Importing the libraries

In [1]:

```
import pandas as pd
import numpy as np
```

In [2]:

```
data = pd.read_csv("Reviews.csv")
```

In [3]:

```
data.head()
```

Out[3]:

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	
0	1	B001E4KFG0	A3SGXH7AUHU8GW	delmartian	1	1	5	130
1	2	B00813GRG4	A1D87F6ZCVE5NK	dll pa	0	0	1	134
2	3	B000LQOCH0	ABXLMWJIXXAIN	Natalia Corres "Natalia Corres"	1	1	4	121
3	4	B000UA0QIQ	A395BORC6FGVXV	Karl	3	3	2	130
4	5	B006K2ZZ7K	A1UQRSCLF8GW1T	Michael D. Bigham "M. Wassir"	0	0	5	135

In [4]:

```
data.columns
```

Out[4]:

```
Index(['Id', 'ProductId', 'UserId', 'ProfileName', 'HelpfulnessNumerator',
       'HelpfulnessDenominator', 'Score', 'Time', 'Summary', 'Text'],
      dtype='object')
```

In [5]:

```
data["Helpful%"] = np.where(data["HelpfulnessDenominator"]>0,data["HelpfulnessNumerator"
```

In [6]:

```
data.head()
```

Out[6]:

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	
0	1	B001E4KFG0	A3SGXH7AUHU8GW	delmartian	1	1	5	130

1	2	B00813GRG4	A1D87F6ZCVE5NK	dll pa	0	0	1	134
2	3	B000LQOCH0	ABXLMWJIXXAIN	Natalia Corres "Natalia Corres"	1	1	4	121
3	4	B000UA0QIQ	A395BORC6FGVXV	Karl	3	3	2	130
4	5	B006K2ZZ7K	A1UQRSCLF8GW1T	Michael D. Bigham "M. Wassir"	0	0	5	135

