Applied probability models for CS: Exercise 3

Daniel Bazar 314708181

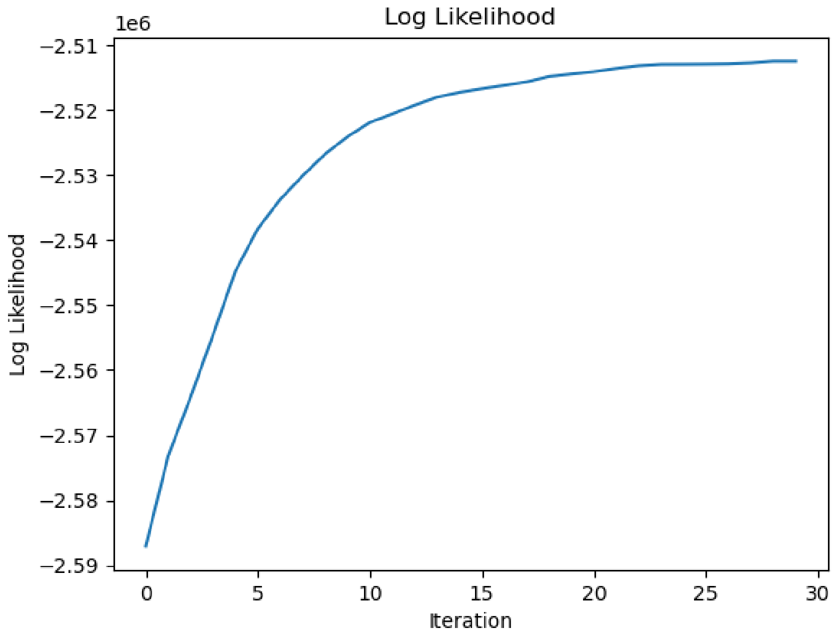
Lior Krengel 315850594

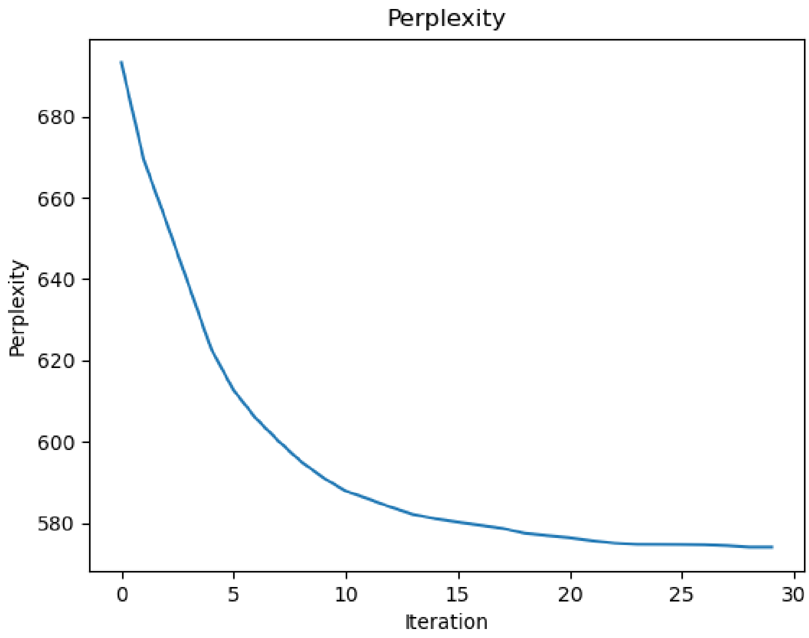
Part 1

We decided the threshold for stopping criteria in the EM algorithm to be **10.** In scale of the large numbers of the log-likelihood, 10 is a very small number. We tested many thresholds and there weren’t big differences.

Part 2

Log-likelihood and mean perplexity plots until convergence (30 iterations):





