# Hw8

# Hw8

- Purpose:
  - Learn to use Mallet package
  - Learn to create feature vectors
- Text classification task
- Three categories: guns, mideast, and misc
- Each category has 1000 files under ~/dropbox/18-19/570/hw8/20_newsgroups/talk.politics.*/

# Data preparation

- Define features

- Create feature vectors for each training/test instance; save them in a text vector format

  ➔ write your own code

- Run “Mallet import-file” to convert the text format to binary format.

- Q1: use “mallet import-dir” to create feature vectors:
  - mallet import-dir –input dir1 –output vectorFile
  - The training/test data are stored under dir1/; files under the same subdir belong to the same class; each file is an instance.
  - vectorFile is the output vector file in the binary format

- Q2-Q4: the same task, but you need to prepare the vectors yourself.

# Features

Given a document

- Skip the header: use the text after the first blank lines.
- Replace any char that is not [a-zA-Z] with whitespace, lowercase everything, and break the lines into tokens by whitespace. These tokens are features.

  => This is different from the typical tokenization

- Feature values are the frequencies of the tokens in the document.

# An example: talk.politics.guns/53293

Xref: cantaloupe.srv.cs.cmu.edu misc.headlines:41568 talk.politics.guns:53293

…

Lines: 38

hambidge@bms.com wrote:

: In article <C4psoG.C6@magpie.linknet.com>, manes@magpie.linknet.com (Steve Manes) writes:

# After “tokenization”

hambidge@bms.comwrote:

:In article<C4psoG.C6@magpie.linknet.com>,
manes@magpie.linknet.com(SteveManes) writes:

hambidge bms comwrote

In article c psog c magpie linknet commanes
magpie linknet com stevemanes writes

# After lowercasing, counting and sorting

- talk.politics.guns/53293 guns a 11 about 2 absurd 1 again 1 an 1 and 5 any 2 approaching 1 are 5 argument 1 art icle 1 as 5 associates 1 at 1 average 2 bait 1 be 6 being 1 betraying 1 better 1 bms 1 by 5 c 2 calculator 1 capita 1 choice 1 chrissakes 1 citizen 1 com 4 crow 1 dangerous 1 deaths 2 die 1 easier 1 eigth 1 enuff 1 …

# Coming up

- Please try the Mallet commands ASAP to ensure it runs for you. Do not wait.