Applied probability models for CS – Exercise 3 Final Report

Ofri Kleinfeld 302893608 Shai Keynan 301687273

1) EM Algorithm Tuning –

Our threshold for the algorithm's stopping criterion was one of the following:

- Likelihood didn't increase compared to last iteration
- Reaching a total cap of 150 iterations

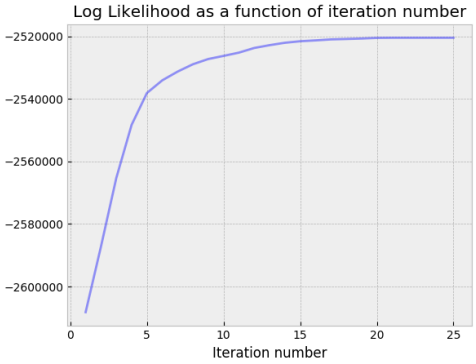
Eventually, the algorithm stopped after 24 iterations, reaching a log-likelihood value of -2,520,046 and a perplexity value of 585.73.

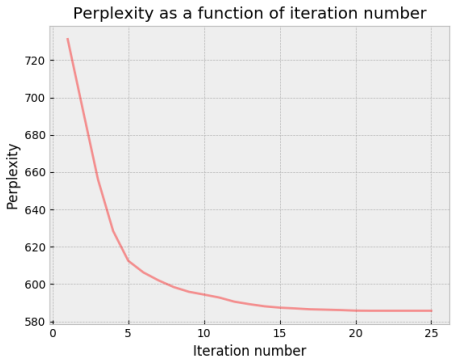
After experimenting few options of different lambda and K values, our final version of the algorithm used lambda=1 and K=10.

Also, our vocabulary size, considering only frequent-enough words is 6800 unique words.

2) Log Likelihood and Perplexity as a function of iteration number –

Indeed, we are able to see that the log likelihood does not decrease after each iteration, and as a result the perplexity also does not increase after each iteration.



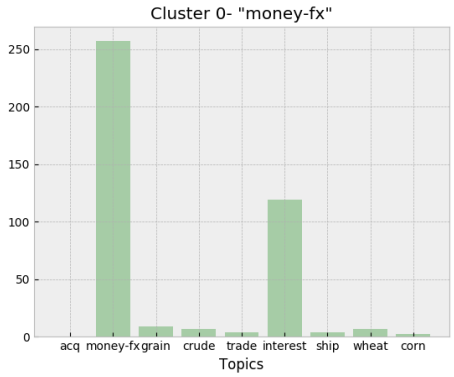


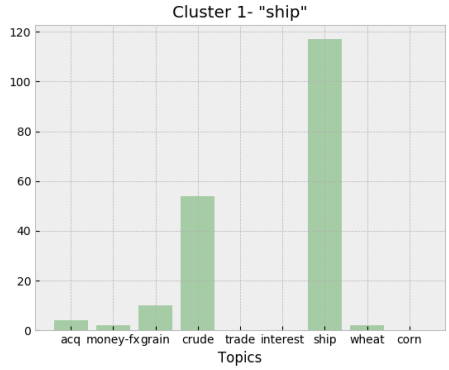
3) Confusion Matrix and Clusters Histograms–

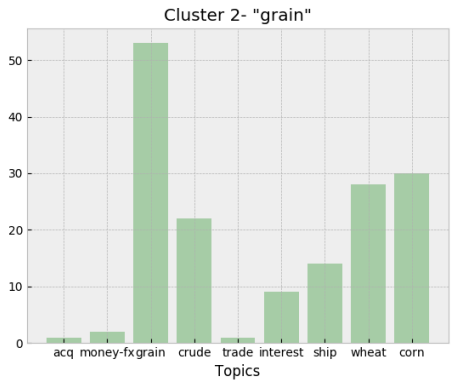
Using the given labels, we were able a create a confusion matrix and a histogram of each cluster - examining the different topic assigned to each cluster.