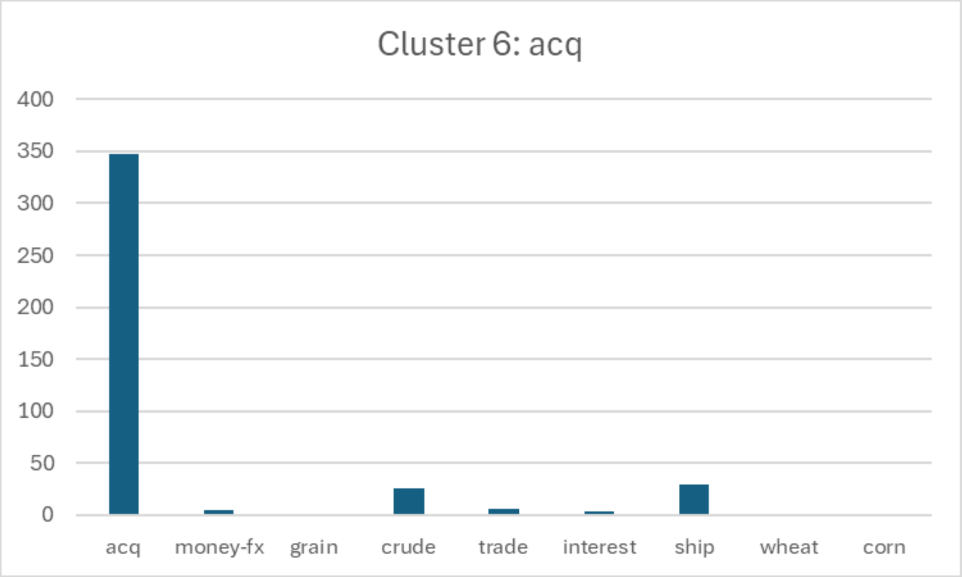
Part 3

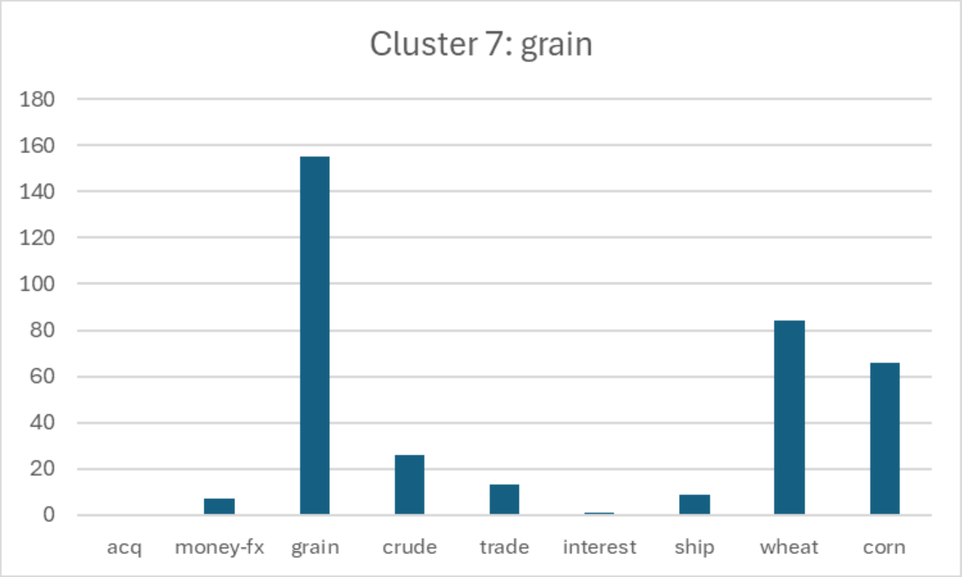
Confusion matrix:

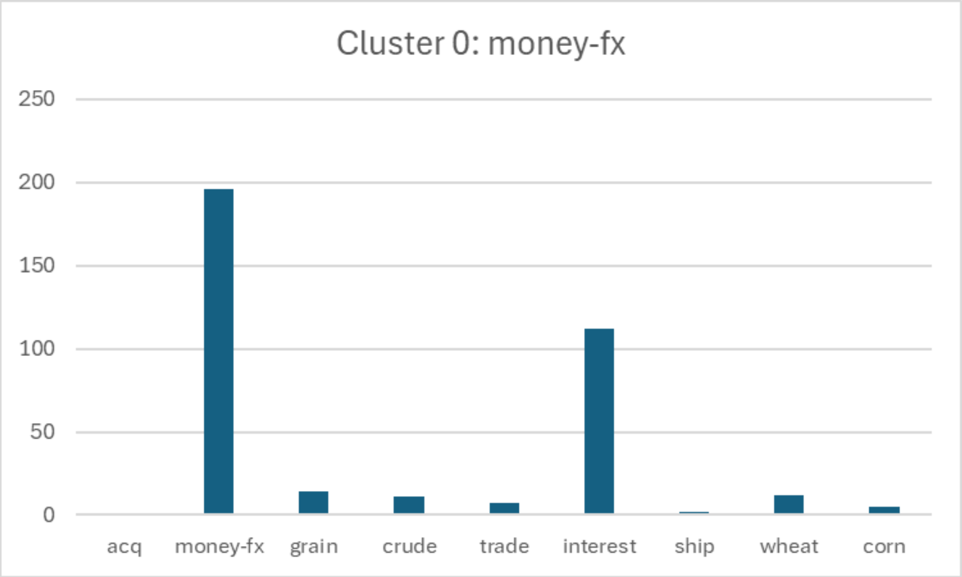
	acq	money-fx	grain	crude	trade	interest	ship	wheat	corn	size
6	347	5	1	26	6	3	29	1	0	418
7	0	7	155	26	13	1	9	84	66	361
0	0	196	14	11	7	112	2	12	5	359
5	7	82	12	32	77	115	3	5	5	338
2	7	7	98	44	14	12	21	60	38	301
8	9	44	16	15	193	7	2	3	7	296
3	2	87	55	16	17	15	18	19	33	262
1	8	13	29	51	5	4	96	10	4	220
4	20	10	8	128	5	6	6	3	2	188

