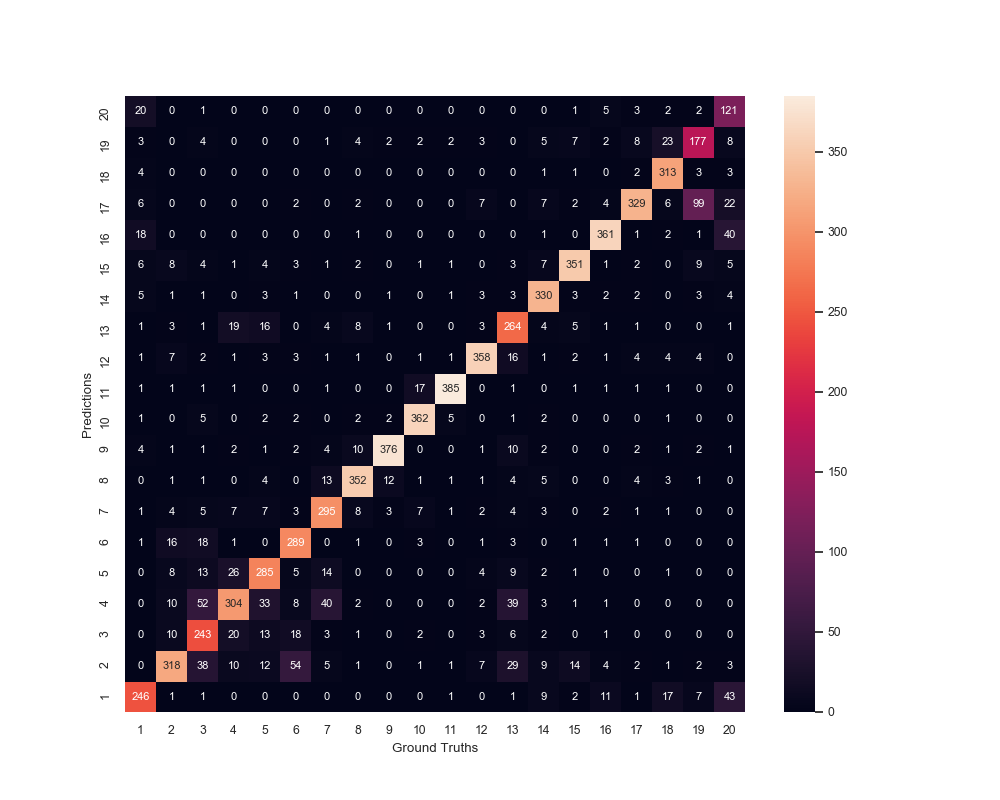
Andrew Sanders

Machine Learning Homework 2

2/14/20

Question 1:

|  |  |
| --- | --- |
| α | Test Accuracy |
| 1 | 0.807328448 |
| 0.1 | 0.810526316 |
| 0.01 | 0.802398401 |
| 0.001 | 0.790939374 |
| 0.0001 | 0.781612258 |
| 0.00001 | 0.777348434 |

Question 2: The algorithm seems to confuse the comp and talk.politics newsgroups (not with eachother, but within their own subgroups). I believe these topics are very similar in nature as compared to other categories. For example, there is very little confusion between sci.med and sci.space due to both topics being in completely different areas of study. This is opposed to comp.sys.ibm.pc.hardware and comp.sys.mac.hardware, or talk.politics.mideast and talk.politics.misc which essentially cover the same topic.

Question 3: The purpose of additive smoothing is to both include our preconceived answer in the probabilities, and to prevent multiplying by zero. When the alpha is too little, the calculated probabilities become multiplied by almost zero, which can heavily distort the probability for a series of words. When the alpha is too big, the calculated probability reaches closer to a flat distribution rather than the actual probability.