LING 572 Hw5 (MaxEnt decoder)
Due: 11pm on Feb 8, 2017

The example files are under /dropbox/16-17/572/hw5/examples/.

Q1 (5 points): Run the Mallet MaxEnt learner (i.e., the trainer's name is MaxEnt) with **train2.vectors.txt** as the training data and **test2.vectors.txt** as the test data.

- You can use *vectors2classify* or "*mallet train-classifier*" plus "*mallet classify-svmlight*".
- Save the model to a file called *q1/m1*.
- Convert the model into the text format with the following command: classifier2info --classifier q1/m1 > q1/m1.txt
- In your note file, write down the command you used, the training accuracy and the test accuracy.

Q2 (40 points): Write a MaxEnt classifier, called **maxent_classify.sh**, that classifies test data given a MaxEnt model learned from training data.

- The format is: maxent_classify.sh test_data model_file sys_output > acc_file
- test_data, sys_output, and acc_file have the same format as in Hw2-Hw4.
- model_file has the same format as q1/m1.txt created in Q1.
- Run "maxent_classify.sh test2.vectors.txt q1/m1.txt q2/res > q2/acc". What is the test accuracy? Is it the same as the test accuracy in Q1? Why or why not?

Q3 (15 points): Write a script, calc_emp_exp.sh, to calculate empiricial expectation.

- The format is: calc_emp_exp.sh training_data output_file
- training_data has the same format as before.
- output_file has the format "class_label feat_name expectation raw_count" (c.f. emp_count_ex): *raw_count* is the number of training instances with that class_label and contains that feat_name; *expectation* is the empirical expectation.
- Run "calc_emp_exp.sh train2.vectors.txt q3/emp_count" and include q3/emp_count in your submission.

Q4 (40 points): Write a script, calc_model_exp.sh, to calculate model expectation.

- The format is: calc_model_exp.sh training_data output_file {model_file}

- training_data has the same format as before.
- output_file has the format "class_label feat_name expectation count" (e.g., **emp_count_ex**): *expectation* is the model expectation; *count* is *expectation* multiplied by the number of training instances.
- model_file is optional. If it is given, it has the same format as in Q2 (e.g., q1/m1.txt) and it is used to calculate $p(y|x_i)$. If it is not given, $p(y|x_i) = 1/|C|$, where $|C|$ is the number of class labels.
- Run "calc_model_exp.sh train2.vectors.txt q4/model_count q1/m1.txt" and include q4/model_count in your submission.
- Run "calc_model_exp.sh train2.vectors.txt q4/model_count2" and include q4/model_count2 in your submission.

Submission: Submit a tar file via CollectIt. The tar file should include the following.

- If you work with a partner, one of you should submit all the files, and the other person submits only one message specifying his/her partner.
- All the files should be saved under hw5_dir/ or its subdirectories.
- In your note file hw5-notes.*, include your answers to Q1-Q4, and any notes that you want the TA to read.
- Shell scripts for Q2-Q4 and related source and binary code should be stored under hw5_dir/.
- The model and output files created in Q1-Q4 should be stored under hw5_dir/q[1-4]/ (i.e., q1/m1, q1/m1.txt, q2/res, q2/acc, q3/emp_count, q4/model_count, and q4/model_count2).