In [7]:

```
data["Helpful%"].unique()
```

Out[7]:

```
array([ 1.        , -1.        ,  0.8       ,  0.        ,  0.5       ,
        0.66666667,  0.25      ,  0.89473684,  0.83333333,  0.75      ,
        0.33333333,  0.3       ,  0.11111111,  0.42857143,  0.875     ,
        0.85714286,  0.2       ,  0.26315789,  0.6       ,  0.71428571,
        0.53846154,  0.57142857,  0.91489362,  0.86666667,  0.82352941,
        0.78571429,  0.74074074,  0.4       ,  0.375     ,  0.28571429,
        0.14285714,  0.77777778,  0.125     ,  0.9       ,  0.94117647,
        0.92307692,  0.7       ,  0.45454545,  0.88888889,  0.83870968,
        0.9047619 ,  0.92857143,  0.90909091,  0.91666667,  0.84615385,
        0.10526316,  0.98214286,  0.97826087,  0.7518797 ,  0.3125    ,
        0.1       ,  0.18518519,  0.88      ,  0.69230769,  0.625     ,
        0.54545455,  0.41666667,  0.45833333,  0.22222222,  0.81818182,
        0.8125    ,  0.16666667,  0.93103448,  0.88235294,  0.23529412,
        0.63636364,  0.81481481,  0.95652174,  0.64285714,  0.58333333,
        0.94444444,  0.921875  ,  0.86574074,  0.96      ,  0.91304348,
        0.64705882,  0.95833333,  0.09090909,  0.13333333,  0.52941176,
        0.96969697,  0.36363636,  0.07142857,  0.72727273,  0.18181818,
        0.96666667,  0.99074074,  0.97297297,  0.80645161,  0.64102564,
        0.55555556,  0.4375    ,  0.76923077,  0.28      ,  0.15384615,
        0.44444444,  0.5625    ,  0.53333333,  0.47058824,  0.47222222,
        0.23076923,  0.25925926,  0.98876404,  0.88372093,  0.19047619,
        0.94594595,  0.84313725,  0.96629213,  0.72222222,  0.05882353,
        0.27272727,  0.97959184,  0.26666667,  0.30769231,  0.94736842,
        0.27777778,  0.6875    ,  0.92      ,  0.90566038,  0.95      ,
        0.9375    ,  0.9137931 ,  0.82857143,  0.86363636,  0.85      ,
        0.96428571,  0.95238095,  0.08333333,  0.97560976,  0.93333333,
        0.46666667,  0.96153846,  0.24      ,  0.92682927,  0.93548387,
        0.86956522,  0.06666667,  0.98461538,  0.97      ,  0.97619048,
        0.925     ,  0.88461538,  0.61538462,  0.09375   ,  0.79166667,
        0.70588235,  0.45      ,  0.93939394,  0.90322581,  0.68      ,
        0.95454545,  0.04166667,  0.89655172,  0.88571429,  0.38461538,
        0.07692308,  0.12121212,  0.92237443,  0.92156863,  0.36585366,
        0.88095238,  0.84      ,  0.61904762,  0.96129032,  0.96385542,
0.90588235,  0.87878788,  0.05555556,  0.80952381,  0.20689655,
0.07407407,  0.35      ,  0.77272727,  0.91428571,  0.04545455,
0.76470588,  0.70833333,  0.73333333,  0.93650794,  0.8671875 ,
0.75949367,  0.65957447,  0.57692308,  0.41176471,  0.40909091,
0.34693878,  0.30263158,  0.16176471,  0.65      ,  0.96296296,
0.96808511,  0.94915254,  0.98290598,  0.9893617 ,  0.95744681,
0.96268657,  0.98305085,  0.61111111,  0.59183673,  0.98913043,
0.98809524,  0.92982456,  0.78947368,  0.75757576,  0.82608696,
0.96491228,  0.84507042,  0.98412698,  0.96551724,  0.87341772,
0.73913043,  0.7037037 ,  0.98888889,  0.7826087 ,  0.17647059,
0.96226415,  0.94339623,  0.97058824,  0.57894737,  0.47368421,
0.5106383 ,  0.97777778,  0.92352941,  0.78378378,  0.97674419,
0.35714286,  0.94805195,  0.94285714,  0.86538462,  0.43478261,
0.99186992,  0.8627451 ,  0.97142857,  0.98484848,  0.73076923,
0.68181818,  0.63333333,  0.64583333,  0.96774194,  0.05263158,
0.36842105,  0.82926829,  0.92045455,  0.34782609,  0.85365854,
0.91803279,  0.97222222,  0.46153846,  0.2173913 ,  0.82051282,
0.29032258,  0.95754717,  0.91176471,  0.04761905,  0.65714286,
0.13636364,  0.77142857,  0.953125  ,  0.92592593,  0.0862069 ,
0.80555556,  0.20512821,  0.29411765,  0.9925187 ,  0.98564593,
0.99253731,  0.80487805,  0.82142857,  0.76      ,  0.21428571,
0.31914894,  0.02702703,  0.20833333,  0.92105263,  0.78125   ,
0.61290323,  0.97435897,  0.07894737,  0.72413793,  0.03125   ,
0.68421053,  0.97979798,  0.38888889,  0.975     ,  0.80769231,
0.06060606,  0.93023256,  0.97260274,  0.90769231,  0.31372549,
0.15789474,  0.32258065,  0.95959596,  0.21052632,  0.84210526,
0.32      ,  0.92631579,  0.03703704,  1.5       ,  0.11428571,
0.88333333,  0.1875    ,  0.96875   ,  0.64      ,  0.30434783,
0.93150685,  0.88709677,  0.75609756,  0.60606061,  0.54166667,
0.52380952,  0.98275862,  0.98630137,  0.76190476,  0.85106383,
0.79069767,  0.8974359 ,  0.93617021,  0.87234043,  0.0625    ,
0.075     ,  0.39393939,  0.74107143,  0.49090909,  0.90243902,
0.56521739,  0.27027027,  0.03846154,  0.31147541,  0.24528302,
0.97727273,  0.60714286,  0.98360656,  0.95918367,  0.94      ,
0.72      ,  0.15      ,  0.12903226,  0.35294118,  0.14084507,
0.13888889,  0.08219178,  0.03636364,  0.13043478,  0.55172414,
0.64516129,  0.98      ,  0.76271186,  0.98333333,  0.95384615,
0.85294118,  0.13513514,  0.32142857,  0.87912088,  0.82758621,
0.72881356,  0.73684211,  0.86111111,  0.81355932,  0.72839506,
0.73809524,  0.74193548,  0.51612903,  0.7109375 ,  0.69565217,
0.02941176,  0.17391304,  0.85185185,  0.06451613,  0.92727273,
0.08695652,  0.03333333,  0.475     ,  0.32352941,  0.22727273,
0.98113208,  0.42307692,  3.        ,  0.90625   ,  0.8404908 ,
0.72093023,  0.98181818,  0.69047619,  0.05660377,  0.93159609,
0.95604396,  0.95348837,  0.98823529,  0.95774648,  0.94520548,
0.62068966,  0.22058824,  0.25827815,  0.86842105,  0.82222222,
0.89041096,  0.78846154,  0.63157895,  0.98717949,  0.93406593,
0.11538462,  0.04      ,  0.86206897,  0.38095238,  0.95555556,
0.97402597,  0.94230769,  0.47619048,  0.99166667,  0.98387097,
0.93589744,  0.89915966,  0.88489209,  0.89285714,  0.8989899 ,
0.84415584,  0.02857143,  0.98496241,  0.96590909,  0.91240876,
0.94545455,  0.90526316,  0.67924528,  0.109375  ,  0.89090909,
0.98795181,  0.02777778,  0.94642857,  0.18918919,  0.67605634,
0.55      ,  0.53030303,  0.45098039,  0.05454545,  0.96363636,
0.06122449,  0.98039216,  0.99443414,  0.98688525,  0.27586207,
0.1025641 ,  0.11764706,  0.05128205,  0.81395349,  0.69387755,
0.98611111,  0.99466192,  0.98951782,  0.98723404,  0.97122302,
0.97183099,  0.75862069,  0.98550725,  0.97368421,  0.56      ,
0.98657718,  0.90196078,  0.77419355,  0.65625   ,  0.87012987,
0.25581395,  0.21153846,  0.71794872,  0.52      ,  0.02222222,
0.15625   ,  0.05      ,  0.10714286,  0.8902439 ,  0.79310345,
0.65384615,  0.94174757,  0.65116279,  0.59459459,  0.58823529,
0.0952381 ,  0.10638298,  0.20430108,  0.89361702,  0.65217391,
0.84090909,  0.92753623,  0.89156627,  0.89333333,  0.890625  ,
