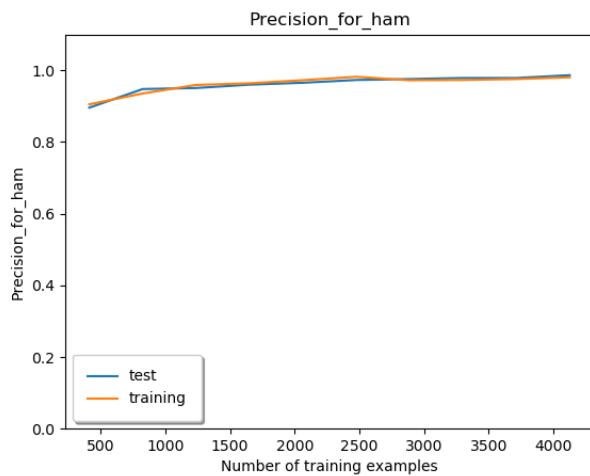
1	x	test	training
2	412	0.8956743002544529	0.9050847457627119
3	825	0.9476439790575916	0.9352159468438538
4	1238	0.9505208333333334	0.9590909090909091
5	1650	0.9598930481283422	0.9635901778154107
6	2063	0.9651474530831099	0.972356599861783
7	2476	0.9728997289972899	0.981903093987157
8	2888	0.975609756097561	0.97216796875
9	3301	0.978494623655914	0.9730077120822622
10	3714	0.978494623655914	0.9754977029096478
11	4127	0.9863013698630136	0.9803786574870912

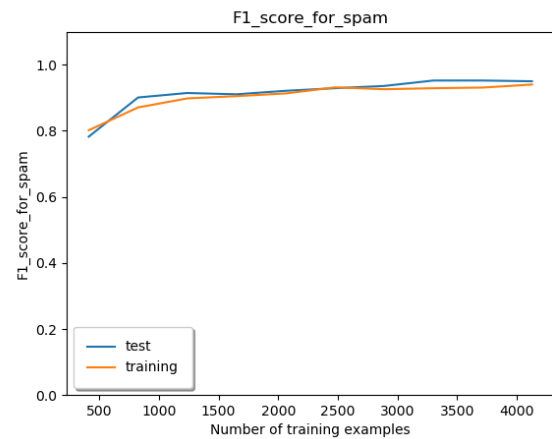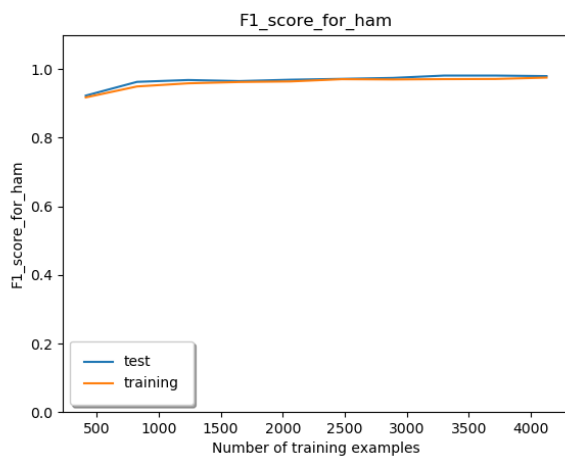| 1	x	test	training
2	412	0.8548387096774194	0.8290598290598291
3	825	0.9407407407407408	0.905829596412556
4	1238	0.9624060150375939	0.8966480446927374
5	1650	0.9230769230769231	0.9019189765458422
6	2063	0.9305555555555556	0.8944805194805194
7	2476	0.9256756756756757	0.9069462647444299
8	2888	0.9324324324324325	0.9202380952380952
9	3301	0.9586206896551724	0.9234746639089969
10	3714	0.9586206896551724	0.9210526315789473
11	4127	0.9342105263157895	0.9279869067103109

Recall:

1	x	test	training
2	412	0.9513513513513514	0.9303135888501742
3	825	0.9783783783783784	0.964041095890411
4	1238	0.9864864864864865	0.9580022701475596
5	1650	0.9702702702702702	0.9611486486486487
6	2063	0.972972972972973	0.9558423913043478
7	2476	0.9702702702702702	0.959498003422704
8	2888	0.972972972972973	0.967444120505345
9	3301	0.9837837837837838	0.9684434968017057
10	3714	0.9837837837837838	0.9669829222011386
11	4127	0.972972972972973	0.9700272479564033

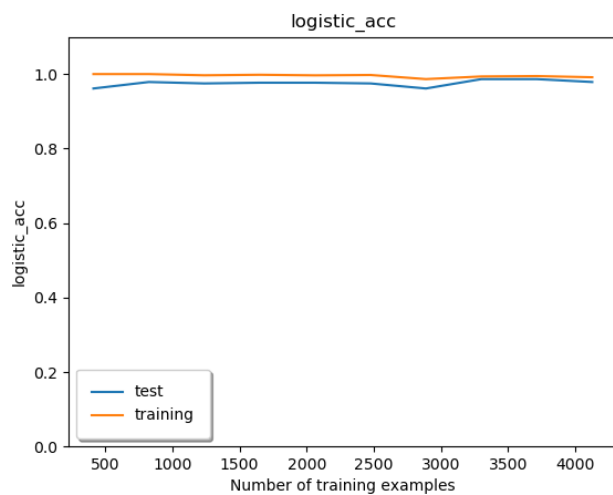| 1	x	test	training
2	412	0.7210884353741497	0.776
3	825	0.8639455782312925	0.8381742738589212
4	1238	0.8707482993197279	0.8991596638655462
5	1650	0.8979591836734694	0.907725321888412
6	2063	0.9115646258503401	0.9323181049069373
7	2476	0.9319727891156463	0.9571230982019364
8	2888	0.9387755102040817	0.9313253012048193
9	3301	0.9455782312925171	0.9341004184100419
10	3714	0.9455782312925171	0.9406858202038925
11	4127	0.9659863945578231	0.9521410579345088

F1:

F1_score_for_ham

F1_score_for_spam

	x	test	training
1	x	test	training
2	412	0.9226736566186107	0.9175257731958762
3	825	0.9627659574468085	0.9494097807757167
4	1238	0.9681697612732096	0.9585462805224304
5	1650	0.9650537634408601	0.962367864693446
6	2063	0.9690444145356663	0.9640287769784173
7	2476	0.9715832205683356	0.9705712637045585
8	2888	0.9742895805142084	0.9698002922552362
9	3301	0.981132075471698	0.9707202393673862
10	3714	0.981132075471698	0.9712216504669334
11	4127	0.979591836734694	0.9751754836500599

	x	test	training
1	x	test	training
2	412	0.7822878228782288	0.8016528925619835
3	825	0.900709219858156	0.8706896551724138
4	1238	0.9142857142857143	0.897902097902098
5	1650	0.9103448275862069	0.904812834224599
6	2063	0.9209621993127148	0.9130074565037283
7	2476	0.9288135593220338	0.9313593539703903
8	2888	0.9355932203389831	0.9257485029940119
9	3301	0.952054794520548	0.9287571502860114
10	3714	0.952054794520548	0.9307657038055936
11	4127	0.9498327759197324	0.9399088271860754

scitik(Perceptron):

logistic_acc

	x	test	training
1	x	test	training
2	412	0.9613152804642167	1.0
3	825	0.9787234042553191	1.0
4	1238	0.9748549323017408	0.9967689822294022
5	1650	0.97678916827853	0.9981818181818182
6	2063	0.97678916827853	0.9966068831798351
7	2476	0.9748549323017408	0.9975767366720517
8	2888	0.9613152804642167	0.9864958448753463
9	3301	0.9864603481624759	0.9936382914268403
10	3714	0.9864603481624759	0.9946149703823371
11	4127	0.9787234042553191	0.9915192633874486