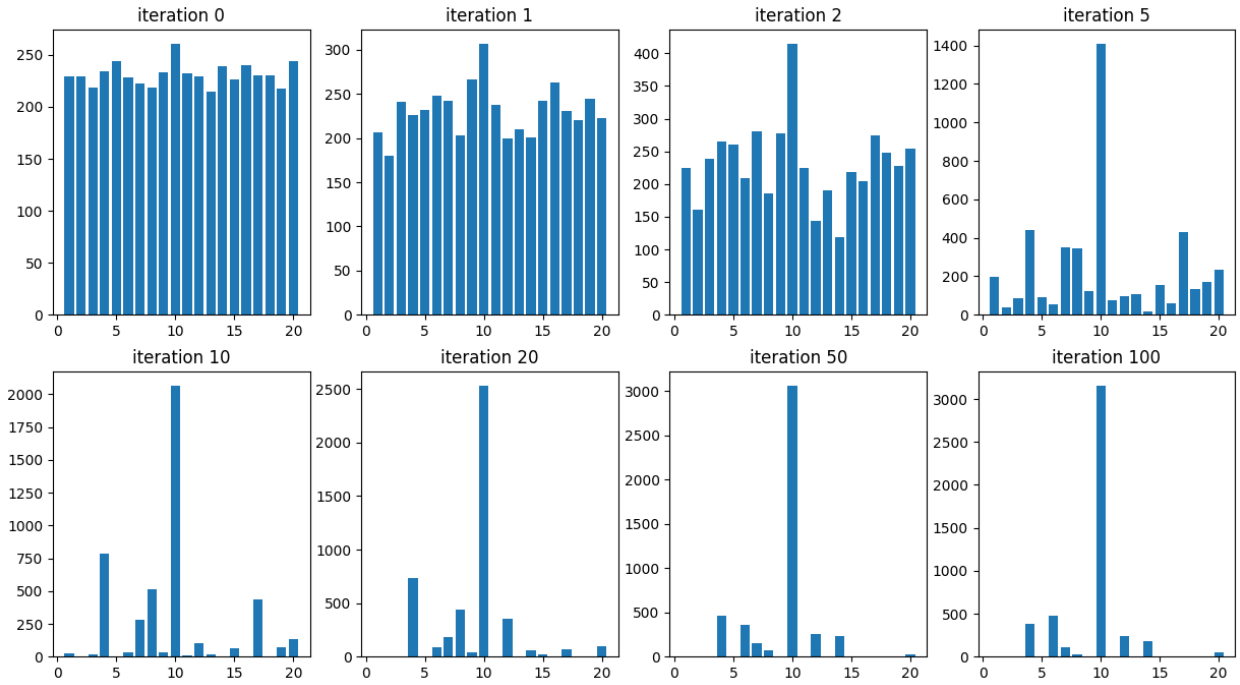
LDA Report

Task#1.

Resulted counts are saved in files cw.csv, ck.csv.

