# Bogazici University

## CmpE493
## Homework 2

Spring 2018

Irmak Kavasoglu
2013400090

Apr 5, 2018

# 1. Implementation Summary

For this project, I have used Java programming language with Eclipse IDE. In this chapter, I will try to explain the steps I have taken to finish the project and explain details of the implementation. You can find the project progress steps in my github.

The preprocessing of the files are the same as the previous project, you can check my previous report as I will not write the same process again in this report.

## 1.1. Calculations without mutual information

The *calculateTopicProbabilities* function calculates the topic probabilities by counting each topic and dividing by the total count of the proper training documents.

The *countTermsForTopic* function counts the terms for each topic, using the term counts in each story.

The *calculateTermProbabilities* function calculates the term probabilities for each topic, using the term counts of each topic.

The *classifyTestDocuments* function classified the test documents using the calculated term probabilities and topic probabilities.

All of the probabilities are kept as the *log probabilities* and the logs are *natural logarithms* with base *e*.

## 1.2. Calculations with mutual information

The *calculateMutualInformation* function returns the most distinctive 50 words of each topic. These are listed in the following section.

The *distinctiveTerms* set is constucted from these 250 words. There are overlaps, so the new term count is around 170.

The *updateDocumentsWithWords* function discards all words that are not distinct.

The steps above are repeated with updated dictionary and documents, including recalculation of term probabilities and classification.

Within the classification function, the statistics are printed.

# 2. Statistical Information

This section answers the questions from the project description.

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

While reading the documents, I am discarding the documents which do not follow the topic criteria, that is, has more than one of the five topics or does not belong to any of them.

After this filtering, I am left with **8091** documents.

Looking at the *lewissplit* values of these well-topiced documents, we have **5791** documents for **training** and **2300** documents for **testing**.

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

The most discriminating words for each topic are listed as follows:

- **money-fx**: [reserv, corp, fed, monei, nation, japan, bank, monetari, shortag, econom, band, net, england, rev, level, todai, stg, stabil, bill, qtr, market, at, shr, exchang, interven, vs, currenc, against, treasuri, compani, around, dollar, foreign, central, rate, germani, share, sai, pari, deficit, inc, yen, economist, intervent, ct, trade, repurchas, dealer, assist, further]
- **earn**: [year, govern, payout, about, had, offer, div, loss, bank, would, prior, price, jan, record, net, export, mth, rev, agreement, todai, 4th, qtr, market, purchas, at, were, shr, exchang, vs, acquir, note, qtly, not, 1st, avg, dividend, wheat, sai, said, profit, pct, tender, more, sell, ct, agre, bui, offici, ha, to]
- **grain**: [sorghum, sourc, corp, corn, program, enhanc, agricultur, acreag, season, farm, grain, depart, net, export, ec, rev, shipment, usda, qtr, feed, cereal, subsidi, shr, vs, grower, crop, import, winter, compani, ussr, tonn, wheat, barlei, share, commod, inc, soybean, harvest, hectar, union, rice, bushel, soviet, ct, china, maiz, offici, farmer, to, under]
- **acq**: [corp, disclos, year, own, complet, outstand, offer, div, loss, record, net, subsidiari, group, rev, sharehold, 4th, qtr, purchas, unit, transact, shr, vs, acquir, note, qtly, undisclos, compani, acquisit, avg, file, common, takeov, dividend, share, term, said, profit, inc, pct, sell, control, stake, ct, agre, bui, ha, to, bid, approv, merger]
- **crude**: [minist, opec, dai, refin, output, countri, would, price, quota, state, net, petroleum, earthquak, rev, last, iran, iraq, gasolin, bbl, qtr, saudi, at, were, explor, shr, produc, vs, suppli, ecuador, refineri, import, natur, sea, crude, oil, bpd, gulf, sai, ga, said, inc, barrel, kuwait, product, pipelin, drill, energi, ct, report, to]

*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 = 1.0.*

The values are as follows.

*Without mutual information*

- Macro-averaged precision:
  Precision for topic **earn**: 0.9896810506566605
  Precision for topic **acq**: 0.9667128987517337
  Precision for topic **money-fx**: 0.9672131147540983
  Precision for topic **grain**: 0.9932432432432432
  Precision for topic **crude**: 0.9560439560439561
  **Macro-average**: 0.9745788526899384

- Micro-averaged precision:
  True positives for topic **money-fx**: 177
  True positives for topic **earn**: 1055
  True positives for topic **grain**: 147
  True positives for topic **acq**: 697
  True positives for topic **crude**: 174
  **Micro-average**: 2250/2300 = 0.9782608695652174

- Macro-Recall:
  Recall for topic **money-fx**:0.9943820224719101
  Recall for topic **earn**: 0.9732472324723247
  Recall for topic **grain**: 0.9932432432432432
  Recall for topic **acq**: 0.9816901408450704
  Recall for topic **crude**:0.9666666666666667
  **Macro-Recall**: 0.9818458611398431

