(a) Report the number of documents in each class in the training and test sets.

number of articles in the training set: 5791

number of articles in the test set: 2300

number of unique words: 16958

training set: {'earn': 2848, 'acq': 1617, 'money-fx': 536, 'grain': 429, 'crude': 361}

test set: {'earn': 1084, 'acq': 710, 'money-fx': 178, 'grain': 148, 'crude': 180}

(b) Report the k most discriminating words (where k = 50) for each class based on Mutual Information.

earn: {'offer', 'loss', 'ha', 'would', 'u', 'dividend', 'vs', 'sell', 'offici', 'prior', 'year', 'by', 'todai', 'about', 'purchas', 'profit', 'exchang', 'export', 'net', 'rev', 'qtly', 'pct', 'ct', 'to', 'qtr', 'avg', 'mth', 'said', 'acquir', 'sai', 'lt', 'shr', '4th', 'div', '1', 'market', 'that', 'bui', 'at', 'not', 'price', 'agre', 'note', 's', 'had', '31', 'record', 'bank', 'agreement', '2'}

---

acq: {'offer', 'loss', 'ha', 'subsidiari', 'acquisit', 'dividend', 'sell', 'undisclos', 'year', 'unit', 'purchas', 'common', 'stake', 'commiss', 'profit', 'group', 'corp', 'net', 'rev', 'outstand', 'bid', 'ct', 'to', 'qtr', 'pct', 'complet', 'compani', 'said', 'merger', 'share', 'sharehold', 'acquir', 'transact', 'lt', 'shr', 'approv', 'file', '1', 'bui', 'agre', 'note', 'inc', 'term', '31', 'record', 'control', 'agreement', 'own', 'takeov', 'disclos'}

---

money-fx: {'currenc', 'treasuri', 'dollar', 'reserv', '000', 'england', 'u', 'bill', 'stabil', 'nation', 'trade', 'deficit', 'economi', 'dealer', 'monetari', 'todai', 'foreign', 'repurchas', 'yen', 'japan', 'exchang', 'k', 'assist', 'around', 'net', 'corp', 'monei', 'further', 'ct', 'stg', 'compani', 'rate', 'share', 'sai', 'germani', 'against', 'lt', 'economist', 'market', 'interven', 'at', 'fed', 'inc', 'central', 'bank', 'pari', 'shortag', 'intervent', 'band', 'econom'}

---

grain: {'union', 'crop', 'farmer', 'grower', 'commod', 'season', 'maiz', 'corn', 'u', 'ussr', 'shipment', 'vs', 'depart', 'offici', 'bushel', 'wheat', 'sorghum', 'tonn', 'soviet', 'winter', '87', 'cereal', 'agricultur', 'soybean', 'china', 'program', 'enhanc', 'export', 'net', 'rice', 'corp', 'ccc', 'farm', 'ct', 'to', 'compani', 'share', 'import', 'grain', 'lt', 'harvest', 'hectar', 's', 'inc', 'barlei', 'subsidi', 'ec', 'acreag', 'feed', 'usda'}

---

crude: {'would', 'gasolin', 'output', 'ga', 'explor', 'ecuador', 'were', 'suppli', 'barrel', 'refineri', 'by', 'earthquak', 'gulf', 'iraq', 'last', 'iran', 'bpd', 'opec', 'state', 'kuwait', 'net', 'sea', 'bbl', 'ct', 'to', 'qtr', 'said', 'countri', 'crude', 'product', 'import', 'pipelin', 'sai', 'lt', 'minist', 'natur', 'produc', 'price', 'that', 'refin', 'dai', 'quota', 'at', 'oil', 'inc', 'drill', 'report', 'petroleum', 'saudi', 'energi'}

(c) Report the macro-averaged and micro-averaged precision, recall, and F-measure values obtained by your two classifiers on the test set, as well as the performance values obtained for each class separately by using Laplace smoothing with α = 1.

