Vincent Chee
Natural Language Processing
Professor Kristina Striegnitz
22/05/15

Problem Set 6

Question 2

1. How many reviews are in the data set?
There are 11914 reviews in the data set.

2. What is the total number of words in all reviews?
There is 1758733 words in all reviews.

3. What are the different product categories?
Camera, books, software, music, health and DVD.

Question 4

1. The classifier uses everything but the first 2000 reviews for training, so the last 9914
2. The classifier uses the first 2000 reviews for training.
3. Line 44 is list comprehension and it extracts all the features of the reviews from all the reviews in the file.

Question 5

2. This classifier has a 0.534 accuracy using the NaiveBayes classifier.
3. I will examine the most informative features for each of the six categories. The show_most_informative_features shows the 20 top ranked features according to the ratio of one label to another, for example if there are five times as many positive documents containing this word as negative ones, it will have a 5.0:1.0 pos:neg ratio. The most informative feature for software given by this classifier was if the review contained the word 'learning', it has a 64.1:1.0 pos:neg ratio with music. The most informative feature for cameras were reviews containing the word 'video', which makes sense, it has a 51.0:1.0 pos:neg ratio with health. The most informative feature for music were reviews containing the word 'music', it has a 44.7:1.0 pos:neg ratio with camera. DVD's most informative feature was if it contained the word 'watched', it had a ratio of 44.4:1.0 pos:neg with camera. Health's most informative feature was the word 'cream' and had a ratio of 24.3:1.0 pos:neg with DVD, this doesn't make quite as much sense as the previous 4. Books' most informative feature was if the review contained the word 'research' which is not that informative, it has a 23.9:1.0 pos:neg ratio. Some features that might be problematic is the word 'saying' and 'parents' in DVD reviews, however these aren't that problematic given that most of the other informative made sense and there is nothing which is utterly problematic.

Results:

```
Total number of docs: 11914
Total number of words: 1758699
Categories: ['CAMERA', 'BOOKS', 'SOFTWARE', 'MUSIC', 'HEALTH', 'DVD']
0.534
Most Informative Features
    contains(learning) = True          SOFTWA : MUSIC  =     64.2 : 1.0
       contains(video) = True          CAMERA : HEALTH =     51.0 : 1.0
       contains(music) = True           MUSIC : CAMERA =     44.7 : 1.0
     contains(watched) = True             DVD : CAMERA =     44.4 : 1.0
      contains(photos) = True          CAMERA : MUSIC  =     35.3 : 1.0
        contains(role) = True             DVD : HEALTH =     30.0 : 1.0
   contains(installed) = True          SOFTWA : BOOKS  =     29.7 : 1.0
      contains(saying) = True             DVD : CAMERA =     28.2 : 1.0
        contains(belt) = True          CAMERA : HEALTH =     27.7 : 1.0
    contains(internet) = True          SOFTWA : MUSIC  =     26.0 : 1.0
       contains(cream) = True          HEALTH : DVD    =     24.3 : 1.0
    contains(research) = True           BOOKS : MUSIC  =     23.9 : 1.0
       contains(learn) = True          SOFTWA : MUSIC  =     23.1 : 1.0
    contains(episodes) = True             DVD : BOOKS  =     22.8 : 1.0
     contains(shutter) = True          CAMERA : DVD    =     22.0 : 1.0
        contains(rich) = True           BOOKS : SOFTWA =     21.6 : 1.0
     contains(parents) = True             DVD : CAMERA =     20.9 : 1.0
 contains(complicated) = True          SOFTWA : MUSIC  =     20.4 : 1.0
     contains(singles) = True           MUSIC : DVD    =     20.4 : 1.0
       contains(world) = True           BOOKS : HEALTH =     20.1 : 1.0
```