Practical exercice on ME

Joan Andreu Sánchez

Departamento de Sistemas Informáticos y Computación
Universidad Politècnica de Valencia

URL: http://www.dsic.upv.es/∼jandreu
e-mail: jandreu@prhlt.upv.es

Introduction to MALLET

MALLET keeps an internal representation for the data. For a given text, the text is transformed a vector in which the position is defined with a mapping and the position in the vector stores the ocurrences of each word in the text. Each word is a feature.

Example

Supose that MALLET uses a `hash` for the mapping funciont `f`:

```
f("This") = 345    f("is") = 174      f("and") = 705
f("a") = 5         f("table") = 15
f("this") = 798    f("chair") = 191
```

Then a sentence is coded as follows:

```
This  is a table and this  is a chair
345  174 5  15   705 798  174 5  191
```

Introduction to MALLET

MALLET can keep internally the information in several formats.

- As a sequence of features:

```
345, 174, 5, 15, 705, 798, 174, 5, 191
```

- As a bag of features:

```
5 15 174 191 345 705
2  1  2   1   1   1
```

MALLET registers the following information of each instance:

- Instance name
- Data (as explained above)
- Label
- Source (th original data)

Introduction to MALLET

MALLET support two different ways of working

- Scripts

```
$ bin/mallet --help
```

- Java classes

```
$ java -cp "class/:lib/mallet-deps.jar" \
  cc.mallet.classify.tui.Csv2Vectors --help
```

Option `--cp` specifies the CLASS PATH variable

The API is available at http://mallet.cs.umass.edu/api/

MALLET can read data from

- Directories
- Files

Introduction to MALLET

Convert some date to MALLET format (script)

```
$ tree sample-data/web/
sample-data/web/
|-- de
|---|-- apollo8.txt
|---|-- fiv.txt
...
|-- en
    |-- elizabeth_needham.txt
    |-- equipartition_theorem.txt
    ...

$ bin/mallet import-dir --input sample-data/web/* --output web.mallet
```

Convert some date to MALLET format (Java classes)

```
$ java -cp \
  "class/:lib/mallet-deps.jar" cc.mallet.classify.tui.Text2Vectors \
  --input sample-data/web/* --output web.mallet
```

MALLET will use the directory names as labels and the filenames as instance names

Introduction to MALLET

MALLET accepts other formats. The most interesting is one file with one line per instance. The format is similar to this:

```
instance0 label0 w01 w02 ...
instance1 label1 w11 w12 ...
...
```

The MALLET script command are:

```
$ bin/csv2vectors --input myfile --output myfile.vectors
$ bin/mallet import-file --input myfile --output myfile.mallet \
  --use-pipe-from myfile.vectors
```

Option `--use-pipe-from` specifies that the word coding is stored in the file `myfile.vectors` for further use in other files different from file `myfile`
Other relevant options are:

- `--keep-sequence`: If true, final data will be a FeatureSequence rather than a FeatureVector. Default is false.
- `--preserve-case` Do not force all strings to lowercase. Default is false.
- `--token-regex`: To define tokens!!

Creating a classifier with MALLET

For creating a classifier with MALLET, we can proceed as follows:

```
$ bin/mallet train-classifier --input web.mallet \
           --trainer MaxEnt --output-classifier myfile
```

The classifier is stored in binary format. If you want to see the classifer:

```
$ bin/classifier2info --classifier myfile
FEATURES FOR CLASS de
 <default> 0.021118921395153 81
 die 0.09235271137405299
 ...
```

Each line has the weight associated to each feature.
Additional options can be seen with one of the following commands:

```
$ bin/mallet train-classifier --help
$ java -cp "class/:lib/mallet-deps.jar" \
  cc.mallet.classify.tui.Vectors2Classify --help
```

Creating a classifier with MALLET

See data description here

Let us create a ME classifier with the chromosome task.

```
$ cd /tmp/tools/
$ wget http://www.dsic.upv.es/~jandreu/data.tgz
$ tar zxvf data.tgz
$ cd data
$ ../mallet-2.0.7/bin/csv2vectors --input cromosTr \
  --output cromosTr.vectors
$ ../mallet-2.0.7/bin/mallet import-file --input cromosTr --output  \
  cromosTr.mallet --use-pipe-from cromosTr.vectors
$ ../mallet-2.0.7/bin/mallet train-classifier --input cromosTr.mallet \
  --trainer MaxEnt --output-classifier cromosTr.classifier
$ ../mallet-2.0.7/bin/mallet import-file --input cromosTe --output  \
  cromosTe.mallet --use-pipe-from cromosTr.vectors
$ ../mallet-2.0.7/bin/vectors2classify --input cromosTr.classifier \
  --training-file cromosTr.mallet --testing-file \
  cromosTe.mallet --trainer MaxEnt --report test:confusion
```

Additional options can be obtained with:

```
$ ../mallet-2.0.7/bin/vectors2classify --help
```

Creating a classifier with MALLET

You can train a classifier and then to use it as follows:

```
$ ../mallet-2.0.7/bin/mallet train-classifier --input cromosTr.mallet \
  --trainer MaxEnt --output-classifier cromosTr.classifier

$ ../mallet-2.0.7/bin/mallet classify-file --input cromosTe.mallet \
  --output - --classifier cromosTr.classifier
```

But in such case, the test results are not reported directly and you have to compute it with script.

Available Resources: Datasets

- Autoritas dataset

  https://github.com/autoritas/RD-Lab/archive/master.zip

- http://www.hlt.utdallas.edu/~sajib/multifacetedText.html

- https://archive.ics.uci.edu/ml/datasets.html

- http://data.princeton.edu/wws509/datasets/

- http://komarix.org/ac/ds/

- http://www.umass.edu/statdata/statdata/stat-logistic.html

- http://www.statsci.org/datasets.html