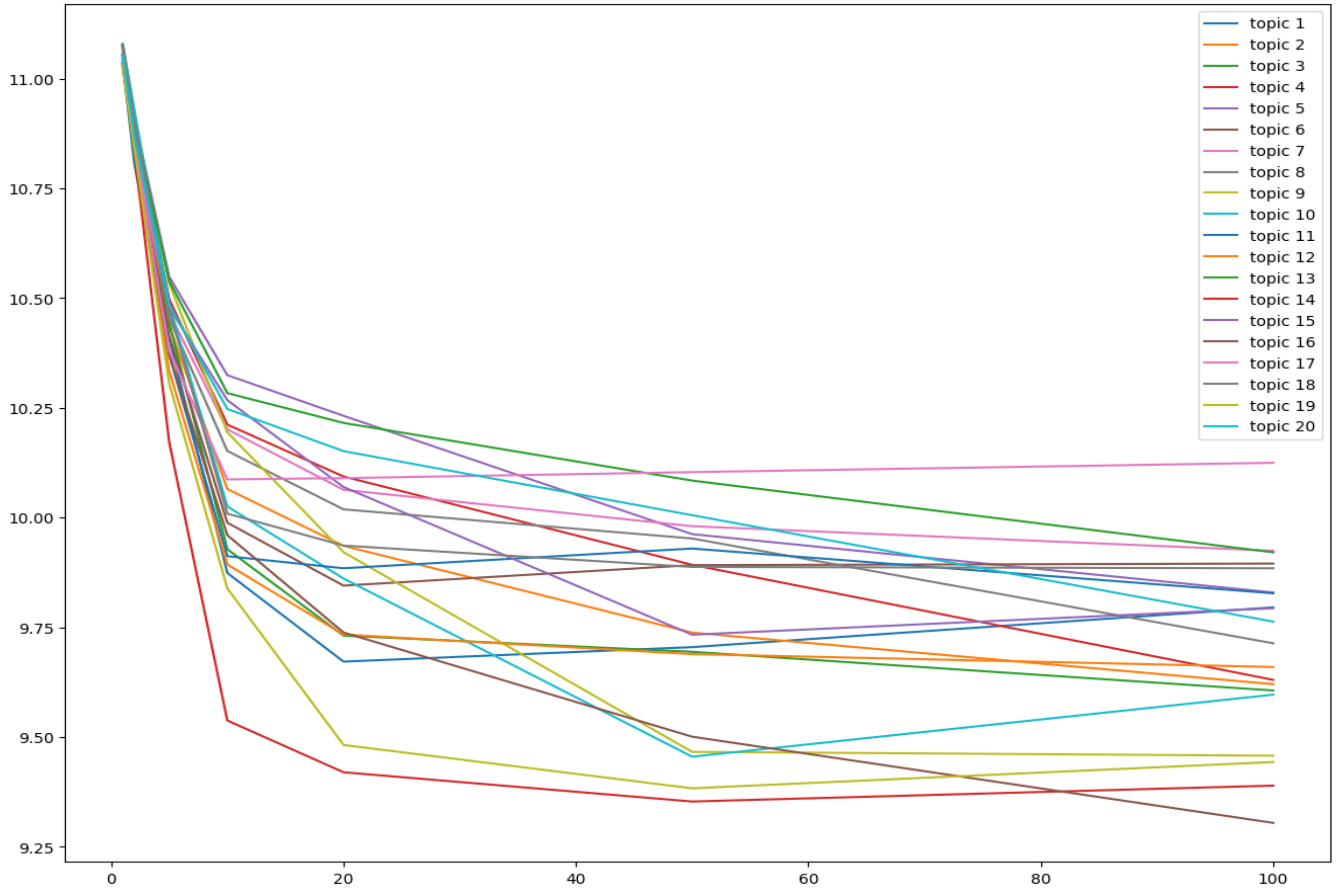
The model's per topic perplexity is [888.72, 787.19, 779.39, 792.53, 909.82, 952.36, 971.72, 839.68, 703.09, 774.24, 908.69, 808.84, 969.41, 670.67, 887.45, 632.34, 1116.62, 945.27, 696.13, 868.94]

Task #2.



As it is shown on the graph above, at the beginning, the distribution over topics was close to uniform and during the training, it was moving more and more to one dominated topic for the current document.

Task #3.



