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:

- Encountering 5 consecutive iterations without likelihood improvement
- Reaching a total cap of 150 iterations

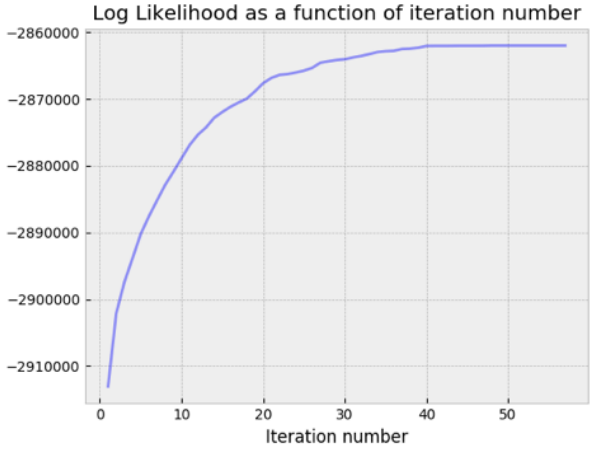
Eventually, the algorithm stopped after 57 iterations, reaching a log-likelihood value of -2,861,996.17 and a perplexity value of 986.25.

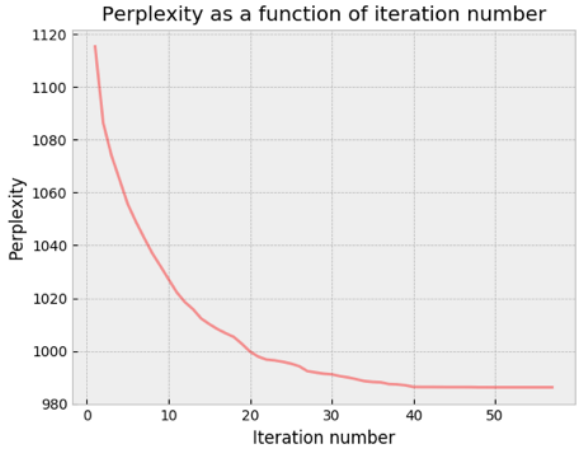
After experimenting few options of different lambda and K values, our final version of the algorithm used lambda=0.01 and K=10.

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.



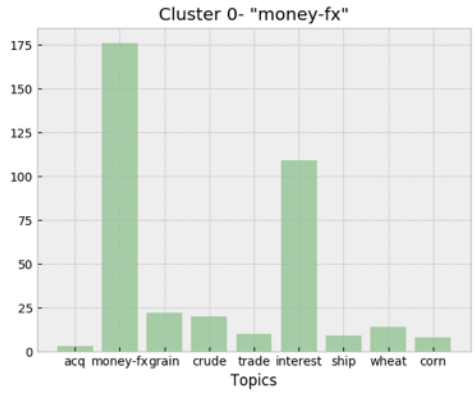


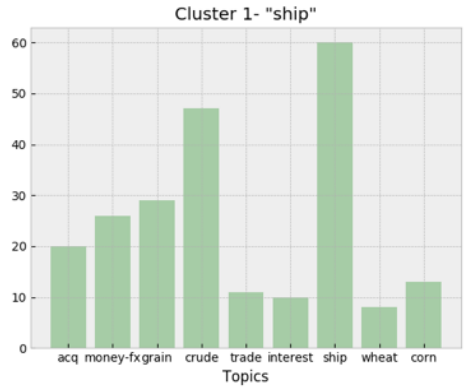
3) Confusion Matrix and Clusters Histograms–

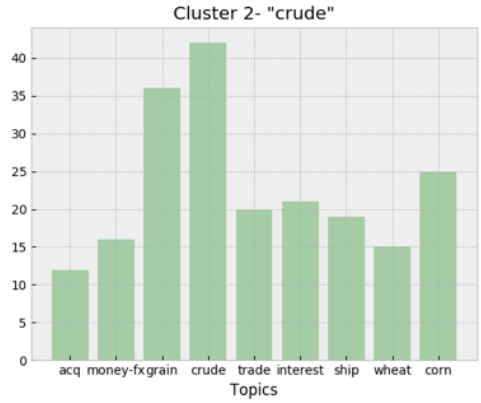
Using the given labels, we were able a create a confusion matrix and a histogram of each cluster - examining the different topic assigned to each cluster.

