LING 572: Hw1
Due: 11pm on 1/11 (Wed)
Total points: 100

The goal of this assignment is to use the Mallet package for the text classification task. All the data files are under dropbox/16-17/572/hw1/ on patas. Let $dataDir be hw1/20_newsgroups, and $exDir be hw1/examples/. Note:

- When you type the commands, you need to replace $dataDir with hw1/20_newsgroups and $exDir with hw1/examples.
- All the options of Mallet commands (e.g., "--input") start with two "-"s, not one "-".
- Use the Mallet package on Patas, which is the correct version for this assignment.

Q1 (20 points): Learning the Mallet commands

(a) 2 points: Check out Mallet website at http://mallet.cs.umass.edu/ and focus on the classification part. Go over the mallet slides at http://courses.washington.edu/ling572/winter2017/teaching_slides/class0-mallet.pdf and set up your PATH and CLASSPATH on patas properly.

(b) 2 points: Run the following command to create a data vector, **politics.vectors**, using the data from the three talk.politics.* newsgroups:

mallet import-dir --input $dataDir/talk.politics.* --skip-header --output politics.vectors

(c) 2 points: Run the following command to convert **politics.vectors** to the text format **politics.vectors.txt**.

vectors2info --input politics.vectors --print-matrix siw > politics.vectors.txt

(d) 2 points: Run the following command to split **politics.vectors** into training (90% of the data) and testing files (10% of the data):
vectors2vectors --input politics.vectors --training-portion 0.9 --training-file train1.vectors --testing-file test1.vectors

(e) 2 points: Run the following command to train and test. The training and test accuracy is at the end of dt.stdout.
vectors2classify --training-file train1.vectors --testing-file test1.vectors --trainer DecisionTree > dt.stdout 2>dt.stderr

(f) 10 points: Run vectors2classify to classify the data with five learners and complete Table 1.

- Use the train.vectors and test.vectors **under $exDir** for this classification task.
- The names of the five learners are: NaiveBayes, MaxEnt, DecisionTree, Winnow, and BalancedWinnow.
- The command for classification is:
  vectors2classify --training-file $exDir/train.vectors --testing-file $exDir/test.vectors --trainer $zz > $zz.stdout 2>$zz.stderr

  whereas $zz is the name of a learner (e.g., MaxEnt).

Table 1: Classification results for Q1(e)

	Training accuracy	Test accuracy
NaiveBayes		
MaxEnt		
DecisionTree		
Winnow		
BalancedWinnow		

Q2 (30 points): Write a script, **proc_file.sh**, that processes a document and prints out the feature vectors.

- The command line is: proc_file.sh input_file targetLabel output_file
- The input_file is a text file (e.g., **input_ex**).
- The output_file has only one line with the format (e.g., **output_ex**):
  instanceName targetLabel f1 v1 f2 v2 ....
  - The instanceName is the filename of the input_file.
  - The targetLabel is the second argument of the command line.
- To generate the feature vector, the code should do the following:
  - First, skip the header; that is, the text before the first blank line should be ignored.
  - Next, replace all the chars that are not [a-zA-Z] with whitespace, and lowercase all the remaining chars.
  - Finally, break the text into token by whitespace, and each token will become a feature.
  - The feature values will be the frequency of the sequences.
  - The (featname, value) pairs are ordered by the spelling of the featname.
- For instance, running "proc_file.sh $exDir/input_ex c1 output_ex" will produce output_ex as the one under the $exDir.

Q3 (30 points): Write a script, **create_vectors.sh**, that creates training and test vectors from several directories of documents. This script has the same function as "mallet import-dir", except that the vectors produced by this script are in the text format and the training/test split is not random.

- The command line is: create_vectors.sh train_vector_file test_vector_file ratio dir1 dir2 ...
  That is, the command line should include one or more directories.
- ratio is the portion of the training data. For instance, if the ratio is 0.9, then the FIRST 90% of the FILES in EACH directory should be treated as the training data, and the remaining 10% should be treated as the test data. By the first x%, we mean the top x% when one runs "**ls dir**".
- train_vector_file and test_vector_file are the output files and they are the training and test vectors in the text format (the same format as the output_file in Q2).

- The class label is the basename of an input directory. For instance, if a directory is hw1/20_newsgroups/talk.politics.misc, the class label for every file under that directory should be talk.politics.misc.

Q4 (20 points): Classify the documents in the talk.politics.* groups under $dataDir.

- Run create_vectors.sh from Q3 with the ratio being **0.9**, and the directories being talk.politics.guns, talk.politics.mideast, and talk.politics.misc.
- Run "**mallet import-file**" to convert the vectors from the text format to the binary format, and **vectors2classify** for training (with MaxEnt trainer) and for testing.
- Suppose you run "**mallet import-file**" first on train_vector_file and create train.vectors. When you run "**mallet import-file**" next on the test_vector_file, remember to use the option "--use-pipe-from train.vectors". That way, the two vector files will use the same mapping to map feature names to feature indexes.
- Save all the files (the vectors in text format and binary format, the MaxEnt model, the classification output) under a directory called **q4**. You can call the MaxEnt model file **me-model**, and the classification output **me.stdout** and **me.stderr**.
- What are the training and test accuracy?

Submission: In your submission, include the following:

- Shell scripts for proc_file.sh and create_vectors.sh, and the code called by the shell scripts.
- The subdirectory q4/ created in Q4.
- The note file that includes the following:
  - Table 1
  - Training and test accuracy in Q4
  - Any note that you want the grader to read
- No need to submit anything for Q1 except for Table 1 in the note file.