0.38709677,  0.60869565,  0.65853659,  0.42105263,  0.88405797,
0.92473118,  0.86486486,  0.02985075,  0.40625   ,  0.97916667,
0.52631579, 0.19230769, 0.98571429, 0.53571429, 0.12765957,
0.97333333, 0.67857143, 0.93506494, 0.88976378, 0.87037037,
0.81081081, 0.12244898, 0.51724138, 0.89502762, 0.51851852,
0.93181818, 0.82692308, 0.73529412, 0.22857143, 0.62962963,
0.31034483, 0.9787234 , 0.96078431, 0.45714286, 0.98033708,
0.93877551, 0.86904762, 0.98268398, 0.98850575, 0.98148148,
0.56756757, 0.99145299, 0.17948718, 0.71641791, 0.91111111,
0.82653061, 0.98734177, 0.984375  , 0.58064516, 0.47826087,
0.44      , 0.39130435, 0.20454545, 0.98351648, 0.95714286,
0.96503497, 0.86263736, 0.12      , 0.76595745, 0.86440678,
0.89873418, 0.91525424, 0.91071429, 0.88636364, 0.48275862,
0.03571429, 0.98591549, 0.69090909, 0.9516129 , 0.48780488,
0.13793103, 0.08108108, 0.11904762, 0.80597015, 0.9273743 ,
0.89308176, 0.3960396 , 0.98245614, 0.99415205, 0.98969072,
0.96721311, 0.64788732, 0.23333333, 0.99196787, 0.87603306,
0.86567164, 0.87096774, 0.83269962, 0.84057971, 0.82978723,
0.78333333, 0.80434783, 0.78787879, 0.95121951, 0.13157895,
0.96923077, 0.08571429, 0.98701299, 0.775     , 0.90163934,
0.9245283 , 0.34285714, 0.14814815, 0.83529412, 0.79487179,
0.74666667, 0.73239437, 0.74285714, 0.63855422, 0.63414634,
0.04347826, 0.84782609, 0.81632653, 0.94871795, 0.04301075,
0.22580645, 0.98924731, 0.84375   , 0.94047619, 0.91891892,
0.85416667, 0.67307692, 0.97468354, 0.74545455, 0.7311828 ,
0.45945946, 0.97938144, 0.92405063, 0.97101449, 0.68493151,
0.58      , 0.79104478, 0.68888889, 0.99278846, 0.87179487,
0.36111111, 0.36206897, 0.21621622, 0.08510638, 0.62745098,
0.48387097, 0.25806452, 0.98633257, 0.97761194, 0.9789916 ,
0.97530864, 0.96470588, 0.95588235, 0.89705882, 0.85483871,
0.72463768, 0.05405405, 0.8630137 , 0.16129032, 0.29166667,
0.70454545, 0.23255814, 0.94623656, 0.95412844, 0.75555556,
0.67647059, 0.08      , 0.98837209, 0.4137931 , 0.59375   ,
0.52777778, 0.48      , 0.46551724, 0.34146341, 0.36619718,
0.24137931, 0.30188679, 0.265625  , 0.09756098, 0.13559322,
0.91440953, 0.91509434, 0.89622642, 0.86086957, 0.85436893,
0.85245902, 0.81609195, 0.8030303 , 0.78      , 0.77647059,
0.79545455, 0.76521739, 0.77966102, 0.72321429, 0.72972973,
0.67741935, 0.62222222, 0.98652291, 0.78873239, 0.26086957,
0.71875   , 0.39285714, 0.87804878, 0.69444444, 0.79411765,
0.992     , 0.97647059, 0.31578947, 0.31707317, 0.88679245,
0.79591837, 0.9261745 , 0.8629174 , 0.98666667, 0.26923077,
0.17857143, 0.38235294, 0.99180328, 0.15942029, 0.90277778,
0.36      , 0.98507463, 0.7721519 , 0.04651163, 0.68965517,
0.95890411, 0.06766917, 0.56603774, 0.69767442, 0.93442623,
0.97807757, 0.52173913, 0.75471698, 0.70967742, 0.98076923,
0.23809524, 0.95522388, 0.87142857, 0.74418605, 0.83783784,
0.75510204, 0.59090909, 0.89711934, 0.87301587, 0.89795918,
0.73493976, 0.99122807, 0.96644295, 0.95876289, 0.86046512,
0.07954545, 0.76666667, 0.16216216, 0.02739726, 0.09677419,
0.27659574, 0.83636364, 0.65306122, 0.53521127, 0.97580645,
0.93478261, 0.7755102 , 0.98672566, 0.99619772, 0.9876161 ,
0.97169811, 0.92957746, 0.97534247, 0.97123894, 0.97191011,
0.97969543, 0.96478873, 0.95491803, 0.03508772, 0.73584906,
0.96341463, 0.69135802, 0.61764706, 0.74358974, 0.92428198,
0.93421053, 0.78723404, 0.37931034, 0.95762712, 0.92783505,
0.16      , 0.34615385, 0.76388889, 0.63461538, 0.68518519,
0.67567568, 0.675     , 0.98394495, 0.40540541, 0.57575758,
0.89189189, 0.86330935, 0.18604651, 0.98897059, 0.9673913 ,
0.9379562 , 0.93243243, 0.98347107, 0.19444444, 0.91139241,
0.84444444, 0.93220339, 0.968     , 0.53125   , 0.71698113,
0.80672269, 0.02325581, 0.21875   , 0.89519651, 0.71604938,
0.2826087 , 0.07462687, 0.97887324, 0.58974359, 0.33928571,
0.21818182, 0.06779661, 0.28947368, 0.9625    , 0.95081967,
0.91549296, 0.10344828, 0.99212598, 0.84848485, 0.07317073,
0.97831325, 0.97972973, 0.96511628, 0.9202454 , 0.90140845,
0.14492754, 0.37837838, 0.46511628, 0.98765432, 0.98697068,
0.91875   , 0.84274194, 0.84693878, 0.828125  , 0.6969697 ,
       0.02439024,  0.99090909,  0.94078947,  0.94666667,  0.98979592,
       0.81132075,  0.87654321,  0.15277778,  0.68085106,  0.82022472,
       0.88321168,  0.96703297,  0.08955224,  0.31818182,  0.96610169,
       0.95302013,  0.94375   ,  0.81578947,  0.98319328,  0.98639456,
       0.89719626,  0.99107143,  0.64864865,  0.88947368,  0.78688525,
       0.825     ,  0.34482759,  0.95098039,  0.04444444,  0.95804196,
       0.18461538,  0.97663551,  0.97196262,  0.72641509,  0.3877551 ,
       0.85964912,  0.43333333,  0.34920635,  0.09195402,  0.98529412,
       0.36666667,  0.9695122 ,  0.83928571,  0.75675676,  0.5862069 ,
       0.98780488,  0.48648649,  0.03448276,  0.04285714,  0.96039604,
       0.01098901,  0.43902439,  0.31428571,  0.95327103,  0.98695652,
       0.94949495,  0.89230769,  0.83673469,  0.75438596,  0.24242424,
       0.25961538,  0.75409836,  0.14634146,  0.9627907 ,  0.30232558,
       0.97163121,  0.37037037,  0.0212766 ,  0.5952381 ,  0.995     ,
       0.99029126,  0.99689441,  0.9691358 ,  0.74      ,  0.34210526,
       0.38596491,  0.97857143,  0.90697674,  0.20588235,  0.97807018,
       0.97905759,  0.77570093,  0.11666667,  0.85384615,  0.92380952,
       0.76829268,  0.91612903,  0.91025641,  0.99019608,  0.58181818,
       0.46875   ,  0.73170732,  0.97321429,  0.70731707,  0.59259259,
       0.10810811,  0.89830508,  0.44827586,  0.97457627,  0.8961039 ,
       0.54285714,  0.07843137,  0.98251748,  0.90234375,  0.02830189,
       0.98947368,  0.98684211,  0.17073171,  0.69724771,  0.976     ,
       0.98165138,  0.97590361,  0.90438247,  0.40740741,  0.8490566 ,
       0.96410256,  0.9380531 ,  0.77358491,  0.17777778,  0.44117647,
       0.03076923,  0.87755102,  0.0962963 ,  0.91836735,  0.87951807,
       0.80851064,  0.99141104])
```

In [8]:

```
data["%upvote"]=pd.cut(data["Helpful%"],bins=[-1,0,0.2,0.4,0.6,0.8,1],labels=["Empty","0
```

In [9]:

```
data.head()
```

Out[9]:

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	
0	1	B001E4KFG0	A3SGXH7AUHU8GW	delmartian	1	1	5	130
1	2	B00813GRG4	A1D87F6ZCVE5NK	dll pa	0	0	1	134
2	3	B000LQOCH0	ABXLMWJIXXAIN	Natalia Corres "Natalia Corres"	1	1	4	121
3	4	B000UA0QIQ	A395BORC6FGVXV	Karl	3	3	2	130
4	5	B006K2ZZ7K	A1UQRSCLF8GW1T	Michael D. Bigham "M. Wassir"	0	0	5	135

Question no 1

Analyse upvotes for different different scores

In [10]:

```
data.groupby(["Score","%upvote"]).agg("count")
```

Out[10]:

