Project 2 Report
Karthik Unnikrishnan

Pre-processing:

As a part of pre-procesing, all the given instructions were followed in order and expected output of 8090 documents and 5618 tokens was generated. The following additional preprocessing steps were taken:

- removed the trailing empty tokens that were occurring after the document was split.
- Replaced escape characters for " ' ", "&",and "\" in addition to ">" and "<"

The top 20 generated topics are as follows:

'earn', 'acq', 'crude', 'trade', 'mone-fx', 'interest', 'money-supply', 'ship', 'sugar', 'coffee', 'gold', 'gnp', 'cpi', 'cocoa', 'jobs', 'copper', 'reserves', 'grain', 'alum', 'ipi'

Clustering Performance(SSE):

File: freq

Cluster size	20	40	60
Elapsed time	16382 s	17421 s	14562 s
Entropy	1.1832	0.8575	0.7402
Purity	0.7456	0.8202	0.8393

File: log2freq

Cluster size	20	40	60
Elapsed time	17294 s	14653 s	13085 s
Entropy	1.0163	0.7368	0.6405
Purity	0.7822	0.8360	0.8618

File: sqrtfreq

Cluster size	20	40	60
Elapsed time	13874 s	14576 s	15612 s
Entropy	0.8267	0.6142	0.6183
Purity	0.8187	0.8741	0.8601

Clustering Performance(I2):

File: freq Criterion: I2

Cluster size	20	40	60
Elapsed time	13228 s	12986 s	13788 s
Entropy	1.1167	0.8505	0.7931
Purity	0.7494	0.8270	0.8338

File: log2freq Criterion: I2

Cluster size	20	40	60
Elapsed time	14836 s	13944 s	12844 s
Entropy	0.9633	0.6540	0.6173
Purity	0.7888	0.8709	0.8773

File: sqrtfreq Criterion: I2

Cluster size	20	40	60
Elapsed time	14776 s	12842 s	14882 s
Entropy	-0.8405	0.59097	0.5566
Purity	0.8250	0.8876	0.8859

Clustering Solution

Cluster size : 20 Criterion : SSE File: freq

Cluster 1:

Term	Count
gold	1
acq	7
earn	6
crude	258
cpi	1
trade	1
Total	274

Cluster 2:

Term	Count
earn	224
Total	224

Cluster 3:

Term	Count
alum	1
acq	1
ship	1
gnp	3
Reserves	32
money-fx	7
interest	2
earn	117
crude	3
money-supply	36
trade	33
Total	236

Cluster 4:

Term	Count
alum	1
money-fx	11
reserves	1
copper	1
grain	4
earn	31
sugar	6
crude	11

trade	4
gold	3
acq	24
gnp	53
ship	3
interest	57
ipi	37
jobs	32
money-supply	35
cpi	61
Total	375

Cluster 5:

Term	Count
earn	496
Total	496

Cluster 6:

Term	Count
gold	2
acq	504
earn	63
crude	1
Total	570

Cluster 7:

Term	Count
gold	2
acq	504
earn	63
crude	1
Total	365

Cluster 8:

Term	Count
coffee	104
alum	33
money-fx	88
grain	35
copper	39
earn	24
sugar	118
crude	48
trade	25
gold	62
acq	336
ship	126
gnp	10
interest	20
ipi	3
jobs	15
cocoa	52
money-supply	4
cpi	5
Total	1147

Cluster 9:

Term	Count
earn	616
Total	616

Cluster 10:

Term	Count
Acq	12
gnp	1
copper	1
earn	306
crude	2
Total	322

Cluster 11:

Term	Count
acq	1
earn	358
Total	359

Cluster 12:

Term	Count
Coffee	2
acq	7
ship	2
money-fx	49
grain	3
interest	1
earn	148

sugar	5
crude	9
cocoa	1
trade	2
money-supply	1
Total	230

Cluster 13:

Term	Count
acq	1
earn	197
Total	198

Cluster 14:

Term	Count
acq	182
earn	1
Total	183

Cluster 15:

Term	Count
coffee	3
alum	5
reserves	4
Money-fx	3
copper	5
grain	1
earn	78
sugar	1
crude	3

trade	9
gold	10
acq	415
ship	10
gnp	2
interest	3
cocoa	2
money-supply	10
Total	570

Cluster 16:

Term	Count
acq	1
earn	234
Total	235

Cluster 17:

Term	Count
gold	16
acq	565
alum	3
ship	12
copper	8
grain	1
earn	99
sugar	2
crude	14
trade	3
Total	723

Cluster 18:

Term	Count
coffee	2
alum	2
money-fx	17
grain	1
crude	3
trade	3
gold	253
acq	4
gnp	8
ship	1
interest	1
ipi	1
Total	297

Cluster 19:

Term	Count
earn	344
Total	344

Cluster 20:

Term	Count
coffee	3
reserves	4
money-fx	84
earn	29
crude	3
trade	3
gold	1
acq	61

ship	3
interest	1
jobs	137
money-supply	1
cpi	11
Total	332