Naïve-Bayes Classifier Analysis:

Naïve-Bayes Classifier is a method used in Machine Learning with supervised learning. The way Naïve-Bayes Classifier works is that we use labeled data to train the classifier. In this assignment, we have a dataset of comment with the label negative or positive. Theoretically, we would train the classifier with vocabulary and labels. The algorithm would classify the appearance of certain vocabularies based on the label to generate a conditional probability. And use this probability to determine the label of any dataset given in the future. Since Naïve-Bayes Classifier is based on probabilities, the results can not be 100% accurate.

From our python code, we have generated a text file listing the label and corresponding line number. By comparing the original file and the generated file, we find out that the number of negative labels from the original file is 1282, and for positive label is 1101; number of negative labels from the generated file is1239, and for positive label is 1153, and the index of line are listed below:

[10, 11, 13, 15, 16, 18, 26, 28, 34, 39, 40, 43, 51, 54, 55, 58, 59, 62, 64, 70, 80, 91, 93, 98, 99, 103, 111, 114, 116, 122, 125, 127, 130, 135, 151, 159, 162, 164, 170, 173, 174, 175, 181, 193, 227, 232, 237, 242, 265, 273, 285, 289, 301, 305, 306, 309, 310, 318, 319, 323, 326, 328, 336, 343, 346, 351, 352, 356, 358, 360, 367, 373, 374, 379, 390, 405, 413, 415, 417, 424, 434, 437, 448, 449, 469, 471, 473, 475, 481, 484, 491, 496, 498, 505, 512, 519, 522, 525, 530, 533, 542, 546, 547, 550, 553, 560, 562, 569, 571, 573, 580, 583, 595, 597, 601, 609, 611, 618, 631, 646, 647, 650, 655, 659, 665, 667, 670, 674, 677, 678, 679, 719, 724, 731, 734, 736, 741, 743, 746, 747, 748, 773, 778, 783, 789, 795, 800, 801, 802, 805, 807, 811, 814, 818, 824, 825, 833, 839, 842, 847, 848, 850, 855, 862, 863, 864, 873, 875, 876, 877, 878, 880, 894, 903, 904, 905, 913, 914, 928, 939, 950, 951, 957, 962, 970, 975, 976, 985, 1016, 1021, 1029, 1034, 1048, 1057, 1064, 1079, 1082, 1084, 1092, 1093, 1104, 1113, 1115, 1129, 1138, 1142, 1148, 1150, 1152, 1160, 1165, 1168, 1170, 1180, 1192, 1193, 1220, 1223, 1226, 1229, 1230, 1233, 1236, 1238, 1239, 1257, 1258, 1259, 1285, 1288, 1294, 1295, 1304, 1310, 1313, 1333, 1334, 1339, 1340, 1344, 1347, 1356, 1360, 1363, 1370, 1374, 1379, 1380, 1381, 1384, 1385, 1392, 1396, 1397, 1398, 1413, 1414, 1420, 1431, 1443, 1460, 1479, 1480, 1481, 1484, 1490, 1492, 1496, 1501, 1508, 1512, 1513, 1518, 1519, 1529, 1532, 1533, 1537, 1539, 1548, 1550, 1552, 1555, 1559, 1568, 1570, 1585, 1588, 1608, 1609, 1612, 1614, 1616, 1617, 1620, 1622, 1627, 1628, 1634, 1651, 1653, 1664, 1665, 1683, 1684, 1685, 1686, 1698, 1703, 1704, 1715, 1718, 1720, 1725, 1728, 1731, 1741, 1742, 1757, 1771, 1781, 1782, 1787, 1792, 1796, 1797, 1807, 1814, 1825, 1826, 1831, 1835, 1837, 1846, 1848, 1857, 1865, 1872, 1876, 1883, 1890, 1904, 1915, 1916, 1919, 1920, 1922, 1935, 1940, 1941, 1947, 1949, 1964, 1980, 1981, 1984, 1985, 1986, 1990, 1993, 1996, 2003, 2010, 2018, 2023, 2031, 2035, 2047, 2051, 2059, 2062, 2065, 2070, 2071, 2073, 2080, 2081, 2086, 2089, 2090, 2094, 2096, 2099, 2100, 2102, 2105, 2107, 2110, 2112, 2114, 2123, 2133, 2135, 2136, 2139, 2150, 2160, 2164, 2167, 2168, 2170, 2182, 2183, 2190, 2193, 2197, 2204, 2211, 2225, 2227, 2229, 2233, 2236, 2243, 2247, 2248, 2255, 2263, 2266, 2271, 2273, 2281, 2283, 2295, 2296, 2298, 2299, 2313, 2323, 2332, 2341, 2347, 2348, 2357, 2368, 2372, 2373, 2378, 2379, 2380]

We can see that Naïve-Bayes Classification is not 100% accurate. In case of a label difference between the original file and the generated file, the cause might be that some appearance of certain words in the instance make the classifier to think that it belongs to the opposite label. For example, line #10 state: *“i agree with other reviewers that it feels good and does n't smell too much, however , i 've experimented with it several times to confirm my findings , and it turns out to give me really bad blackheads . i 'm 25 with an oily t-zone and very dry facial skin . on mornings after using this cream, i have nasty blackheads on my forehead and chin . there are better products out there”.* By reading the whole instance we know that this is a negative review, but Naïve-Bayes classifier only look at each word in this instance, here we have the appearance of the word *“good”.* The word good should appear in a lot of positive reviews so that it makes the classifier think that this is a positive review rather than a negative review. Another example, one the line #880, it states: *“i am having the same problem . i ordered this product from amazon on august 26. it is now october 26. i have not recieved it yet . i have never had this problem with amazon before . i may have to cancel my order too and go somewhere else . it is a shame , because amazon does have the best price i have found so far . i gave it 5 stars only because i had to give it a rating. i wanted to be fair to the company , but i have honestly never used one”*. We can see that the reviewer was unhappy about the delivery, but he gives a good review of the product. The reason that the classifier senses negative label for the review is that there are vocabularies such as “problem”, “shame”, make the classifier decides that this is a negative review.