1. <u>MNB without FS by MI:</u>

   96.73913043478261 percent: 2225 correct out of 2300

   Evaluation Table:

{'earn': eval_tuple(tp=1045, fp=16, fn=39, tn=1200), 'acq': eval_tuple(tp=691, fp=36, fn=19, tn=1554), 'money-fx': eval_tuple(tp=176, fp=10, fn=2, tn=2112), 'grain': eval_tuple(tp=139, fp=1, fn=9, tn=2151), 'crude': eval_tuple(tp=174, fp=12, fn=6, tn=2108)}

precision: {'earn': 0.9849198868991518, 'acq': 0.9504814305364512, 'money-fx': 0.946236559139785, 'grain': 0.9928571428571429, 'crude': 0.9354838709677419}

recall: {'earn': 0.9640221402214022, 'acq': 0.9732394366197183, 'money-fx': 0.9887640449438202, 'grain': 0.9391891891891891, 'crude': 0.9666666666666667}

macro_avg_precision: 0.9619957780800545

macro_avg_recall: 0.9663762955281593

micro_avg_precision: 0.967391304347826

micro_avg_recall: 0.967391304347826

F_macro: 0.9641810613812164

F_micro: 0.967391304347826

2. MNB with FS by MI:

96.47826086956522 percent: 2219 correct out of 2300

Evaluation Table:

{'earn': eval_tuple(tp=1030, fp=10, fn=54, tn=1206), 'acq': eval_tuple(tp=695, fp=51, fn=15, tn=1539), 'money-fx': eval_tuple(tp=176, fp=7, fn=2, tn=2115), 'grain': eval_tuple(tp=147, fp=3, fn=1, tn=2149), 'crude': eval_tuple(tp=171, fp=10, fn=9, tn=2110)}

precision: {'earn': 0.9903846153846154, 'acq': 0.9316353887399463, 'money-fx': 0.9617486338797814, 'grain': 0.98, 'crude': 0.9447513812154696}

recall: {'earn': 0.9501845018450185, 'acq': 0.9788732394366197, 'money-fx': 0.9887640449438202, 'grain': 0.9932432432432432, 'crude': 0.95}