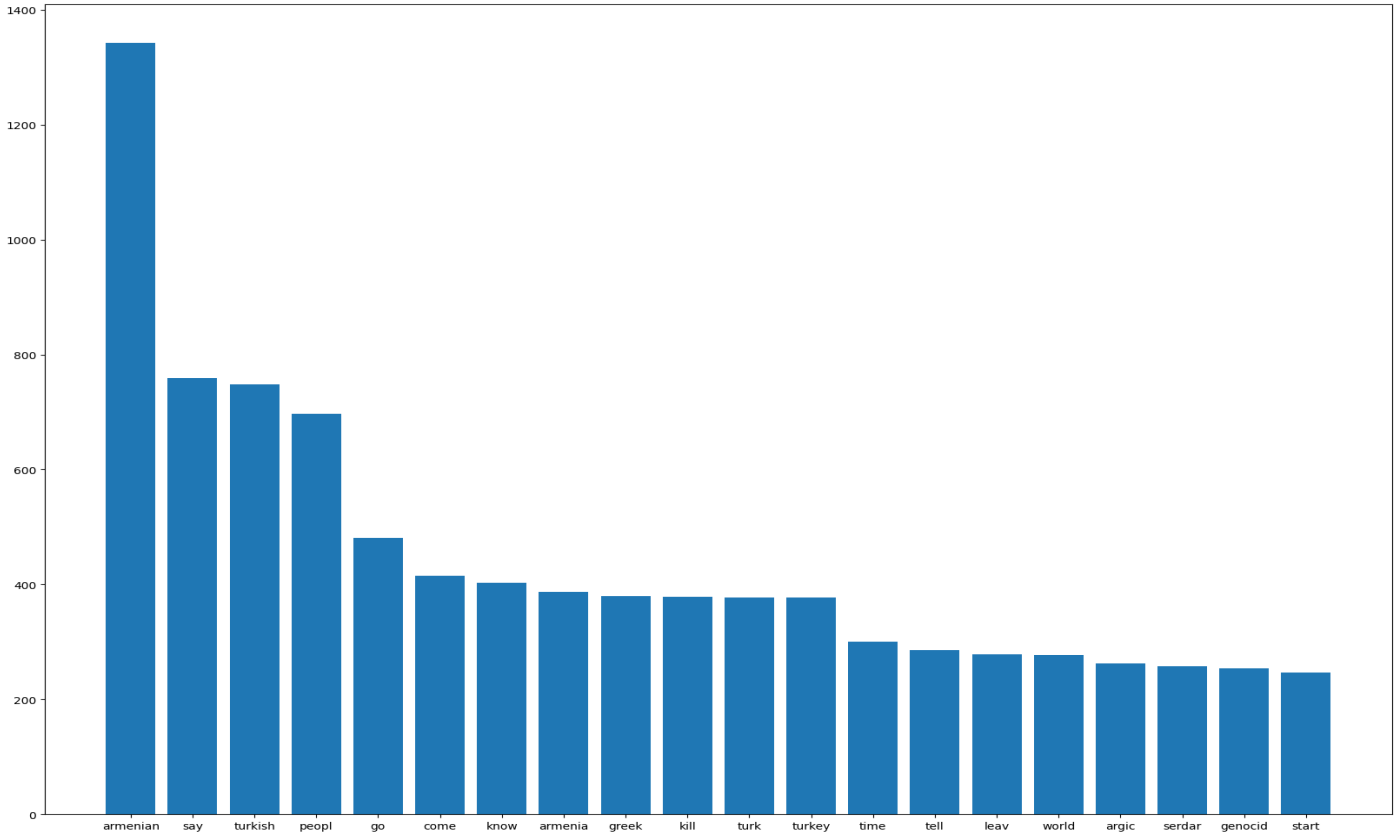
Task #4.

Most frequent words for topic 2: [armenian, say, turkish, people, go, come, know, armenia, greek, kill, turk, turkey, time, tell, leav, world, argic, serdar, genocide, start]