Score	%upvote	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Tim
1	**Empty**	8060	8060	8060	8060	8060	8060	806
	0-20%	2338	2338	2338	2338	2338	2338	233
	20-40	4649	4649	4649	4649	4649	4649	464
	40-60%	6586	6586	6586	6586	6586	6586	658
	60-80%	5838	5838	5838	5836	5838	5838	583
	80-100%	12531	12531	12531	12531	12531	12531	1253
2	**Empty**	4234	4234	4234	4234	4234	4234	423
	0-20%	762	762	762	762	762	762	76
	20-40	1618	1618	1618	1618	1618	1618	161
	40-60%	3051	3051	3051	3051	3051	3051	305
	60-80%	2486	2486	2486	2486	2486	2486	248
	80-100%	7014	7014	7014	7014	7014	7014	701
3	**Empty**	5062	5062	5062	5062	5062	5062	506
	0-20%	474	474	474	474	474	474	47
	20-40	1506	1506	1506	1506	1506	1506	150
	40-60%	3384	3384	3384	3384	3384	3384	338
	60-80%	2754	2754	2754	2754	2754	2754	275
	80-100%	11037	11037	11037	11037	11037	11037	1103
4	**Empty**	4780	4780	4780	4780	4780	4780	478
	0-20%	116	116	116	116	116	116	11
	20-40	909	909	909	909	909	909	90
	40-60%	3185	3185	3185	3185	3185	3185	318
	60-80%	2941	2941	2941	2941	2941	2941	294
	80-100%	26707	26707	26707	26707	26707	26707	2670
5	**Empty**	11638	11638	11638	11638	11638	11638	1163
	0-20%	432	432	432	432	432	432	43
	20-40	2275	2275	2275	2275	2275	2275	227
	40-60%	10312	10312	10312	10312	10312	10312	1031
	60-80%	11060	11060	11060	11060	11060	11060	1106

80-100%	140661	140661	140661	140659	140661	140661	14066

In [11]:

```
data_s=data.groupby(["Score","%upvote"]).agg({"Id": "count"}).reset_index()
```

In [12]:

```
data_s
```

Out[12]:

	Score	%upvote	Id
0	1	Empty	8060
1	1	0-20%	2338
2	1	20-40	4649
3	1	40-60%	6586
4	1	60-80%	5838
5	1	80-100%	12531
6	2	Empty	4234
7	2	0-20%	762
8	2	20-40	1618
9	2	40-60%	3051
10	2	60-80%	2486
11	2	80-100%	7014
12	3	Empty	5062
13	3	0-20%	474
14	3	20-40	1506
15	3	40-60%	3384
16	3	60-80%	2754
17	3	80-100%	11037
18	4	Empty	4780
19	4	0-20%	116
20	4	20-40	909
21	4	40-60%	3185
22	4	60-80%	2941
23	4	80-100%	26707
24	5	Empty	11638
25	5	0-20%	432
26	5	20-40	2275
27	5	40-60%	10312
28	5	60-80%	11060
29	5	80-100%	140661

Create Pivot table and Heat Map

In [13]:

```
pivot_table=data_s.pivot(index="%upvote",columns="Score")
```

In [14]:

```
pivot_table
```

Out[14]:

	Id				
Score	**1**	**2**	**3**	**4**	**5**
%upvote					
Empty	8060	4234	5062	4780	11638
0-20%	2338	762	474	116	432
20-40	4649	1618	1506	909	2275
40-60%	6586	3051	3384	3185	10312
60-80%	5838	2486	2754	2941	11060
80-100%	12531	7014	11037	26707	140661

In [15]:

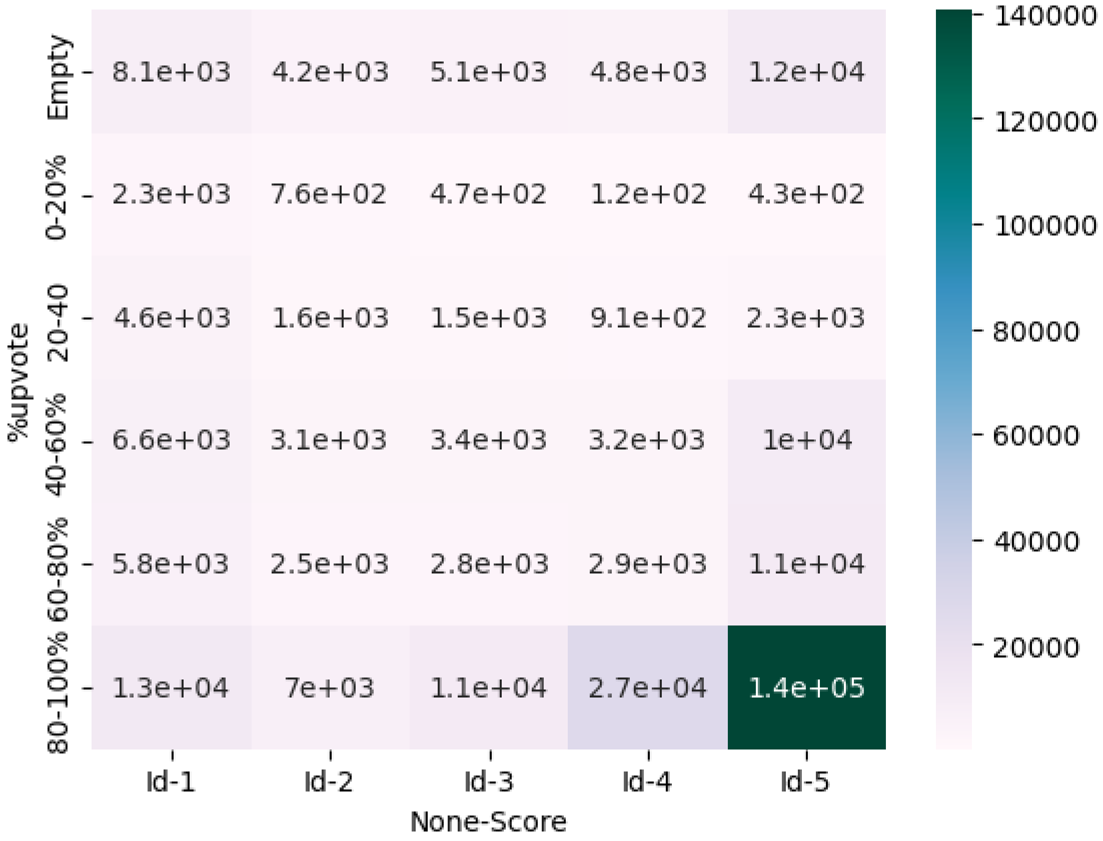
```
import seaborn as sns
```

In [16]:

```
sns.heatmap(pivot_table,annot=True,cmap="PuBuGn")
```

Out[16]:

```
<AxesSubplot:xlabel='None-Score', ylabel='%upvote'>
```



In [17]:

```
data["Score"].unique()
```

Out[17]:

```
array([5, 1, 4, 2, 3], dtype=int64)
```

In [24]:

```
final_data = data[data["Score"]!=3]
```

In [25]:

```
X= final_data["Text"]
```

In [26]:

```
y_dict = {1:0,2:0,4:1,5:1}
Y= final_data["Score"].map(y_dict)
```

In [27]:

```
X
```

Out[27]:

```
0         I have bought several of the Vitality canned d...
1         Product arrived labeled as Jumbo Salted Peanut...
2         This is a confection that has been around a fe...
3         If you are looking for the secret ingredient i...
4         Great taffy at a great price.  There was a wid...
                                ...                        
568449    Great for sesame chicken..this is a good if no...
568450    I'm disappointed with the flavor. The chocolat...
568451    These stars are small, so you can give 10-15 o...
568452    These are the BEST treats for training and rew...
568453    I am very satisfied ,product is as advertised,...
Name: Text, Length: 525814, dtype: object
```

In [28]:

```
Y
```

Out[28]:

```
0         1
1         0
2         1
3         0
4         1
         ..
568449    1
568450    0
568451    1
568452    1
568453    1
Name: Score, Length: 525814, dtype: int64
```

Bag of words

In [29]:

```
from sklearn.feature_extraction.text import CountVectorizer
```

In [30]:

```
count_vec = CountVectorizer(stop_words="english")
```

In [31]:

```
count_vec_X = count_vec.fit_transform(X)
```

In [32]:

```
count_vec_X.shape[1]
```

Out[32]:

```
114969
```

Model Evaluation

In [33]:

```
from sklearn.model_selection import train_test_split
import warnings
warnings.filterwarnings("ignore")
```

In [34]:

```
X_train,X_test,Y_train,Y_test = train_test_split(count_vec_X,Y)
```

In [35]:

```
print(X_train.shape,X.shape,X_test.shape)
```

```
(394360, 114969) (525814,) (131454, 114969)
```