macro_avg_precision: 0.9617040038439626

macro_avg_recall: 0.9722130058937403

micro_avg_precision: 0.9647826086956521

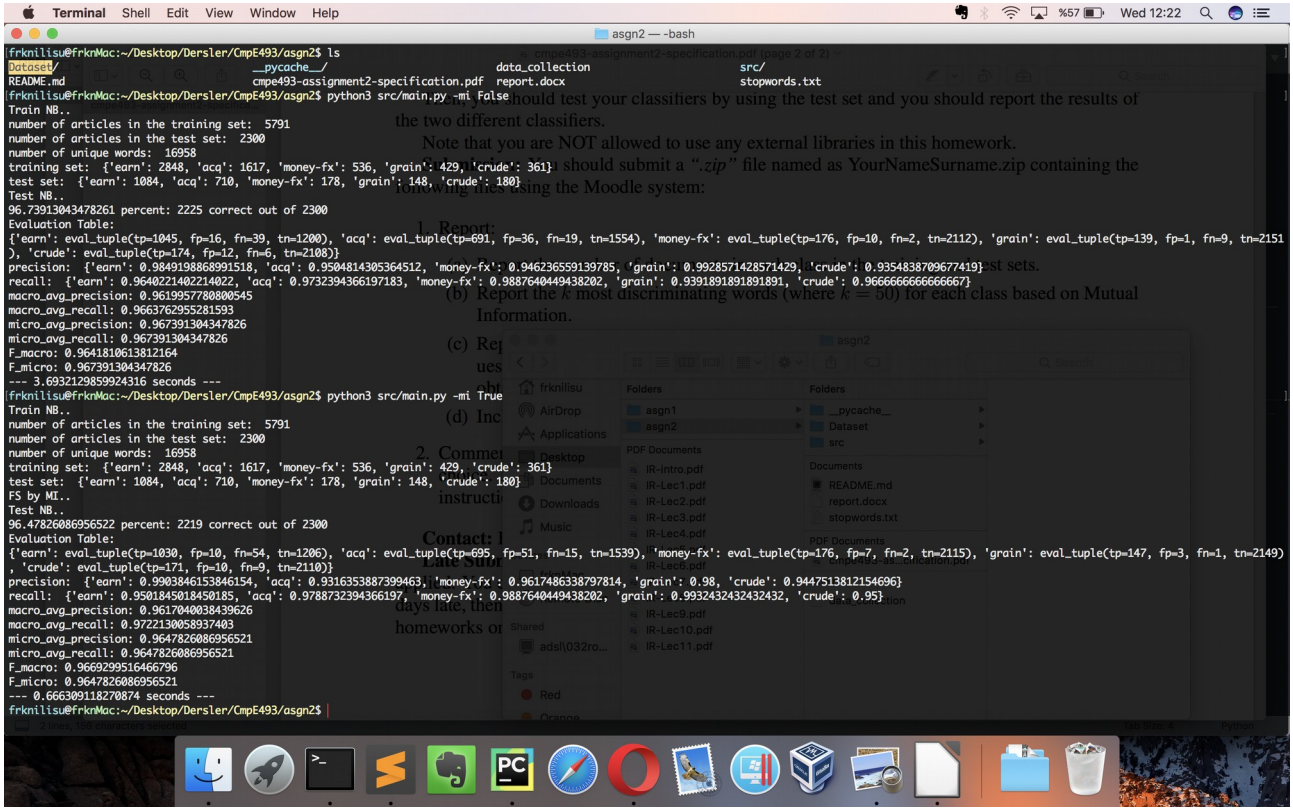
micro_avg_recall: 0.9647826086956521

F_macro: 0.9669299516466796

F_micro: 0.9647826086956521

(d) Include screenshots showing sample runs of your two programs.

at below


Terminal Shell Edit View Window Help
%57 Wed 12:22
asgn2 — -bash
frknilisu@frknMac:~/Desktop/Dersler/CmpE493/asgn2$ ls
Dataset/ __pycache__/ data_collection src/
README.md cmpe493-assignment2-specification.pdf report.docx stopwords.txt
frknilisu@frknMac:~/Desktop/Dersler/CmpE493/asgn2$ python3 src/main.py -mi False
Train NB..
number of articles in the training set: 5791
number of articles in the test set: 2300
number of unique words: 16958
training set: {'earn': 2848, 'acq': 1617, 'money-fx': 536, 'grain': 429, 'crude': 361}
test set: {'earn': 1084, 'acq': 710, 'money-fx': 178, 'grain': 148, 'crude': 180}
Test NB..
96.73913043478261 percent: 2225 correct out of 2300
Evaluation Table:
{'earn': eval_tuple(tp=1045, fp=16, fn=39, tn=1200), 'acq': eval_tuple(tp=691, fp=36, fn=19, tn=1554), 'money-fx': eval_tuple(tp=176, fp=10, fn=2, tn=2112), 'grain': eval_tuple(tp=139, fp=1, fn=9, tn=2151), 'crude': eval_tuple(tp=174, fp=12, fn=6, tn=2108)}
precision: {'earn': 0.9849198868991518, 'acq': 0.9504814305364512, 'money-fx': 0.946236559139785, 'grain': 0.9928571428571429, 'crude': 0.935483870967741}
recall: {'earn': 0.9640221402214022, 'acq': 0.9732394366197183, 'money-fx': 0.9887640449438202, 'grain': 0.9391891891891891, 'crude': 0.9666666666666667}
macro_avg_precision: 0.9619957780800545
macro_avg_recall: 0.9663762955281593
micro_avg_precision: 0.967391304347826
micro_avg_recall: 0.967391304347826
F_macro: 0.9641810613812164
F_micro: 0.967391304347826
--- 3.6932129859924316 seconds ---
frknilisu@frknMac:~/Desktop/Dersler/CmpE493/asgn2$ python3 src/main.py -mi True
Train NB..
number of articles in the training set: 5791
number of articles in the test set: 2300
number of unique words: 16958
training set: {'earn': 2848, 'acq': 1617, 'money-fx': 536, 'grain': 429, 'crude': 361}
test set: {'earn': 1084, 'acq': 710, 'money-fx': 178, 'grain': 148, 'crude': 180}
FS by MI..
Test NB..
96.47826086956522 percent: 2219 correct out of 2300
Evaluation Table:
{'earn': eval_tuple(tp=1030, fp=10, fn=54, tn=1206), 'acq': eval_tuple(tp=695, fp=51, fn=15, tn=1539), 'money-fx': eval_tuple(tp=176, fp=7, fn=2, tn=2115), 'grain': eval_tuple(tp=147, fp=3, fn=1, tn=2149), 'crude': eval_tuple(tp=171, fp=10, fn=9, tn=2110)}
precision: {'earn': 0.9903846153846154, 'acq': 0.9316353887399463, 'money-fx': 0.9617486338797814, 'grain': 0.98, 'crude': 0.9447513812154696}
recall: {'earn': 0.9501845018450185, 'acq': 0.9788732394366197, 'money-fx': 0.9887640449438202, 'grain': 0.9932432432432432, 'crude': 0.95}
macro_avg_precision: 0.9617040038439626
macro_avg_recall: 0.9722130058937403
micro_avg_precision: 0.9647826086956521
micro_avg_recall: 0.9647826086956521
F_macro: 0.9669299516466796
F_micro: 0.9647826086956521
--- 0.6663091182708740 seconds ---
frknilisu@frknMac:~/Desktop/Dersler/CmpE493/asgn2$