Dataset Description

Topic	Training	Test
Atheism	480	319
Graphics	584	389
MSwindows	572	393
PC	590	392
Mac	578	385
Xwindows	593	392
Forsale	585	390
Autos	594	395
Motorcycles	598	398
Baseball	597	397
Hockey	600	399
Cryptology	595	396
Electronics	591	393
Medicine	594	396
Space	593	394
Christianity	598	398
Guns	545	364
MideastPolitics	564	376
Politics	465	310
Religion	377	251
Totals	**11293**	**7527**

- The file "`forumTraining.data`" is the training dataset for your Naïve-Bayes classifier. It has 11293 lines. Each line is a full document. The first 480 lines consist of 480 documents that have been tagged with the category "Atheism", and so on. There are 20 categories. The first word of each line represents the tag, or category.

- The file "`forumTest.data`" is the test dataset to see how well your classifier performs on documents it has not seen. It follows an identical format and consists of 7527 documents. So, for example, achieving 50% accuracy with your classifier would mean that you had correctly classified ~3764 documents from the test set (random assignment would expect 5% accuracy).

- The two "stemmed" files are versions of the original two files which have been pre-processed as described in the project specifications.