In [36]:

```
from sklearn.linear_model import LogisticRegression
```

In [37]:

```
log_reg = LogisticRegression()
```

In [38]:

```
log_reg.fit(X_train,Y_train)
```

Out[38]:

```
LogisticRegression()
```

In [39]:

```
log_reg.score(X_test,Y_test)
```

Out[39]:

```
0.93670029059595
```

Fetch 20 positive words and 20 negative words

In [40]:

```
w =count_vec.get_feature_names()
w
```

Out[40]:

```
['00',
 '000',
 '0000',
 '000001',
 '00001',
 '000013',
 '0000soo',
 '0001',
 '000111052',
 '0002251337',
 '0003',
 '0004',
 '000iu',
 '000kwh',
 '000mg',
 '000mi',
 '000s',
 '000su',
 '000usd',
 '000v',
 '001',
 '00100',
 '00127',
 '00128',
 '00129',
 '00130',
 '00131',
 '00132',
 '00134',
 '00136',
 '00139',
 '001bru',
 '002',
 '0020100604',
 '00202',
 '00227',
 '0023',
 '003',
 '004',
 '00493',
 '005',
 '00533',
 '0060187654',
 '0060721855',
 '0060928115',
 '0060959584',
 '0061658197',
 '006176793x',
 '0067575986',
 '0069615',
 '007',
 '00703',
 '00704',
 '0071468633',
 '0071477845',
 '0071486011',
 '0071499849',
 '008',
 '0099',
 '00a',
 '00am',
 '00b',
 '00gr',
 '00lb',
 '00m',
 '00pm',
 '00s',
 '00something',
 '00z',
 '01',
 '010',
 '0100',
 '01014',
 '01069',
 '011',
 '012',
 '013',
 '01317',
 '01318',
 '014',
 '0140444254',
 '0140446680',
 '0143114964',
 '014mg',
 '015',
 '017',
 '0174',
 '018',
 '0188',
 '019',
 '01915',
 '0199232768',
 '0199535892',
 '0199536066',
 '02',
 '020',
 '02027c',
 '02043',
 '02115',
 '022313',
 '022413',
 '022813',
 '023',
 '024600017558',
 '025',
 '025913',
 '025968680a',
 '02604',
 '026220',
 '027',
 '0273',
 '02762',
 '028',
 '02jan2012',
 '02oq8o2bbaxaacmkmmmf4tby',
 '03',
       '030',
       '0300',
       '030113',
       '0303',
       '030712',
       '030713',
       '0307261581',
       '0307460169',
       '0307474291',
       '0307720497',
       '0307887960',
       '0307949435',
       '030813',
       '031231521x',
       '0312362919',
       '0312377428',
       '0312545525',
       '0312629818',
       '0312649940',
       '031513',
       '0316074284',
       '031612737x',
       '0316129445',
       '031613',
       '031613290x',
       '0316735507',
       '032',
       '032513',
       '032712',
       '032813',
       '034',
       '03430',
       '034549458x',
       '035',
       '03510',
       '035273',
       '0373',
       '0373892322',
       '0373892349',
       '03755',
       '0375757996',
       '0376024933',
       '0377',
       '038',
       '038542017x',
       '03885',
       '039',
       '0393066304',
       '0393070212',
       '03jan12',
       '03oz',
       '04',
       '040',
       '040413',
       '040513',
       '040612',
       '041113',
       '041213',
       '041224721302',
       '041313',
       '041913',
       '042013',
       '042113',
       '042513',
       '0425204138',
       '042608460503',
       '042813',
       '043',
       '0439',
       '044',
       '044000025298',
       '0440129613',
       '0446675385',
       '04472700',
       '045',
       '0451155505',
       '0451184963',
       '0452285801',
       '0461',
       '04691',
       '0470041153',
       '0470194960',
       '047028188x',
       '0470425245',
       '0470440856',
       '0470913029',
       '0470916575',
       '0471168572',
       '0471202711',
       '0472066978',
       '0475',
       '048',
       '04820',
       '04830',
       '049',
       '04mg',
       '04oz',
       '05',
       '050',
       '050411',
       '050913',
       '050oz',
       '051',
       '051013',
       '051213',
       '05140',
       '051513',
       '0517452502',
       '051813',
       '051913',
       '052',
       '0520269926',
       '052100071336',
       '052213',
       '052313',
       '052413',
       '052913',
       '053229',
       '0545044251',
       '055',
       '0553380788',
       '056',
       '05608',
       '0563214546',
       '05715',
       '059',
       '0595318452',
       '05oz',
       '06',
       '060',
       '060113',
       '060713',
       '060813',
       '061',
       '0610',
       '061113',
       '061313',
       '061413',
       '061513',
       '062',
       '062213',
       '0625',
       '06254',
       '062613',
       '062713',
       '062813',
       '063',
       '0631',
       '06312',
       '064',
       '065',
       '066',
       '067',
       '0674005392',
       '0679767959',
       '06799',
       '0684800012',
       '0688',
       '0688092616',
       '0688157920',
       '06g',
       '06l33t',
       '06oz',
       '06sep11',
       '06sep12',
       '07',
       '070',
       '07002',
       '07003',
       '07036',
       '070590080010',
       '07065',
       '07078',
       '0709',
       '070913',
       '0712',
       '071313',
       '071813',
       '071913',
       '0735814414',
       '0738213861',
       '073821423x',
       '0738551856',
       '0738713597',
       '074',
       '0740765353',
       '0740779729',
       '0740784137',
       '0743223225',
       '0743246268',
       '0743292545',
       '0745',
       '075',
       '0761135812',
       '076114479x',
       '0761160981',
       '0762426020',
       '0762435259',
       '0764135775',
       '0764544659',
       '0764559176',
       '0764569112',
       '0764578650',
       '07652',
       '0767908236',
       '0768ppm',
       '0777',
       '0785383239',
       '078799622x',
       '0790700506',
       '0790742780',
       '0792165055',
       '07m2402',
       '08',
       '080',
       '0804835942',
       '0805089586',
       '0806',
       '080942925x',
       '0810',
       '0811845354',
       '0811867811',
       '0816524882',
       '0816650187',
       '08204',
       '0821811896051997357371911700758600066844838040192l',
       '0825305845',
       '083',
       '083613561x',
       '083619263x',
       '0836192648',
       '0836194942',
       '084',
       '0848734378',
       '0848806832',
       '085',
       '0865477043',
       '0872203492',
       '088',
       '0880015047',
       '08873',
       '089',
       '089036241397',
       '089036241434',
       '0892810998',
       '0892817267',
       '0894803123',
       '0895264641',
       '08m07c0',
       '08m13c8',
       '08oz',
       '09',
       '090',
       '090214',
       '090227921',
       '091',
       '091012',
       '0913',
       '0913990604',
       '092',
       '09244791666',
       '0929',
       '0929173252',
       '0929173309',
       '093',
       '0937',
       '0939165422',
       '0939165562',
       '0941599604',
       '0943151201',
       '0947',
       '095',
       '0961959533',
       '0963796836',
       '0967089735',
       '0967089751',
       '0969276818',
       '097',
       '0970245823',
       '097221643x',
       '0972722726',
       '0976896931',
       '0979201802',
       '0979676460',
       '097977280x',
       '0979780802',
       '098',
       '0980137659',
       '0981623506',
       '098214251x',
       '0982207700',
       '0982207786',
       '0982428669',
       '0982428693',
       '098253311x',
       '0982708319',
       '09mg',
       '0arty',
       '0c',
       '0cm',
       '0d',
       '0f',
       '0g',
       '0g12',
       '0g5',
       '0gr',
       '0grams',
       '0iu',
       '0l',
       '0lb',
       '0lp',
       '0mcg',
       '0mg',
       '0mg0',
       '0n',
       '0ne',
       '0nly',
       '0o',
       '0on',
       '0oz',
       '0p',
       '0pen',
       '0ppm',
       '0r',
       '0ther',
       '0unce',
       '0ver',
       '0xk6hzpjrkaed855hewp',
       '0z',
       '10',
       '100',
       '1000',
       '10000',
       '100000',
       '1000000',
       '1000000000000000000000000',
       '100001',
       '10000km',
       '10000x',
       '1000g',
       '1000grams',
       '1000iu',
       '1000mg',
       '1000ml',
       '1000mph',
       '1000s',
       '1000th',
       '1000w',
       '1000watt',
       '1000x',
       '1001',
       '10010',
       '1002',
       '10041224721309',
       '100432',
       '1007',
       '100cal',
       '100calories',
       '100cals',
       '100count',
       '100ct',
       '100ea',
       '100f',
       '100ft',
       '100g',
       '100gm',
       '100gms',
       '100gr',
       '100gram',
       '100k',
       '100kg',
       '100lb',
       '100lbs',
       '100lu',
       '100m',
       '100mcg',
       '100mg',
       '100mile',
       '100ml',
       '100mph',
       '100o',
       '100pc',
       '100piece',
       '100pk',
       '100plus',
       '100psi',
       '100s',
       '100sheet',
       '100th',
       '100ug',
       '100w',
       '100watt',
       '100x',
       '100x3',
       '100year',
       '100°',
       '101',
       '1010',
       '101012',
       '1010mg',
       '1011',
       '1012',
       '10134',
       '10145camp09',
       '1015',
       '1017',
       '101812',
       '101mg',
       '102',
       '1020',
       '1020mg',
       '1021',
       '102308',
       '1024',
       '1025g',
       '102612',
       '102lbs',
       '103',
       '1030',
       '10312',
       '1035',
       '1035985',
       '104',
       '1040mg',
       '1047',
       '104oz',
       '105',
       '1050',
       '10504',
       '1050mgs',
       '1054',
       '1056',
       '105f',
       '105g',
       '105lbs',
       '105mg',
       '105°',
       '106',
       '10620',
       '107',
       '1072',
       '1073',
       '1074',
       '1076',
       '1077',
       '1078',
       '1079',
       '108',
       '1080',
       '1080p',
       '1082',
       '1083',
       '1084',
       '1085',
       '1086',
       '1087',
       '1088',
       '1089',
       '108g',
       '108mg',
       '109',
       '1090',
       '1091',
       '1092',
       '1093',
       '1095',
       '10954012',
       '1096',
       '1099',
       '10_',
       '10a',
       '10am',
       '10b',
       '10bucks',
       '10c',
       '10can',
       '10cents',
       '10days',
       '10ft',
       '10g',
       '10gm',
       '10gr',
       '10gram',
       '10grams',
       '10grms',
       '10hr',
       '10hrs',
       '10ib',
       '10in',
       '10ish',
       '10jul18',
       '10k',
       '10kg',
       '10lb',
       '10lbs',
       '10m',
       '10mar12',
       '10mcg',
       '10mg',
       '10mg0',
       '10mgs',
       '10min',
       '10mins',
       '10minutes',
       '10ml',
       '10mo',
       '10month',
       '10months',
       '10mos',
       '10oz',
       '10p',
       '10pc',
       '10pcs',
       '10pk',
       '10pks',
       '10pm',
       '10pound',
       '10s',
       '10sec',
       '10seconds',
       '10tbsp',
       '10th',
       '10ths',
       '10tiny',
       '10w',
       '10wks',
       '10x',
       '10x10',
       '10x12x12',
       '10x5',
       '10x7x3',
       '10year',
       '10years',
       '10yo',
       '10yr',
       '10yrs',
       '10°',
       '11',
       '110',
       '1100',
       '1100mg',
       '1100w',
       '1100yrs',
       '1101',
       '1103',
       '1105',
       '110576',
       '110583',
       '1106',
       '110712',
       '110812',
       '110912',
       '110lb',
       '110lbs',
       '110m3',
       '110mg',
       '110v',
       '111',
       '11105',
       '1111',
       '1112',
       '111412',
       '1116',
       '11164',
       '11173407',
       '1118',
       '111812',
       '112',
       '1120',
       '1120mg',
       '1121',
       '1122',
       '1123',
       '11231',
       '11279p',
       '112812',
       '112g',
       '113',
       '1130',
       '1130am',
       '11370',
       '113g',
       '114',
       '1140091',
       '114159',
       '114316',
       '115',
       '1150',
       '1150mg',
       '1152',
       '1155',
       '1157',
       '115g',
       '115iu',
       '115mg',
       '115mg5',
       '115s',
       '116',
       '11629',
       '1165',
       '1167',
       '117',
       '11713',
       '117lbs',
       '118',
       '1180mg',
       '1186',
       '118lbs',
       '119',
       '1190mg',
       '1197',
       '1198',
       '11am',
       '11aug12',
       '11g',
       '11gm',
       '11gms',
       '11gr',
       '11gs',
       '11in',
       '11ish',
       '11k',
       '11l',
       '11lb',
       '11lbs',
       '11m',
       '11mg',
       '11mo',
       '11month',
       '11mths',
       '11ounces',
       '11oz',
       '11pm',
       '11s',
       '11t',
       '11th',
       '11wks',
       '11x13',
       '11year',
       '11yo',
       '11yr',
       '11yrs',
       '12',
       '120',
       '1200',
       '1200cal',
       '1200candy',
       '1200mg',
       '1200v',
       '1200w',
       '1200x',
       '120112',
       '120312',
       '1203123',
       '120322fs1d',
       '120374',
       '1206',
       '1208084',
       '1209',
       '120c28',
       '120cals',
       '120ct',
       '120g',
       '120hz',
       '120ish',
       '120lb',
       '120lbs',
       '120mg',
       '120ml',
       '120s',
       '120th',
       '120v',
       '120z',
       '121',
       '1219bestby05dec2012',
       '122',
       '1220',
       '1220mg',
       '12290',
       '122nd',
       '123',
       '1230',
       '1234',
       '123411286',
       '12356',
       '124',
       '1240',
       '12424',
       '12442909',
       '1245',
       '1246',
       '124th',
       '125',
       '1250',
       '1250ml',
       '125g',
       '125lbs',
       '125mg',
       '125th',
       '125v',
       '126',
       '1260255555',
       '1260257694',
       '1266273814',
       '127',
       '1270744039',
       '1279',
       '127lb',
       '128',
       '1280',
       '128077',
       '128078',
       '1280mg',
       '1281996897',
       '128516',
       '128522',
       '128523',
       '128525',
       '128541',
       '1288868059',
       '1289',
       '128fl',
       '128g',
       '128gm',
       '128mg',
       '128oz',
       '129',
       '1290',
       '1290838229',
       '1290mg',
       '1291331352',
       '12931',
       '1294674431',
       '1297',
       '1297861877',
       '1298166499',
       '12am',
       '12cal',
       '12cans',
       '12ct',
       '12cup',
       '12cups',
       '12de07',
       '12ea',
       '12fl',
       '12g',
       '12gm',
       '12gms',
       '12gprotein',
       '12gram',
       '12grams',
       '12gs',
       '12h2o',
       '12hr',
       '12hrs',
       '12in',
       '12inch',
       '12ish',
       '12jul12',
       '12jul2011',
       '12k',
       '12lb',
       '12lbs',
       '12m',
       '12min',
       '12mo',
       '12month',
       '12ounce',
       '12ox',
       '12oz',
       '12ozbottle',
       '12ozs',
       '12pack',
       '12packages',
       '12packs',
       '12pak',
       '12pk',
       '12pks',
       '12pm',
       '12pounds',
       '12qt',
       '12qts',
       '12qty',
       '12s',
       '12seconds',
       '12th',
       '12v',
       '12w',
       '12wk',
       '12x',
       '12x10',
       '12x10x2',
        '12x12',
        '12x14',
        '12x16',
        '12x17',
        '12x18x6',
        '12x24',
        '12x25',
        '12x2oz',
        '12x32',
        '12x4',
        '12x5',
        '12x8',
        '12x9x4',
        '12years',
        '12yo',
        '12yr',
        '12yrs',
        '12½',
        '13',
        '130',
        '1300',
        '1300131716',
        '1300373343',
        '1300mg',
        '1300s',
        '1300w',
        '1300watt',
        '1302206977',
        '1306',
        '1308',
        '1308_12',
        '130cal',
        '130calories',
        '130db',
        '130g',
        '130ish',
        '130l',
        '130lb',
        '130lbs',
        '130mg',
        '130x',
        '131',
        '1310712',
        '1316',
        '132',
        '13214',
        '1325488826',
        '1326377612',
        '1328216581',
        '1328216609',
        '1328216637',
        '1328639422',
        '13289',
        '1329666521',
        '132mgs',
        '133',
        '1330',
        '1331447807',
        '1333mg',
        '1334249885',
        '1334324776',
        '1334324953',
        '1334767059',
        '1334882068',
        '1335572191',
        '1336984170',
'1337',
'1337202670',
'1339675640',
'133g',
'133mg',
'134',
'1341338906',
'1342132011',
'1342563776',
'13432',
'1344360665',
'1344600644',
'1346527950',
'1346955176',
'1347903950',
'1348333106',
'1348518159',
'1348706702',
'1349143571',
'1349906835',
'1349906863',
'1349919269',
'134mg',
'135',
'1350',
'13502',
...]
```

