CSC 206: NLP
Spring 2015

Name (Print): _____

Problem Set 6: Document Classification

Due: Friday, May 22 2015, before class

Please create a folder for this problem set, and save all files you need to submit in this folder. Zip up the folder and submit the .zip file on Nexus.

Questions:

1. Getting started

   Download the starter code and data from Nexus and unzip it.

   The data set is a collection of product reviews. Each line in the text file contains first a label indicating the category of the product being reviewed and then the review text. The text has already been cleaned up by lower-casing all words and separating punctuation from words.

   You task is going to be to write/complete Python code for training a classifier that can take a review and determine the category of the product being reviewed.

   What follows is a mix of code reading comprehension and exploration questions as well as some questions that ask you to write some code.

   Please answer all code reading comprehension and exploration questions in writing in one separate document. In the end save/export this document to pdf and include the pdf in your submission.

2. Reading in the data

   Look at the first section of the code, that deals with reading in the data from the file. Read and understand the code. Run it and use further Python queries to answer the following questions.

   1. How many reviews are in the data set?
   2. What is the total number of words in all reviews?
   3. What are the different product categories? (Hint: there are six different categories.)

3. Feature extraction

   Complete the function `document_features` which takes a list of words representing a single review and returns a feature dictionary. (Look at the name classification and movie review classification code that you have seen in class for guidance.)

   Notes:

   1. Start by using the words that occur in the reviews as features.
   2. Start by using only the 2000 most frequent words as features.

3. Do not use stopwords as features.

4. Preparing the training and test sets

   Now look at and run the next section of the code. Answer the following questions.

   1. How many reviews are used for testing?
   2. How many reviews are used for training?
   3. Explain what happens in line 44:

```
docs_as_features = [(document_features(words), category)
                    for (words, category) in all_docs]
```

5. Training and evaluating the classifier

   1. Add code that trains a classifier, evaluates its accuracy, and prints out the most informative features.
   2. What is the accuracy of this classifier?
   3. Comment on the most informative features. Do they make sense? Do you see any features that might be problematic?

What to Submit:

- the Python file `topic-classify.py` with your additions
- a pdf document containing the answers to all questions

Please do **NOT** include the data file in your submission.