- Micro-Recall:
  True positives for topic **money-fx**: 177
  True positives for topic **earn**: 1055
  True positives for topic **grain**: 147
  True positives for topic **acq**: 697
  True positives for topic **crude**: 174
  **Micro-Recall**: 2250/2300 = 0.9782608695652174

- **Macro F-Measure**: 0.9781988605070203

- **Micro F-Measure**: 0.9782608695652174

*With mutual information*

- Macro-averaged precision:
  Precision for topic **earn**: 0.9902248289345064
  Precision for topic **acq**: 0.9195250659630607
  Precision for topic **money-fx**: 0.9615384615384616
  Precision for topic **grain**: 0.9735099337748344
  Precision for topic **crude**: 0.9247311827956989
  **Macro-average**: 0.9539058946013123

- Micro-averaged precision:
  True positives for topic **money-fx**: 175
  True positives for topic **earn**: 1013
  True positives for topic **grain**: 147
  True positives for topic **acq**: 697
  True positives for topic **crude**: 172
  **Micro-average**: 2204/2300 = 0.9582608695652174

- Macro-Recall:
  Recall for topic **money-fx**: 0.9831460674157303
  Recall for topic **earn**: 0.9345018450184502
  Recall for topic **grain**: 0.9932432432432432
  Recall for topic **acq**: 0.9816901408450704
  Recall for topic **crude**: 0.9555555555555556
  **Macro-Recall**: 0.96962737041561

- Micro-Recall:
  True positives for topic **money-fx**: 175
  True positives for topic **earn**: 1013
  True positives for topic **grain**: 147
  True positives for topic **acq**: 697
  True positives for topic **crude**: 172
  **Micro-Recall**: 2204/2300 = 0.9582608695652174

- **Macro F-Measure**: 0.9781988605070203

- **Micro F-Measure**: 0.9582608695652174

# 3. Screenshots

The screenshots of the program are as follows:

The initial steps and calculations

```
Reading documents...
Reading documents DONE.
Tokenizing documents...
Tokenizing document 01/22: [#---------------------]
Tokenizing document 02/22: [##--------------------]
Tokenizing document 03/22: [###-------------------]
Tokenizing document 04/22: [####------------------]
Tokenizing document 05/22: [#####-----------------]
Tokenizing document 06/22: [######----------------]
Tokenizing document 07/22: [#######---------------]
Tokenizing document 08/22: [########--------------]
Tokenizing document 09/22: [#########-------------]
Tokenizing document 10/22: [##########------------]
Tokenizing document 11/22: [###########-----------]
Tokenizing document 12/22: [############----------]
Tokenizing document 13/22: [#############---------]
Tokenizing document 14/22: [##############--------]
Tokenizing document 15/22: [###############-------]
Tokenizing document 16/22: [################------]
Tokenizing document 17/22: [#################-----]
Tokenizing document 18/22: [##################----]
Tokenizing document 19/22: [###################---]
Tokenizing document 20/22: [####################--]
Tokenizing document 21/22: [#####################-]
Tokenizing document 22/22: [######################]
Tokenizing documents DONE.
Creating dictionary...
Creating dictionary DONE.
Calculating probabilities of terms...
Calculating probabilities of terms DONE.
Classifying test documents...
```

Calculations without mutual information

```
Correctly classified: 2250/2300=0.9782608695652174

Correcly classified documents by topic: {money-fx=177, earn=1055, grain=147, acq=697, crude=174}
Falsely classified documents by their classified topic: {money-fx=6, earn=11, grain=1, acq=24, crude=8}
Falsely classified documents by their actual topic: {money-fx=5, earn=10, grain=1, acq=25, crude=9}

Precision for topic earn: 0.9896810506566605
Recall for topic earn: 0.9906103286384976
Precision for topic acq: 0.9667128987517337
Recall for topic acq: 0.9653739612188366
Precision for topic money-fx: 0.9672131147540983
Recall for topic money-fx: 0.9725274725274725
Precision for topic grain: 0.9932432432432432
Recall for topic grain: 0.9932432432432432
Precision for topic crude: 0.9560439560439561
Recall for topic crude: 0.9508196721311475
```

Calculations with mutual information

```
Calculating mutual information...
Calculating mutual information DONE.
Classifying test documents with mutual information...

Correctly classified: 2204/2300=0.9582608695652174

Correcly classified documents by topic: {money-fx=175, earn=1013, grain=147, acq=697, crude=172}
Falsely classified documents by their classified topic: {money-fx=7, earn=10, grain=4, acq=61, crude=14}
Falsely classified documents by their actual topic: {money-fx=8, earn=8, grain=2, acq=62, crude=15}

Precision for topic earn: 0.9902248289345064
Recall for topic earn: 0.9921645445641528
Precision for topic acq: 0.9195250659630607
Recall for topic acq: 0.9183135704874835
Precision for topic money-fx: 0.9615384615384616
Recall for topic money-fx: 0.9562841530054644
Precision for topic grain: 0.9735099337748344
Recall for topic grain: 0.9865771812080537
Precision for topic crude: 0.9247311827956989
Recall for topic crude: 0.9197860962566845
```