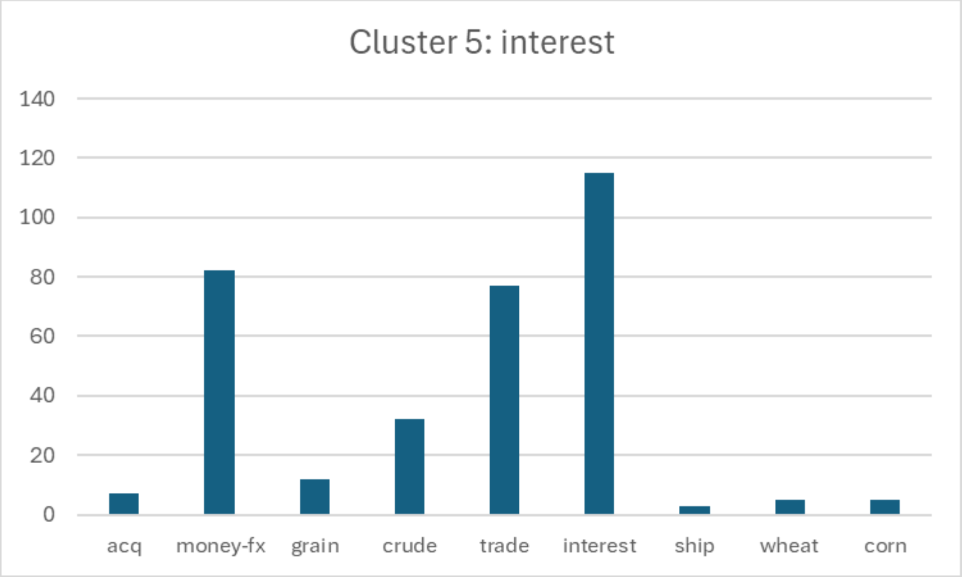
Part 4

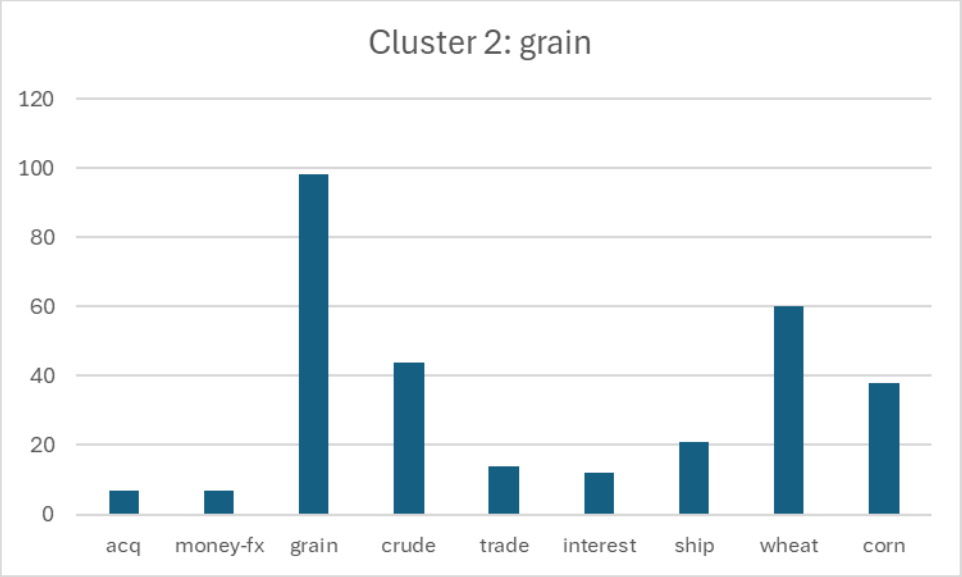
Topics histograms:

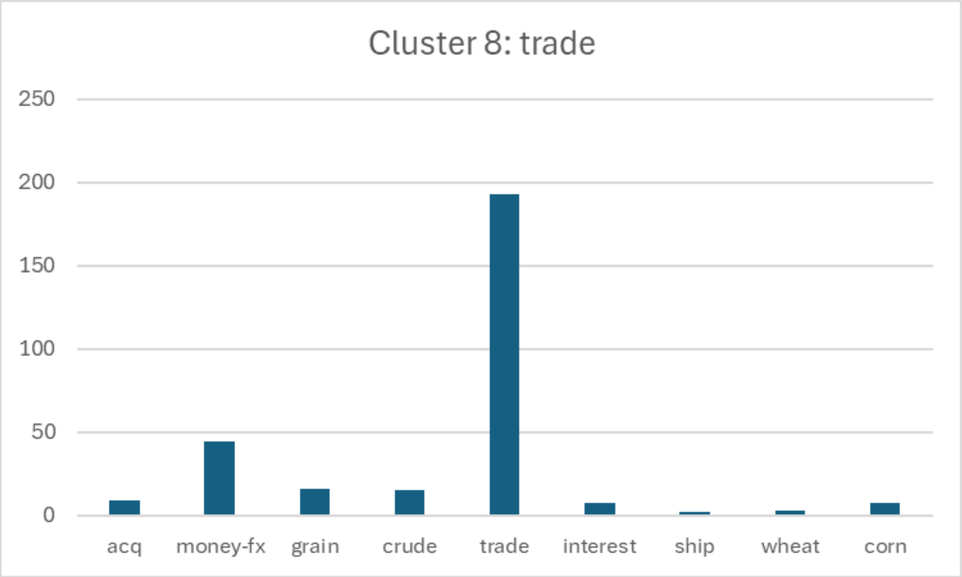


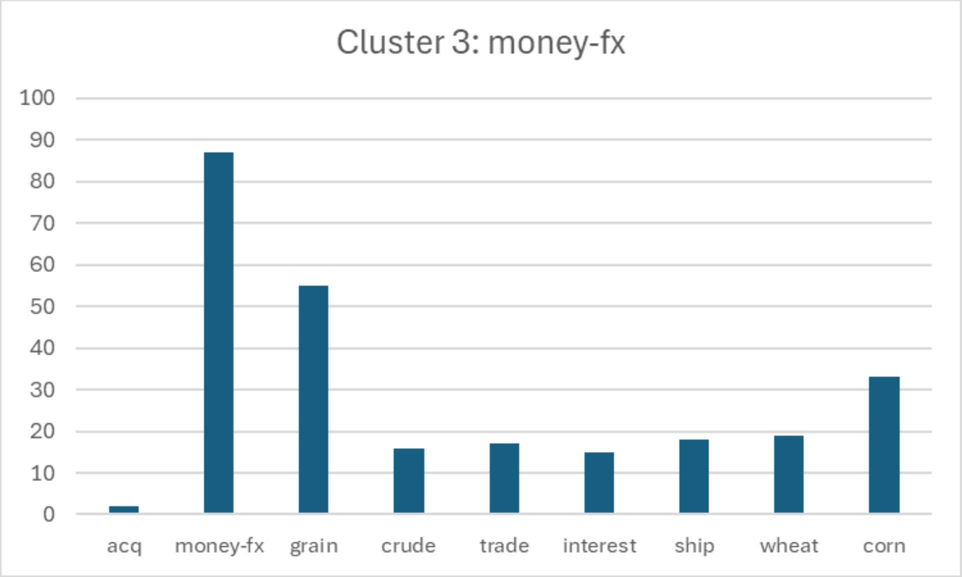

Cluster 7: grain
180
160
140
120
100
80
60
40
20
0
acq
money-fx
grain
crude
trade
interest
ship
wheat
corn