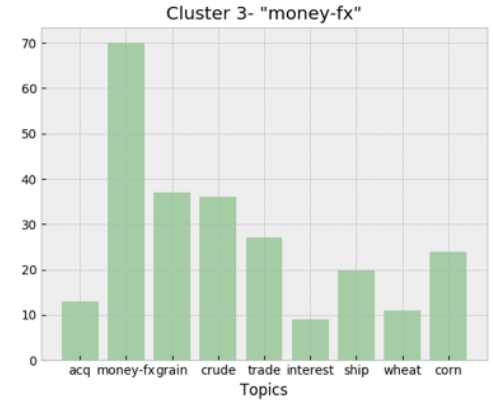
Culster Number / Topic	acq	money-fx	grain	crude	trade	interest	ship	wheat	corn	total_in_cluster
0	3	176	22	20	10	109	9	14	8	**371**
1	20	26	29	47	11	10	60	8	13	**224**
2	12	16	36	42	20	21	19	15	25	**206**
3	13	70	37	36	27	9	20	11	24	**247**
4	27	22	24	70	14	12	11	12	5	**197**
5	12	78	20	37	74	104	5	8	10	**348**
6	299	12	8	52	7	0	43	5	3	**429**
7	5	12	175	26	14	6	12	112	58	**420**
8	9	39	37	19	160	4	7	12	14	**301**

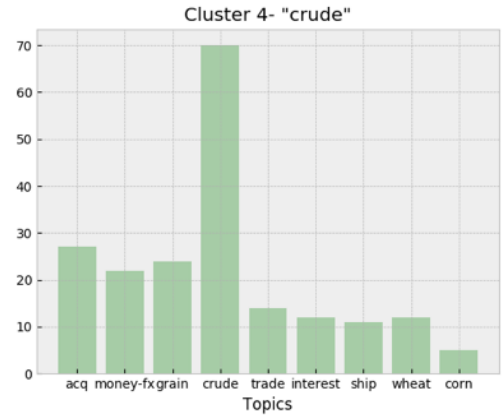
Cells in yellow represents most frequent topic, in each cluster.

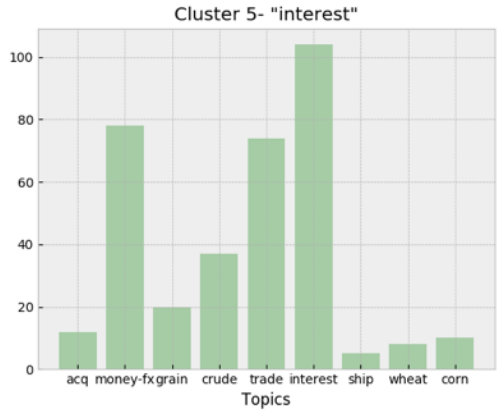


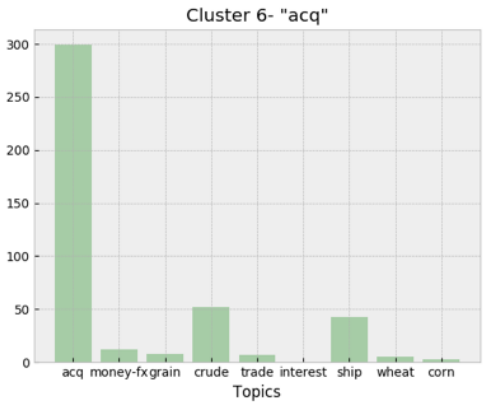


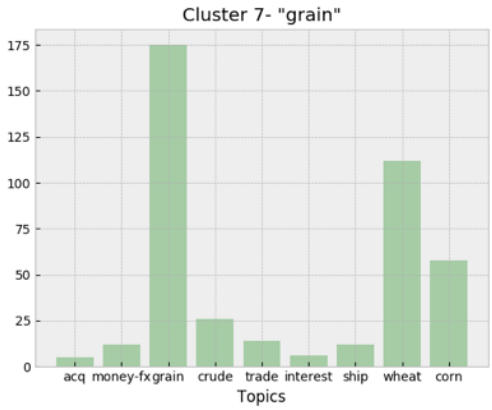


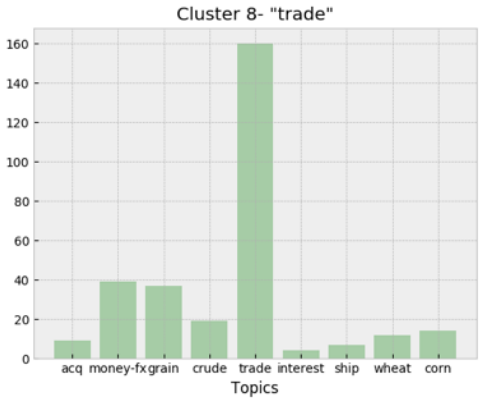












4) Accuracy of Classification–

Using most frequent/dominant topic of each cluster as it's class, we are to present the model's accuracy – 54.42% (considering an assignment as correct if the assigned topic is one of it's topics)