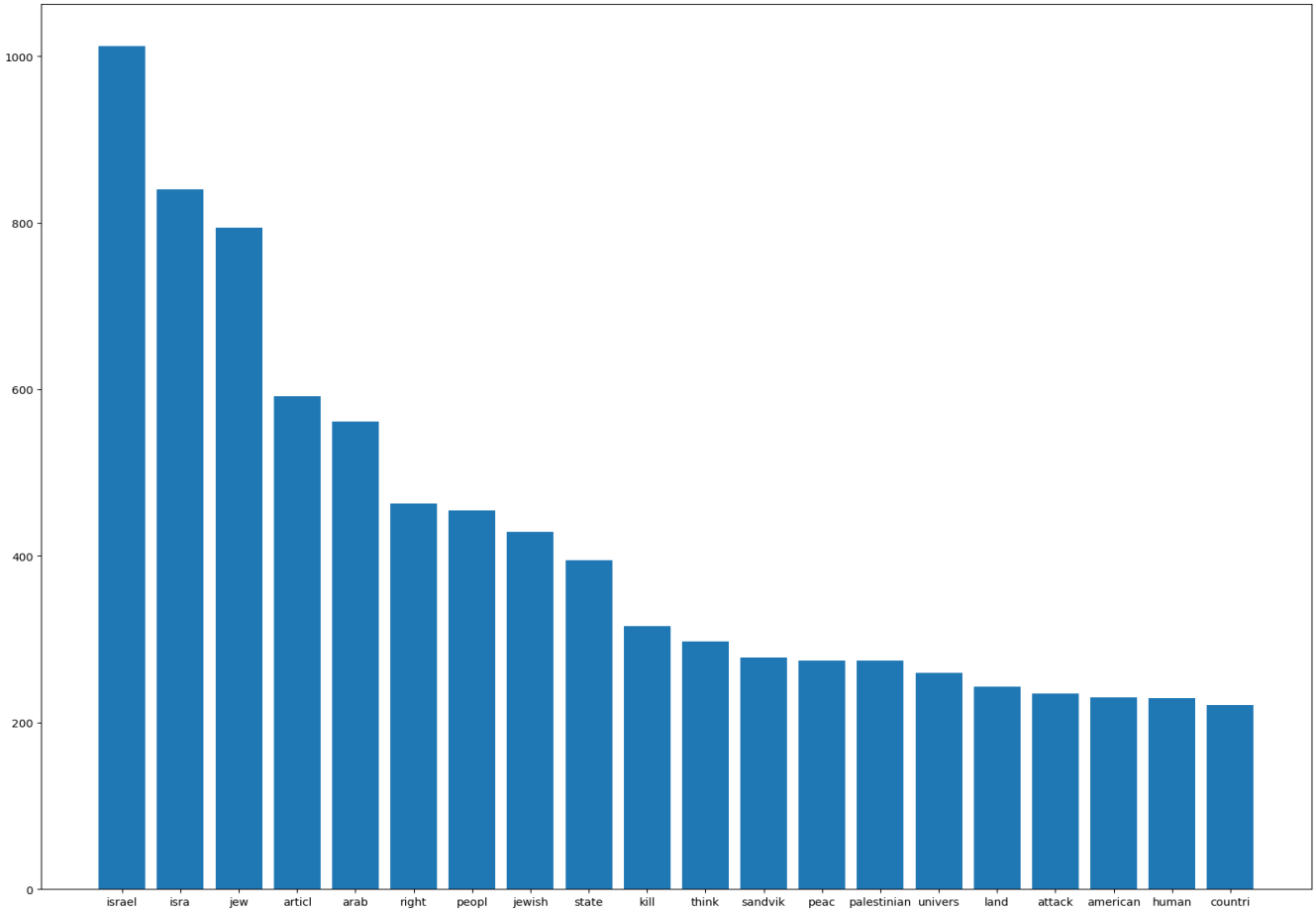
Most frequent words for topic 10: [israel, isra, jew, articl, arab, right, peopl, jewish, state, kill, think, sandvik, peac, palestinian, univers, land, attack, american, human, countri]

Most frequent words for topic 16: [articl, bike, like, drive, look, nntp, host, good, car, engin, speed, think, time, distribut, want, know, price, need, univers, sell]

