CMPE 493
Information Retrieval
Assignment 2

A Simple Document Retrieval System for Boolean Queries

Beyza Gul Gurbuz
2013400081

April 3, 2018

Number Of Documents

For Training Data

- earn: 2832
- acq: 1615
- money-fx: 532
- grain: 424
- crude: 357

For Test Data

- earn: 1077
- acq: 709
- money-fx: 178
- grain: 147
- crude: 177

Most Discriminating Words

- **earn**: 'vs', 'ct', 'shr', 'net', 'said', 'qtr', 'to', 'rev', 'the', 'loss', '4th', 'ha', 'div', 'at', 'profit', 'note', 'dividend', 'record', 'bui', '31', 'avg', 'qtly', 'u.', 'prior', 'pct', 'market', 'not', 'agreement', 'year', 'would', 'acquir', 'mth', 'offer', 'agre', 'offici', 'had', 'sell', 'purchas', 'sai', 'exchang', 'bank', 'price', 'about', 'export', '1st', 'were', 'todai', 'jan', 'wheat', 'tender'
- **acq**: 'vs', 'ct', 'shr', 'said', 'acquir', 'net', 'qtr', 'rev', 'acquisit', 'stake', 'to', 'bui', 'compani', 'merger', 'year', 'loss', '4th', 'ha', 'complet', 'offer', 'sell', 'record', 'share', 'note', 'common', 'div', 'group', 'own', 'sharehold', 'outstand', 'corp', 'undisclos', 'inc', 'purchas', 'avg', 'agre', 'approv', 'term', 'file', 'qtly', 'profit', 'subsidiari', 'bid', 'takeov', 'dividend', 'control', 'transact', 'disclos', '31', 'unit'

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

- **grain**: 'wheat', 'agricultur', 'tonn', 'grain', 'corn', 'usda', 'export', 'crop', 'u.', 'farmer', 'depart', 'ct', 'soybean', 'vs', 'soviet', 'farm', 'inc', 'commod', 'maiz', 'barlei', 'net', 'shr', 'program', '1986/87', 'qtr', 'rice', 'ec', 'bushel', 'feed', 'ussr', 'winter', 'share', 'shipment', 'cereal', 'corp', 'compani', 'rev', 'harvest', 'to', 'union', 'grower', 'the', 'sorghum', 'season', 'import', 'china', 'ccc', 'enhanc', 'subsidi', 'offici'

- **crude**: 'oil', 'barrel', 'crude', 'bpd', 'petroleum', 'opec', 'energi', 'dai', 'vs', 'price', 'refineri', 'product', 'shr', 'ga', 'minist', 'drill', 'output', 'explor', 'quota', 'ct', 'gasolin', 'net', 'said', 'to', 'ecuador', 'last', 'qtr', 'sea', 'produc', 'state', 'natur', 'at', 'rev', 'gulf', 'import', 'were', 'pipelin', 'report', 'saudi', 'earthquak', 'countri', 'he', 'inc', 'would', 'sai', 'iranian', 'refin', 'tanker', 'suppli', 'iraq'

Evaluations

With All Words

- Precision for earn: 0.9943019943019943
- Recall for earn: 0.9721448467966574
- F-Value for earn: 0.9830985915492959
- Precision for acq: 0.9589603283173734
- Recall for acq: 0.9887165021156559
- F-Value for acq: 0.9736111111111111
- Precision for money-fx: 0.9619565217391305
- Recall for money-fx: 0.9943820224719101
- F-Value for money-fx: 0.9779005524861879
- Precision for grain: 0.9931972789115646

- Recall for grain: 0.9931972789115646
- F-Value for grain: 0.9931972789115646
- Precision for crude: 0.9884393063583815
- Recall for crude: 0.9661016949152542
- F-Value for crude: 0.9771428571428571
- **Macroaveraged Precision:** 0.9793710859256889
- **Macroaveraged Recall:** 0.9829084690422084
- **Macroaveraged F-Value:** 0.9811365890801147
- **Microaveraged Precision:** 0.9798951048951049
- **Microaveraged Recall:** 0.9798951048951049
- **Microaveraged F-Value:** 0.9798951048951049

With Most Discriminating Words

- Precision for earn: 0.9551098376313276
- Recall for earn: 0.9285051067780873
- F-Value for earn: 0.9416195856873824
- Precision for acq: 0.8965986394557823
- Recall for acq: 0.9294781382228491
- F-Value for acq: 0.9127423822714682
- Precision for money-fx: 0.946969696969697
- Recall for money-fx: 0.702247191011236
- F-Value for money-fx: 0.8064516129032259
- Precision for grain: 0.9776119402985075
- Recall for grain: 0.891156462585034
- F-Value for grain: 0.9323843416370107

- Precision for crude: 0.7208333333333333
- Recall for crude: 0.9774011299435028
- F-Value for crude: 0.829736211031175
- **Macroaveraged Precision:** 0.8994246895377296
- **Macroaveraged Recall:** 0.8857576057081419
- **Macroaveraged F-Value:** 0.8925388310665532
- **Microaveraged Precision:** 0.9125874125874126
- **Microaveraged Recall:** 0.9125874125874126
- **Microaveraged F-Value:** 0.9125874125874126

Appendix

Figure 1: With All Words

```
$ python3 naive_bayes.py 0
Please wait until calculating evaluations, it may take a while.
Precision for earn: 0.9943019943019943
Recall for earn: 0.9721448467966574
F-Value for earn: 0.9830985915492959
----------------------
Precision for acq: 0.9589603283173734
Recall for acq: 0.9887165021156559
F-Value for acq: 0.9736111111111111
----------------------
Precision for money-fx: 0.9619565217391305
Recall for money-fx: 0.9943820224719101
F-Value for money-fx: 0.9779005524861879
----------------------
Precision for grain: 0.9931972789115646
Recall for grain: 0.9931972789115646
F-Value for grain: 0.9931972789115646
----------------------
Precision for crude: 0.9884393063583815
Recall for crude: 0.9661016949152542
F-Value for crude: 0.9771428571428571
----------------------
Macroaveraged Precision: 0.9793710859256889
Macroaveraged Recall: 0.9829084690422084
Macroaveraged F-Value: 0.9811365890801147
Microaveraged Precision: 0.9798951048951049
Microaveraged Recall: 0.9798951048951049
Microaveraged F-Value: 0.9798951048951049
--- 13.579056978225708 seconds ---
```

Figure 2: With Most Discriminating Words

```
$ python3 naive_bayes.py 1
Please wait until calculating evaluations, it may take a while.
Precision for earn: 0.9551098376313276
Recall for earn: 0.9285051067780873
F-Value for earn: 0.9416195856873824
----------------------
Precision for acq: 0.8965986394557823
Recall for acq: 0.9294781382228491
F-Value for acq: 0.9127423822714682
----------------------
Precision for money-fx: 0.946969696969697
Recall for money-fx: 0.702247191011236
F-Value for money-fx: 0.8064516129032259
----------------------
Precision for grain: 0.9776119402985075
Recall for grain: 0.891156462585034
F-Value for grain: 0.9323843416370107
----------------------
Precision for crude: 0.7208333333333333
Recall for crude: 0.9774011299435028
F-Value for crude: 0.829736211031175
----------------------
Macroaveraged Precision: 0.8994246895377296
Macroaveraged Recall: 0.8857576057081419
Macroaveraged F-Value: 0.8925388310665532
Microaveraged Precision: 0.9125874125874126
Microaveraged Recall: 0.9125874125874126
Microaveraged F-Value: 0.9125874125874126
--- 24.48516011238098 seconds ---
```