HW2

The task for Q1-Q3

- Three-way text classification:
  - talk.politics.guns
  - talk.politics.mideast
  - talk.politics.misc

- Training data: 2700 instances (900 per class)
- Test data: 300 instances (100 per class)

- Features: words

- Task:
  - Q1-2: run MALLET DT learner
  - Q3: build your own DT learner

Use MALLET

- mallet import-svmlight --input train.vectors.text --output train.vectors
  - Format of train.vectors.txt: label f1:v1 f2:v2

- Mallet train-classifier --input train.vectors --trainer MaxEnt --output-classifier model1
  - Trains MaxEnt classifierand stores model

- vectors2classify --training-file train.vectors --testing-file test.vectors --trainer DecisionTree --report test:raw test:accuracy test:confusion train:confusion train:accuracy > de1.stdout 2>de1.stderr

-

Q3: build a DT learner

- Each node checks exactly one feature

- Features are all binary: either present or not present
  - This DT is a binary tree

- Quality measure: information gain

Efficiency Issue

- To select best feature, need info gain for each
- Need counts for (c, f) and (c, not f) for each label c and feature f
- Try to be efficient!
- Report running time in Tables 2 and 3. (Order of minutes)

Patas usage

- When testing your code, use small data sets and small depth values first.
  - NB: data is sorted by class, so sample smartly!
- Use condor_submit for anything more than very simple testing
- Monitor jobs!
- For condor: https://www.shane.st/teaching/571/aut19/welcome_to_patas_1920.pdf