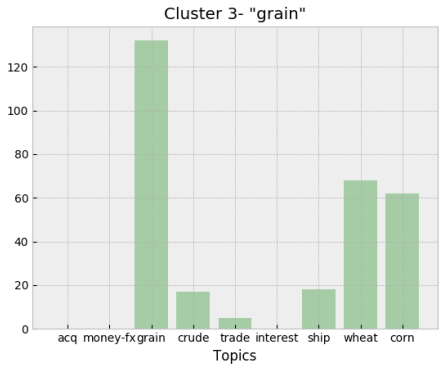
Cluster number / Topic	acq	money-fx	grain	crude	trade	interest	ship	wheat	corn	total_in_cluster
0	0	257	9	7	4	119	4	7	2	**409**
1	4	2	10	54	0	0	117	2	0	**189**
2	1	2	53	22	1	9	14	28	30	**160**
3	0	0	132	17	5	0	18	68	62	**302**
4	14	4	0	199	2	1	1	0	0	**221**
5	2	93	1	10	85	126	2	0	0	**319**
6	369	2	2	21	1	2	15	0	2	**414**
7	1	2	168	6	11	0	14	89	59	**350**
8	9	89	13	13	228	18	1	3	5	**379**

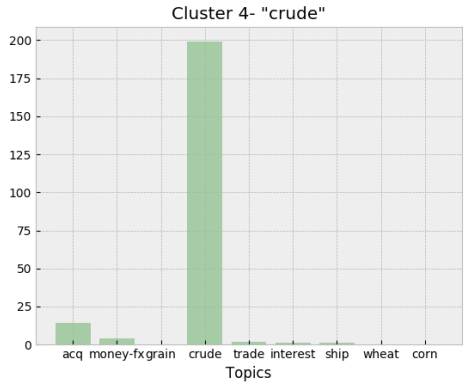
Cells in yellow represents most frequent topic, in each cluster.

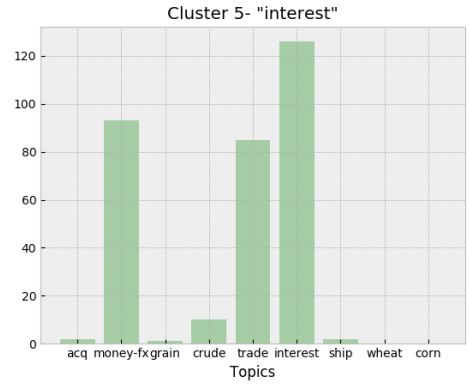


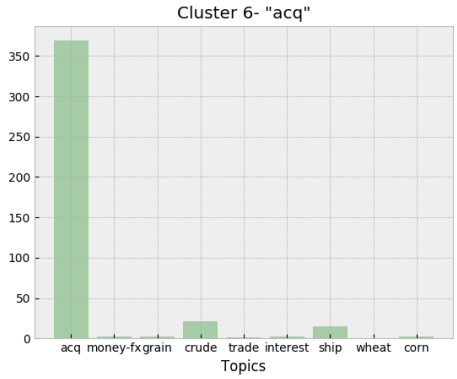


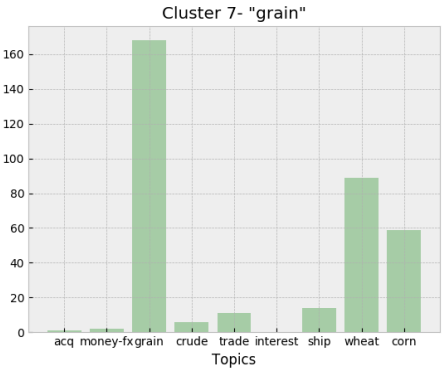


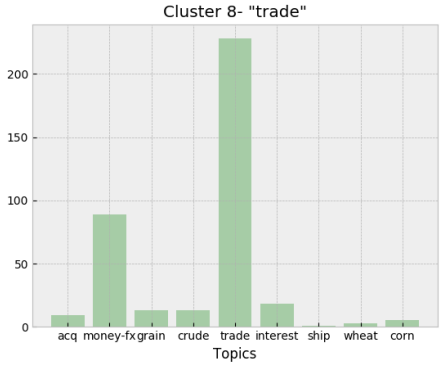












4) Accuracy of Classification–

Using most frequent/dominant topic of each cluster as it's class, we are to present the model's accuracy – 77.63% (considering an assignment as correct if the assigned topic is one of its topics)