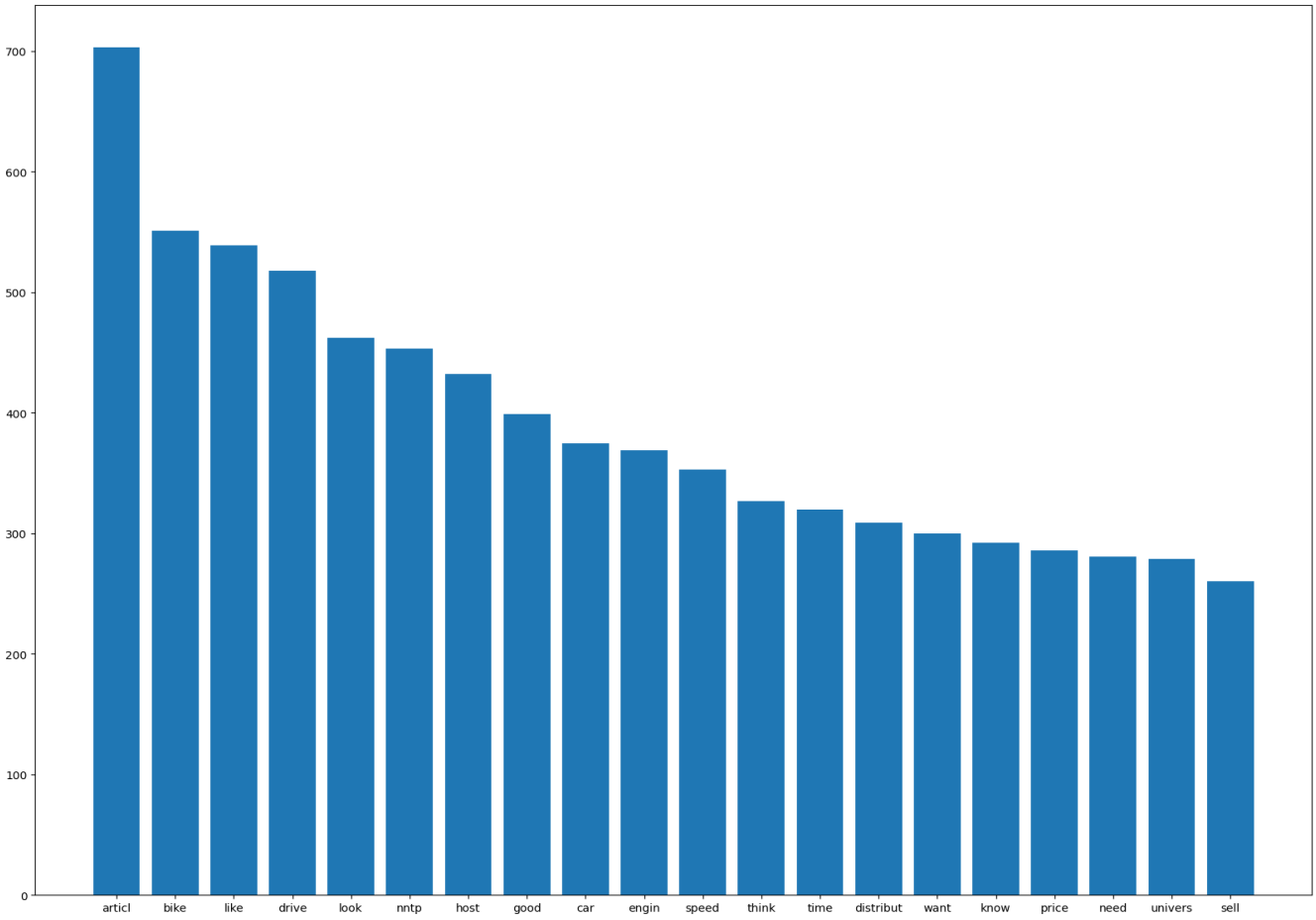
Topic 2.



Topic 10.



Topic 16.



I chose topics 2, 10, and 16, which shown on graphs above. Looks like the topic 2 is about Armenian-Turkish conflict and Armenian genocide by Turkey. Topic 10 is about Israel-Arabian conflict. Topic 16 is about some vehicles and their selling.

Task #5.

Test perplexity is 1452.50.

Simple model perplexity is 2627.03, which is more than for the LDA model.

Task #6.

#Model perplexity per topic - 50 iterations, k=25, alpha= 0.01, gamma = 0.5

[1441.70, 1691.61, 1180.42, 1499.38, 1257.36, 1540.48 1201.54, 1670.07, 975.40, 1655.83, 1350.40, 1162.24, 1409.34, 1518.36, 1700.41, 1840.55, 1693.16, 1437.16, 1804.08, 1850.70, 1721.66, 1270.36, 1787.18, 1386.01, 1879.16]

#Model perplexity per topic - 50 iterations, k=15, alpha= 0.05, gamma = 0.1

[957.51, 1320.80, 778.70, 1102.61, 962.80, 1197.39, 1334.59, 1105.86, 1089.39, 1048.82, 1070.33, 882.79, 968.81, 1212.58, 808.33]

So, looks like the original model has the best model perplexity.