In [41]:

```
coef=log_reg.coef_.tolist()[0]
coef
```

Out[41]:

```
[-0.3777485941937896,
 0.043005373229101564,
 0.19062935315845883,
 -0.008009032728656345,
 0.0,
 -0.010904172343068126,
 0.00027777349801800025,
 -0.016112293681180476,
 0.0003756088234374901,
 0.00030256849147725654,
 0.001185053782239346,
 -0.00740197592557816,
 0.0,
 0.018052942854458166,
 -0.40936431539792745,
 0.0,
 0.0015506889590511039,
 0.003834688165460538,
 0.0015050230235818369,
 4.138498380379814e-05,
 0.08184176887875508,
 0.027852625093210657,
 0.034299463173945256,
 0.034299463173945256,
 0.034299463173945256,
 0.034299463173945256,
 0.034299463173945256,
 0.034299463173945256,
 0.034299463173945256,
 0.034299463173945256,
 0.034299463173945256,
 -0.00028498392366002275,
 0.049772409891837996,
 0.0001473985776355583,
 -0.09290584407610772,
-0.0005042906076782855,
0.00014585296608859026,
0.0032769101949834626,
0.1158291997895628,
2.1121113325493957e-05,
-0.1062025516624344,
0.0,
3.188941050693572e-05,
0.006328066235872159,
0.0,
8.67084168657343e-05,
0.010619556307999123,
3.342586624401128e-05,
0.0001473985776355583,
0.0,
0.02100132325248696,
0.0003208483621280013,
0.0003208483621280013,
0.04075749004416965,
9.90020041673721e-05,
9.239480440807413e-06,
0.0015246279832288098,
-0.06869454127266657,
-0.06216814687901886,
0.07111455346407354,
-0.051412692704754855,
0.018798626079830204,
0.002298925169765569,
1.4410774016027776e-05,
0.00032310120204511193,
0.029256514266450557,
0.0007868547353343861,
0.00616148429977211,
-0.01648790250461784,
0.35378822017396055,
0.1283587462874257,
0.0,
-0.0007426552791492729,
-0.0020341809892687355,
0.0004768255007102661,
0.001435709867418438,
0.0,
0.042735822735085964,
0.042735822735085964,
-0.0072124582424216935,
0.0,
0.0,
-0.014678792751677276,
8.760473689350791e-06,
0.0014193621928547462,
0.0,
0.07436701601145383,
0.0,
0.008971401499149628,
0.00018951768315652996,
0.06028092125802582,
0.0005274069973606182,
0.0012214570512148978,
0.00011947135834908432,
-0.8863165388451218,
0.0002482790822714311,
0.0,
0.0007616841978360522,
0.0,
-0.051770853853842974,
-0.025885426926921487,
    -0.025885426926921487,
    0.00019422310599651212,
    0.0319562198126151,
    0.01569741724065896,
    0.04218512681713085,
    -0.008241722641986401,
    -0.09290584407610772,
    0.0007669511912434961,
    0.0,
    0.00044339183519912797,
    0.0003016581329720721,
    0.00034433540991150885,
    0.047408224700764764,
    0.0,
    -1.093451358840694,
    -0.15019466730008021,
    8.495096495121436e-05,
    -0.025885426926921487,
    0.0,
    -0.06963072171117288,
    -0.025885426926921487,
    0.00013559911391236806,
    0.0,
    0.005041533810533275,
    0.0016668524046326767,
    0.0,
    0.012797229893989915,
    -0.07765628078076431,
    1.3910560758002313e-05,
    0.00013754330593871458,
    0.00010047754088018956,
    0.0001162408591986539,
    8.610170879522352e-05,
    0.00014230317144595054,
    0.0007616841978360522,
    2.379726995519913e-05,
    0.00019121532768932278,
    8.67084168657343e-05,
    -0.025885426926921487,
    0.005404248341710005,
    0.00018258064201369094,
    -0.009556482488767302,
    -0.07765628078076431,
    0.009612867575933347,
    -0.025885426926921487,
    0.0023593858433608376,
    0.0,
    0.016375979402096997,
    0.012182311321955788,
    0.02982385398044501,
    0.00216859522783909,
    -0.00740197592557816,
    0.0,
    0.0,
    -0.0021224184644187323,
    0.0,
    0.001535121729611459,
    -0.00740197592557816,
    0.0006461002334477703,
    0.0,
    0.0,
    0.012276038073063454,
    0.0,
    0.05220969563312436,
    1.7443778117284716e-05,
    -0.06252208140120712,
-0.2424266034735899,
0.07783087252803467,
-0.051770853853842974,
-0.025885426926921487,
0.0,
-0.025885426926921487,
-0.025885426926921487,
0.0,
-0.025885426926921487,
-0.051770853853842974,
-0.025885426926921487,
-0.07765628078076431,
-0.051770853853842974,
0.0004847492217230299,
6.801518944322068e-05,
-0.025885426926921487,
0.009742298940041461,
-0.0020393833763429035,
0.0012972101683326306,
0.05547631534151864,
0.00800335504194764,
0.0001849962810051342,
0.05547631534151864,
0.0,
0.0,
6.434651682188196e-05,
0.0001038131590523815,
0.0,
0.0,
-0.16737063489397686,
0.0,
0.0004095197369989343,
0.0001038131590523815,
0.0,
0.15290893648161688,
0.0,
0.000569386265878711,
0.0,
0.035300881965462313,
0.011095754367752255,
0.00018951768315652996,
-0.04925833905581029,
0.00015466648162177274,
0.00018951768315652996,
8.760473689350791e-06,
0.00020939024334011718,
-0.34977395984252846,
0.0007580707326261198,
-0.0027593135180852247,
-0.025885426926921487,
0.0545649503182274 8,
0.00018951768315652996,
-0.025885426926921487,
-0.025885426926921487,
0.01778432787825426,
-0.025885426926921487,
0.010003178825386302,
-0.025885426926921487,
-0.025885426926921487,
0.0013231972392201286,
0.0008565500716414197,
0.0319562198126615 1,
-0.025885426926921487,
-0.051770853853842974,
-0.025885426926921487,
-0.025885426926921487,
-0.0010196916881714518,
-0.04109153102649564,
-0.020652281868525278,
0.00020084696445493038,
-0.08202347623968524,
0.000815596036431582,
0.0004471009393093731,
-0.09377453738920752,
0.0005547662976594469,
0.000183323559270741,
0.002959239067324904,
0.041679009861612194,
-0.07041970678927878,
-0.07765628078076431,
-0.025885426926921487,
-0.051770853853842974,
0.00018951768315652996,
-0.20221540620137965,
-0.025885426926921487,
-0.025885426926921487,
-0.025885426926921487,
-0.051770853853842974,
0.0135360299492321,
-0.051770853853842974,
0.06653833146393061,
-0.17356265790036976,
-0.025885426926921487,
-0.051770853853842974,
-0.051770853853842974,
0.00018951768315652996,
0.016751048518719877,
0.006963479051823973,
0.0011095325953188939,
-0.042584387403119685,
0.0011095325953188939,
0.0063773555708460305,
-0.010671692222335267,
0.00021374144613721358,
0.00013924141413063331,
0.00021458418425482435,
0.0,
8.6685362155306e-05,
0.0,
8.760473689350791e-06,
-0.0034107533039760144,
0.003959066639547791,
0.0,
-0.024237090102958547,
0.9066155340223295,
0.0002084417469208087,
-0.055232978729472526,
-0.08815293475486005,
-0.0027757131775339554,
0.0,
0.0001657105116844669,
0.021303042880614045,
-0.006604719795869233,
-0.025885426926921487,
0.0046814020976034595,
-0.025885426926921487,
-0.025885426926921487,
-0.025885426926921487,
0.0,
3.7839004339281655e-05,
0.0,
0.035300881965462313,
0.008545277830765953,
0.05671876779068503,
6.557973768602571e-05,
0.0001783241751799189,
0.0005403175365818062,
-0.16737063489397686,
-0.08169978921996428,
0.0,
0.010874455720781431,
0.013123476860313954,
0.0,
0.0,
4.124199165129945e-05,
0.0,
0.0033721417403604573,
-0.00997395647469526,
0.0078102312541796,
0.0,
0.0011608371602290666,
0.00010027291157151005,
0.00030440460904834646,
0.0,
0.03396284808279027,
-0.004423671155114804,
0.003212643105493077,
0.0001038131590523815,
0.00040925357153781967,
0.061851787155847375,
0.0012214570512148978,
0.0,
-0.6431200942154551,
0.03641223368269397,
0.00036811502822692245,
6.434651682188196e-05,
0.0,
0.00038383834369307314,
-0.004523766238016968,
-0.07147941030085932,
0.0025032031539048145,
2.528964796889188e-06,
0.0,
0.0,
0.010395856416308975,
0.006976938866743201,
0.06404506257529431,
0.0002823390771846655,
0.0002823390771846655,
0.0,
0.00016480253720973996,
4.0996371940885905e-05,
0.0,
0.061851787155847375,
0.0006334493124915972,
-0.16737063489397686,
0.0,
0.00018951768315652996,
0.0001473985776355583,
-0.0013128528115502278,
-0.12282419187468549,
0.0,
0.0,
0.019489530989475812,
0.0032186703864789646,
9.566132758935807e-05,
0.012422766844719274,
0.2975009586924287,
    -0.04986289777554524,
    -0.013828990114694304,
    -0.036021576412235454,
    0.058958252040550815,
    -0.07554093893815765,
    -0.16212726414022408,
    0.00018951768315652996,
    0.0007616841978360522,
    -0.03668573607068677,
    4.2862453773511866e-05,
    6.974180741407773e-05,
    0.02079171283261795,
    -0.15010686907395843,
    -0.0008095853249623609,
    -0.0008095853249623609,
    0.006811564131851786,
    -0.010910387253267956,
    0.0036516839796621155,
    0.0036516839796621155,
    0.00018297852923721764,
    0.007281296585104129,
    2.7758127045470536e-05,
    -0.11926449906651372,
    0.057757518160486285,
    -0.16737063489397686,
    0.005423629973127209,
    0.00015092945251048511,
    6.735256101786221e-05,
    0.008577126001520377,
    0.0005194181429580365,
    0.0,
    3.342586624401128e-05,
    3.112349100320315e-05,
    -0.16022863279548696,
    0.11810403903059813,
    2.7954641066599802e-05,
    0.006328066235872159,
    0.0,
    0.0011044115568789744,
    0.0,
    0.0006867626323845616,
    0.023059528673896042,
    0.0004766633025656309,
    0.0,
    0.0,
    0.0002772314547838526,
    0.0007845040827010904,
    8.760473689350791e-06,
    0.00036207402167716504,
    0.0,
    0.0,
    0.0013009422148379428,
    0.05307276927431416,
    0.3804044137021817,
    0.0001692855352907126,
    0.0001692855352907126,
    0.00041534187462561917,
    0.0009326028234628342,
    0.0021466962938869586,
    0.061040078515634914,
    0.0001283985377092631,
    0.012048175849524754,
    0.004293392587773917,
    -0.022655188059782127,
    0.0021466962938869586,
    -0.10267197743266705,
            0.0015550763329311627,
            -0.10281339925017788,
            0.0,
            0.008313314170589425,
            -0.09679379596594956,
            0.0013015479856542506,
            0.0,
            0.00036632296890701 63,
            0.009517777036495652,
            -0.04064231158221523,
            7.000079235582695e-05,
            0.0,
            7.441143296767148e-05,
            0.4940554612878257,
            -0.24892223048889334,
            0.07377010389536008,
            0.22999048223341234,
            0.0012572730260441455,
            0.007082047353734795,
            0.003513305495873039,
            0.014217060392273086,
            -0.0010609098230108579,
            0.014037781384891573,
            0.0,
            4.0041044494724127e-05,
            0.0007655182319838198,
            0.033345902947246105,
            0.07949723195004774,
            0.0003555693110993872,
            0.06865532475951898,
            0.005416033381116211,
            -0.003596497250233907,
            0.017029121814739784,
            0.0,
            0.021412183205313674,
            -0.0014517292962967715,
            -0.0006583641295620777,
            -0.1391695172973164,
            0.0,
            0.022860310601228104,
            0.005362570771526081,
            0.008463131410352752,
            0.022731228744121888,