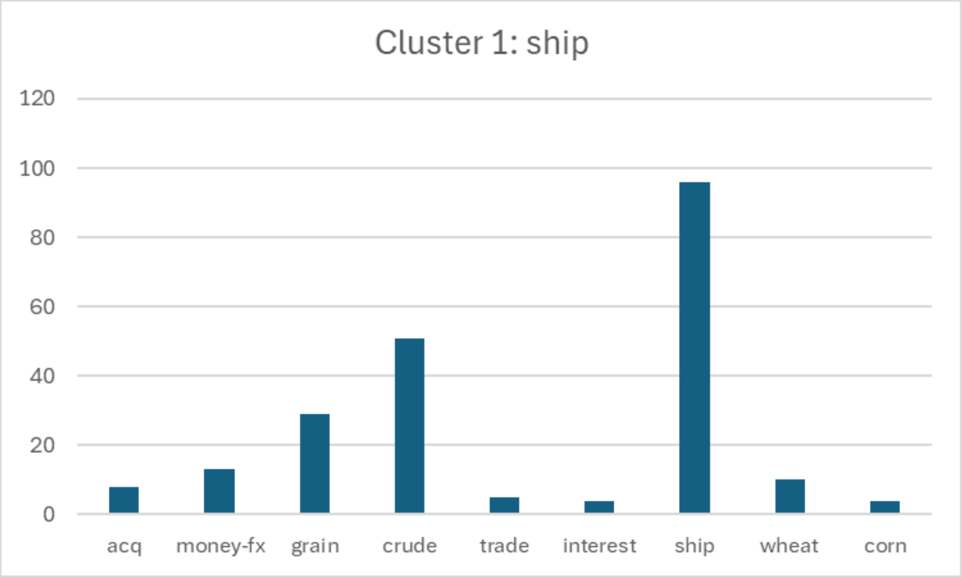

Cluster 0: money-fx
250
200
150
100
50
0
acq
money-fx
grain
crude
trade
interest
ship
wheat
corn

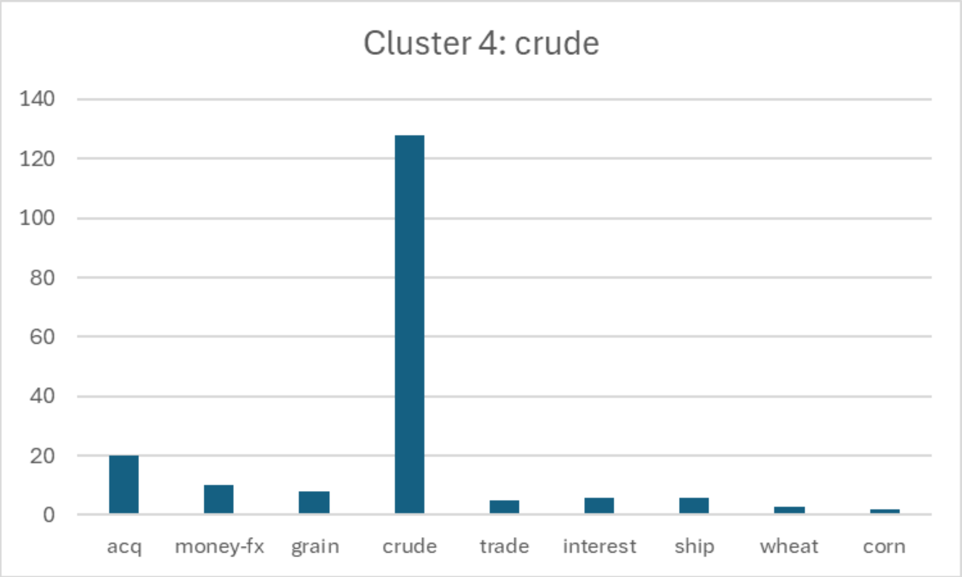

Cluster 5: interest
140
120
100
80
60
40
20
0
acq
money-fx
grain
crude
trade
interest
ship
wheat
corn


Cluster 2: grain
120
100
80
60
40
20
0
acq
money-fx
grain
crude
trade
interest
ship
wheat
corn


Cluster 8: trade
250
200
150
100
50
0
acq
money-fx
grain
crude
trade
interest
ship
wheat
corn


Cluster 3: money-fx
100
90
80
70
60
50
40
30
20
10
0
acq
money-fx
grain
crude
trade
interest
ship
wheat
corn





Part 5

Model accuracy: **66.62%**

Part 6

We tested many lambdas values and λ=0.1 returned the minimum perplexity so we chose it

K	10
vocabulary size after filtering	6800
λ	0.1