            0.05881592246689155,
            0.0,
            0.1372577907726808,
            0.001487510251199516,
            0.06459652825662539,
            0.0016102991202777872,
            -0.00948595601948074,
            0.16802440944518512,
            0.0,
            0.0044223921026904115,
            -0.0008823832144304812,
            0.0008539153635306962,
            0.00018951768315652996,
            -0.4659975361340828,
            0.025495331091741593,
            0.00039904980510033287,
            -0.3991826220685127,
            0.0,
            -0.03283282194515345 4,
            -0.15706246582594788,
            0.003933539260735823,
            0.0,
            0.002647664992121399,
0.0594827388968887,
0.0594827388968887,
0.0,
0.00017947637656000537,
-0.0478065078073136 3,
-0.2320990014892743,
0.003776128431694153,
-0.1665436393017184 2,
0.00010518901111667426,
0.0927486004988428,
0.0077827461161098 6,
0.008353687887113298,
-0.02834635683262855,
0.007261147620764525,
0.00032842292135523324,
-0.2224087152756708,
4.108946467535917e-05,
-0.18119798848845134,
-0.05673401367909314,
0.0283982418842849 4,
0.0008697964213969323,
0.0009110784816995746,
0.0008174237330657645,
0.008364688954463207,
-0.0052135622515392 55,
-0.0517708538538429 74,
-0.06933892084822908,
-0.4176917358202828,
-0.06974877789453685,
0.0023429744660829637,
0.004134268416824309,
0.0,
0.0003091186632232153,
0.01538121617952259 8,
-0.025885426926921487,
0.0020621681721285016,
0.10178461333738233,
0.011284669141779682,
0.0003398041010710006,
-0.0034416955839726812,
0.0,
-0.0013558650904924342,
0.0,
-0.02593028195719666,
0.0,
-0.5139862230095052,
0.0814821622433558 2,
-0.06933661929840833,
-0.06686093025067509,
0.0,
0.00013085752564936 33,
-0.15281554426976116,
0.07684914338497001,
0.0035350985737356766,
-0.12671397952013255,
0.012363418343417424,
0.10740982375532729,
-0.013672591579686482,
0.11151091670113096,
0.0,
0.0,
0.0,
0.0,
0.0,
0.0015820158614780 64,
0.0,
-0.16088858363778386,
0.0012750167199168925,
0.003200143490708276,
0.003400292342564077,
0.001582015861478064,
0.0,
0.0,
0.0,
0.0,
0.0,
0.0,
3.649533807159661e-05,
0.013551730469097262,
-0.12088519878788996,
0.0,
0.0,
0.0,
0.0,
0.0,
-0.0015843112073876482,
0.0,
0.0,
-0.011268439901356203,
0.08198487334606888,
-0.03664289319614327,
0.0034243974241600916,
0.0,
0.0754239118485914,
0.0005903003172317285,
0.004850625570601027,
-0.1267571668895787,
0.0005044129472790322,
-0.1661811641237111,
-0.059596617546063074,
0.003329185382044313,
0.0,
-0.04172474013272207,
0.0,
0.008745701254663785,
0.00909113432818426,
-0.05427151370721145,
0.0040638224751178625,
0.0017284280576792168,
-0.000749600507948465,
0.09627728218201026,
0.0003702409731428432,
-0.0661318903256582,
0.3178778689603336,
0.007384992155266606,
0.0,
0.0,
-0.12685793331942696,
0.0,
0.0,
0.01983712080337063,
-0.18745814886029294,
0.0009881881131721944,
3.125513387139209e-05,
-0.05053788638711969,
0.012294274603744817,
0.1873009584042753,
0.0012811325269893325,
0.7184009744441421,
0.0,
0.003474697297085999,
0.007739402269759765,
-0.08272911504508855,
0.0004641092144880202,
-0.3336381992488732,
0.0001776121747603638,
0.009594707581338922,
0.003971812256506529,
0.00018920684132803334,
0.0,
-0.18774772390150193,
0.017512748019622184,
5.883761684184756e-05,
0.00030353268150145414,
4.9790726426521145e-05,
0.3413882512268582,
0.0011890922779350414,
0.007287634500540897,
-0.006258430289020056,
0.0,
0.00030525209720648,
0.001712023676927424,
0.0,
0.19327472349239544,
0.2386747938060893,
1.9051625411953438e-05,
-0.018144752528760272,
0.3201297432176645,
0.24065060997593973,
0.002501412873241371,
0.002349323725785241,
0.027281319625330275,
0.0,
0.0,
-0.08849137448474935,
-0.006627180156281916,
-0.003313590078140958,
0.025372604546655977,
0.0007616841978360522,
-0.025885426926921487,
-0.0510091696560075,
-0.10808140335601496,
0.0038742611800690802,
-0.02323525929942248,
-0.0002588858307134615,
0.03564261689588727,
0.1392795353201773,
0.0,
0.0009694559605410748,
0.0007933488057895119,
0.0007616841978360522,
0.13589305879434946,
-0.24359718283346382,
-0.21266525606662287,
0.015356015923640942,
-0.025885426926921487,
-0.23152307303444525,
0.003127281202243752,
-0.1013746102431164,
-0.05664869523901477,
-0.08330588125009544,
-0.05664869523901477,
-0.0013128528115502278,
0.15945422995220926,
-0.025885426926921487,
-0.0015554371594113901,
0.4313914281898574,
-0.022588969266654314,
0.0007790657187002492,
0.0,
0.05344863131330327,
-0.24062782500590485,
-0.03706437710391956,
0.05445529004449336,
0.006407572207525267,
0.07279755636617047,
0.07171931932282413,
5.5041173193547784e-05,
0.07833878101275907,
0.00010532563236821495,
0.065899694976859,
-0.20299677775132383,
0.0038426050220965392,
0.04027507886592835,
0.00046511939751552544,
0.0007874606392444323,
0.009399156147295733,
-0.09377453738920752,
0.06954632782309572,
0.0,
-0.1353581829515719,
0.005362570771526081,
0.0,
0.048431309913007245,
0.0,
-0.0015843112073876482,
0.00030035192065512725,
-0.4342033485948993,
7.207646898520568e-05,
0.014952997352784363,
-0.011951803236777374,
0.10058035716184917,
-0.11864493861283337,
0.4374157965520749,
-0.1338060572622142,
0.00539106051088071,
0.0010024803892364892,
0.021116500226052586,
0.11287947553860994,
0.0007111437844232945,
0.0010267567910217754,
0.0017973447395533187,
0.18497007650225047,
-0.028915393792978486,
0.012486471638120163,
0.0,
0.002315385735962438,
0.056040864868404865,
0.002098325613846108,
-0.0500458847187891,
-0.054786786261679585,
-0.3061349291066769,
-0.014134014967256912,
0.002273562243569995,
-0.1856113015727068,
0.03239839408608166,
0.024397738019823013,
0.0003208258653857739,
0.21702919653841968,
0.04536629845694562,
0.0009104419827485629,
-0.09376485952554495,
-0.19661358886545494,
0.2959570180028976,
-0.012287121503710547,
-0.00038353285834889366,
-0.11612097869893014,
-0.0058739509439897355,
0.0020779378323654875,
0.0,
-0.025885426926921487,
-0.025885426926921487,
-0.0013204127567550967,
-0.15360777447828694,
0.0,
-0.0013128528115502278,
0.015399679710008171,
-0.0027757131775339554,
0.11715627139239598,
0.06614972004326017,
0.0006464794780697073,
0.1299136482377238,
0.13646486674493552,
0.0019205213657800895,
-0.028234178025395483,
0.019621038071822465,
0.7120658129735976,
8.774810068052795e-05,
0.0927486004988428,
0.00011510754131541298,
-0.15097890422691515,
0.04045461320524046,
-0.02598057490872505,
-0.0034416955839726812,
0.015416592090762597,
0.11802305863413734,
0.13894311197047104,
0.00015424302055294683,
0.037131106190946915,
0.023810639403281823,
-0.0780290760203204,
0.0,
0.01705084635103099,
0.00015424302055294683,
-0.12941179292821925,
-0.0530441269977418,
0.00015424302055294683,
0.00012565389356939224,
0.0002966423416173455,
0.0002966423416173455,
0.053336547174097415,
0.005431346020460212,
-0.0954102167931453,
0.0,
0.05916784530238596,
0.006569297043488789,
0.10734549622872366,
0.00026498739109549757,
3.841759522614608e-05,
-0.18221590399486534,
0.0004758901120045131,
9.239480440807413e-06,
0.007473244795001668,
-0.18248358382824714,
0.0,
0.09519698269270555,
0.0005246556847883812,
0.23058351054377998,
0.014031848793688762,
0.016207990748090543,
-0.005713659189446577,
0.018363118439591743,
0.00030798626723986067,
0.002091483659599026,
0.0001395698309413747,
6.263470333002105e-05,
0.016207990748090543,
-0.011006222637718615,
0.13056663418758555,
-0.07544228665523293,
0.039221371690318016,
0.014015299261319116,
-0.045953696278747944,
0.08469703612034749,
0.01290972023918657,
-0.2615339974424256,
-0.04898165340223159,
0.0,
-0.04898165340223159,
-0.04376410590694129,
0.0,
0.007052524682390392,
-0.07544228665523293,
0.011331043745979413,
0.0026694540247732094,
3.112349100320315e-05,
0.0009874623018016925,
0.015073522982079202,
-0.1567162371355314,
0.012864081898100531,
-0.13490691152780596,
0.0,
0.0033469992676784153,
0.03556935798817773,
0.44341745359769374,
0.00046141916692079535,
0.0004624250974582319,
0.06614972004326017,
0.0002160117319161011,
0.00044734648144289824,
0.013449754668087602,
0.001026515163926333,
0.008745701254663785,
-0.0838725735534372,
0.0,
0.0005187501570205081,
0.0007153629428645814,
-0.16212726414022408,
0.00011093908573282648,
0.0017899279825726775,
-0.23375779348409376,
0.023457693444362902,
0.00306632179412664,
0.0002721545370175367,
0.0001887030173798397,
0.0008327900761448192,
0.006423297208325694,
-0.04409152893203348,
0.5145617292622782,
0.0,
-0.35220797809897636,
-0.10539493937808017,
-0.29102164572475014,
-0.08195492726696288,
0.0,
0.09484857472318621,
0.001345236333688458,
0.04514961491362601,
0.1198014607702l536,
4.605663772959598e-05,
-0.16700714826372526,
0.005782352156648206,
-0.09429052611194237,
0.00011222140544471545,
0.3208040299708211,
0.006541971146432992,
0.008354584164121713,
-0.061267695943690095,
-0.00020264373440365396,
0.0018299074496781398,
0.010484067238948314,
0.003998617495322853,
-0.03561979675236193,
0.09979909878648437,
-0.0005134378432158695,
0.0584926113238522,
0.00013384157842623046,
0.0013017480607246743,
0.0007073710569948817,
0.0010153871392487408,
-0.004184947600654325,
0.04883752554808659,
0.040356469318901037,
0.0,
0.04106652454103969,
-0.23206826833572833,
-0.016523114573714442,
0.007328764761542015,
0.0021244155077769283,
-0.3942029124363838,
0.578790295650524,
-0.06207456306615545,
0.03262127214083565,
0.1660527556918404,
0.1116779522270919,
0.0005353301361368479,
0.00012478753002080818,
0.0003154126077309339,
0.00340922634442647,
0.0,
-0.012498901102310585,
-0.012498901102310585,
-0.2999407914962728,
0.00016928553529071 26,
0.0006306859403950085,
0.00010562761761324276,
0.015781065928121916,
0.1169923650789755,
0.013947703697871143,
0.024433108182157007,
0.04914336804656683,
0.0,
0.13183206418167023,
-0.08415278597372103,
0.0011892838417381117,
-0.06324755582446438,
0.005089328788786816,
-0.02156369665225641,
0.005845496017254438,
0.0,
0.0,
0.0,
0.1677502214878836,
0.0008772101346785263,
4.404733110587494e-05,
0.0,
0.026765714754870138,
0.0726893100754010 8,
-0.27018881436238595,
0.0,
4.0464637029431407e-05,
-0.12628259910401027,
-0.12628259910401027,
0.0004513137921313412,
0.06916669071487422,
0.00031807129417286517,
0.0,
0.007485527348695209 5,
0.0,
-0.04373814478387655,
0.0007395633252994783,
0.00027484833790438504,
0.011406073442554879,
-0.010685054110515426,
0.0,
-0.00028454958224629153,
0.0,
4.033054036237529e-05,
-0.16677447787497907,
3.3676819984587745e-05,
0.0,
0.0,
-0.029061060347097045,
0.0,
-0.010107851451783301,
0.0,
0.00740320982690904,
0.00740320982690904,
0.0,
2.53162084566882e-05,
-0.6080063246834312,
0.16923072940472542,
0.004491533739805325,
...]
```

In [42]:

```
coef_data=pd.DataFrame({"Word":w,"Coefficent":coef})
coef_data
```

Out[42]:

	Word	Coefficent
0	00	-0.377749
1	000	0.043005
2	0000	0.190629
3	000001	-0.008009
4	00001	0.000000
...	...	...
114964	çaykur	0.000841
114965	çelem	-0.129305
114966	être	0.000000
114967	île	0.010643

114968	ît	0.000395

114969 rows × 2 columns

In [43]:

```
coef_data=coef_data.sort_values(["Coefficent","Word"],ascending=False)
coef_data
```

Out[43]:

	Word	Coefficent
80600	pleasantly	4.131702
39072	downside	3.485272
94667	skeptical	2.846508
5865	addicting	2.747423
113138	worries	2.616647
...	...	...
88945	ripoff	-3.051698
113164	worst	-3.114617
106852	unacceptable	-3.375093
34989	deceptive	-3.545303
107383	undrinkable	-3.893301

114969 rows × 2 columns

In [44]:

```
coef_data.head(20)
```

Out[44]:

	Word	Coefficent
80600	pleasantly	4.131702
39072	downside	3.485272
94667	skeptical	2.846508
5865	addicting	2.747423
113138	worries	2.616647
35726	delish	2.534530
39214	drawback	2.339607
55029	hooked	2.293970
87967	resist	2.200129
68460	met	2.137388
111911	whim	2.092416
35691	delighted	2.054933
103080	thankful	2.049291
102067	tastiest	2.046909
44711	fav	2.033807
19488	beat	2.030245

10992	awesome	2.025304
43222	excellent	2.023071
91064	saves	2.011759
111520	welcome	1.996266

In [45]:

```
coef_data.tail(20)
```

Out[45]:

	Word	Coefficent
21288	blech	-2.298886
35220	defeats	-2.318371
89868	ruins	-2.322227
37563	disappointment	-2.365395
89864	ruined	-2.370502
62574	lame	-2.421404
76576	overpowers	-2.486196
67898	mediocre	-2.504172
65064	lousy	-2.586144
62401	lacked	-2.768649
88351	returnable	-2.807663
41118	embarrassed	-2.825590
37528	disapointed	-2.862184
24949	cancelled	-2.953761
37560	disappointing	-2.988353
88945	ripoff	-3.051698
113164	worst	-3.114617
106852	unacceptable	-3.375093
34989	deceptive	-3.545303
107383	undrinkable	-3.893301

Automate NLP and Machine learning model

In [46]:

```
def text_fit(X,Y,nlp_model,ml_model,coef_show=1):
    count_vec_X = nlp_model.fit_transform(X)
    print("features : {}".format(count_vec_X.shape[1]))
    X_train,X_test,Y_train,Y_test = train_test_split(count_vec_X,Y)
    ml=ml_model.fit(X_train,Y_train)
    acc=ml.score(X_test,Y_test)
    print(acc)

    if coef_show==1:

        w =count_vec.get_feature_names()
        coef=log_reg.coef_.tolist()[0]
        coef_data=pd.DataFrame({"Word":w,"Coefficent":coef})
        coef_data=coef_data.sort_values(["Coefficent","Word"],ascending=False)
        print("\n")
        print("Top 20 positive words")
        print(coef_data.head(20))
        print("\n")
        print("Top 20 negative words")
        print(coef_data.tail(20))
```

In [47]:

```
from sklearn.feature_extraction.text import CountVectorizer
```

In [48]:

```
c=CountVectorizer(stop_words="english")
```

In [49]:

```
from sklearn.linear_model import LogisticRegression
```

In [50]:

```
text_fit(X,Y,c,LogisticRegression())
```

```
features : 114969
0.9362666788382248


Top 20 positive words
              Word  Coefficent
80600   pleasantly    4.131702
39072     downside    3.485272
94667    skeptical    2.846508
5865     addicting    2.747423
113138     worries    2.616647
35726       delish    2.534530
39214     drawback    2.339607
55029       hooked    2.293970
87967       resist    2.200129
68460          met    2.137388
111911        whim    2.092416
35691    delighted    2.054933
103080    thankful    2.049291
102067    tastiest    2.046909
44711          fav    2.033807
19488         beat    2.030245
10992      awesome    2.025304
43222    excellent    2.023071
91064        saves    2.011759
111520     welcome    1.996266


Top 20 negative words
                  Word  Coefficent
21288            blech   -2.298886
35220          defeats   -2.318371
89868            ruins   -2.322227
37563   disappointment   -2.365395
89864           ruined   -2.370502
62574             lame   -2.421404
76576       overpowers   -2.486196
67898         mediocre   -2.504172
65064            lousy   -2.586144
62401           lacked   -2.768649
88351       returnable   -2.807663
41118      embarrassed   -2.825590
37528      disapointed   -2.862184
24949        cancelled   -2.953761
```

```
37560    disappointing   -2.988353
88945           ripoff   -3.051698
113164           worst   -3.114617
106852    unacceptable   -3.375093
34989        deceptive   -3.545303
107383      undrinkable   -3.893301
```

Automate predictions

In [51]:

```python
from sklearn.metrics import confusion_matrix,accuracy_score
def predict(X,Y,nlp_model,ml_model):
    count_vec_X = nlp_model.fit_transform(X)
    X_train,X_test,Y_train,Y_test = train_test_split(count_vec_X,Y)
    ml=ml_model.fit(X_train,Y_train)
    predictions =ml.predict(X_test)
    cm = confusion_matrix(predictions,Y_test)
    print(cm)
    acc= accuracy_score(predictions,Y_test)
    print(acc)
```

In [52]:

```python
c= CountVectorizer()
lr = LogisticRegression()
```

In [53]:

```python
predict(X,Y,c,lr)
```

```
[[ 15439   2808]
 [  5115 108092]]
0.9397279656762061
```

Apply different NLP and machine learning models

In [54]:

```python
from sklearn.dummy import DummyClassifier
```

In [55]:

```python
d=DummyClassifier()
c=CountVectorizer()
```

In [56]:

```python
text_fit(X,Y,c,d,0)
```

```
features : 115282
0.8443333789766763
```

In [57]:

```python
from sklearn.feature_extraction.text import TfidfVectorizer
```

In [58]:

```python
tfd =TfidfVectorizer(stop_words = "english")
```

In [59]:

```python
lr= LogisticRegression()
```

In [60]:

```python
text_fit(X,Y,tfd,lr,0)
```

```
features : 114969
0.9346463401646203
```

Data prepration and Modeling purpose when score is 5

In [61]:

```python
data.head()
```

Out[61]:

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	
0	1	B001E4KFG0	A3SGXH7AUHU8GW	delmartian	1	1	5	130

1	2	B00813GRG4	A1D87F6ZCVE5NK	dll pa	0	0	1	134
2	3	B000LQOCH0	ABXLMWJIXXAIN	Natalia Corres "Natalia Corres"	1	1	4	121
3	4	B000UA0QIQ	A395BORC6FGVXV	Karl	3	3	2	130
4	5	B006K2ZZ7K	A1UQRSCLF8GW1T	Michael D. Bigham "M. Wassir"	0	0	5	135

In [64]:

```
data=data[data["Score"]==5]
```

In [65]:

```
data["%upvote"].unique()
```

Out[65]:

```
['80-100%', NaN, '60-80%', 'Empty', '40-60%', '20-40', '0-20%']
Categories (6, object): ['Empty' < '0-20%' < '20-40' < '40-60%' < '60-80%' < '80-100%']
```

In [67]:

```
data =data[data["%upvote"].isin(['80-100%', '60-80%', '20-40', '0-20%'])]
```

In [70]:

```
X = data["Text"]
```

In [71]:

```
data["%upvote"].unique()
```

Out[71]:

```
['80-100%', '60-80%', '20-40', '0-20%']
Categories (6, object): ['Empty' < '0-20%' < '20-40' < '40-60%' < '60-80%' < '80-100%']
```

In [72]:

```
y_dict = {'80-100%':1, '60-80%':1, '20-40':0, '0-20%':0}
```

In [73]:

```
Y = data["%upvote"].map(y_dict)
```

In [74]:

```
Y.value_counts()
```

Out[74]:

```
1.0    151721
0.0      2707
Name: %upvote, dtype: int64
```

In [75]:

```
from sklearn.feature_extraction.text import TfidfVectorizer
```

In [76]:

```
tf = TfidfVectorizer()
```

In [78]:

```
X_c=tf.fit_transform(X)
```

Handeling Imbalance Data

In [92]:

```
from imblearn.over_sampling import RandomOverSampler
```

In [93]:

```
os = RandomOverSampler()
```

In [95]:

```
os.fit(X_c,Y)
```

Out[95]:

```
RandomOverSampler()
```

In [100…

```
X_resampled,Y_resampled =os.fit_resample(X_c,Y)
```

In [101…

```
print(X_resampled.shape,Y_resampled.shape)
```

```
(303442, 67507) (303442,)
```

In [102…

```
from collections import Counter
```

In [106…

```
print("Orignal dataset shape {}".format(Counter(Y)))
print("Resampled dataset shape {}".format(Counter(Y_resampled)))
```

```
Orignal dataset shape Counter({1.0: 151721, 0.0: 2707})
Resampled dataset shape Counter({1.0: 151721, 0.0: 151721})
```

Cross validation

In [107…

```
from sklearn.linear_model import LogisticRegression
```

In [108…

```
log=LogisticRegression()
```

In [111…

```
np.arange(-2,3)
```

Out[111]:

```
array([-2, -1,  0,  1,  2])
```

In [112…

```
grid ={"C": 10.0**np.arange(-2,3),"penalty":["l1","l2"]}
```

In [110…

```
from sklearn.model_selection import GridSearchCV
```

In [113…

```
clf= GridSearchCV(estimator=log,param_grid =grid,cv=5,n_jobs=-1,scoring="f1_macro")
```

In [114…

```
clf.fit(X_resampled,Y_resampled)
```

Out[114]:

```
GridSearchCV(cv=5, estimator=LogisticRegression(), n_jobs=-1,
             param_grid={'C': array([1.e-02, 1.e-01, 1.e+00, 1.e+01, 1.e+02]),
                         'penalty': ['l1', 'l2']},
             scoring='f1_macro')
```

In [115…

```
from sklearn.model_selection import train_test_split
```

In [116…

```
X_train,X_test,Y_train,Y_test = train_test_split(X_c,Y)
```

In [117…

```
pred = clf.predict(X_test)
```

In [122…

```
from sklearn.metrics import confusion_matrix,accuracy_score
```

In [121…

```
confusion_matrix(Y_test,pred)
```

Out[121]:

```
array([[  688,     0],
       [  559, 37360]], dtype=int64)
```

In [123…

```
accuracy_score(Y_test,pred)
```

Out[123]:

```
0.9855207604838501
```

In [ ]:

```

```