Tim auf der Landwehr
t.auf.der.landwehr@student.rug.nl
S2548682
Last revision: December 10, 2013

Learning from Data
Assignment 5

Summary

This is the assignment 5 in the course Learning from Data at the University of Groningen.

Exercise 1: Cluto-formatter

A script that converts the given file `name_vectors_small.txt` (and later `name_vectors.txt`) to the dense matrix format is provided in `cluto_formatter.py`. This can be run the following way:

```
python3 cluto_formatter.py --in name_vectors.txt --out name_vectors.mat --labels
```

When the parameter `--labels` is used, the files
`[output_filename].rlabel` and
`[output_filename].clabel`
are created in addition.

Exercise 2: vcluster with cluto

The file that was created in exercise 1 is now processed with vluster.

```
vcluster -showtree -sim=cos -clmethod=agglo -plottree=plot.ps name_vectors_small.mat 7
```

Combining the labels with the classes again, and sorting them by class, the following distribution can be obtained:

```
Class 0:
  Donner
  Zalm
  Bos
  Jan
  De_Boer
  De_Vries

Class 1:
  Bush
  Nederland
  Irak
  VS
  IsraÃ«l
  Duitsland
  Frankrijk
  Europa
  Ajax
  PSV
  BelgiÃ«
  ItaliÃ«
  Feyenoord
  China
  Spanje
  Turkije
  Amerika
  Rusland
  Barcelona
  Vlaanderen
  Engeland
  Anderlecht

Class 2:
  Balkenende
  Blair
  SchrÃ¶der
  Kok
  Berlusconi

Class 3:
  Amsterdam
  Rotterdam
  Utrecht
  Den_Haag
  Brussel
  Parijs

Class 4:
  Ahold
  ING
  KLM
  Philips

Class 5:
  A

Class 6:
  VVD
  CDA
  PvdA
```

This shows how well this classifier works already for quite small datasets.

Exercise 3: Large data set

Now the same procedure is used on the big dataset provided in `name_vectors.txt`. The following listing shows the classification in 10 classes, the entities per class are limited to the first 10. From this you can see, that the classification worked quite well.

```
Class 0:
  CD&V
  VLD
  SP.A
  CD&V_/_N-VA
  VVD
  CDA
  Groen_!
  D66
  N-VA
  Vlaams_Belang
Class 1:
  Peeters
  Donner
  Zalm
  Yves_Leterme
  Verdonk
  Janssens
  Frank_Vandenbroucke
  Bernhard
  Kurt
  Filip
Class 2:
  Albert
Class 3:
  Antwerpen
  West-Vlaanderen
  Gent
  Brussel
  Leuven
  Amsterdam
  Brugge
  Oostende
  Limburg
  Kortrijk
Class 4:
  Bos
  FinanciÃ«n
  Justitie
  Sport
  Cultuur
  Groen
  Onderwijs
  Kunst
  Economie
  Jeugd
Class 5:
  Filip_Dewinter
  Lance_Armstrong
Class 6:
  A
  B
  C
  Ahold
  D
  H
  ING
  Fortis
  Delhaize
  Philips
Class 7:
  MÃ¡xima
  Maria
  Astrid
  Diana
  Nathalie
  Electrabel
  Els
  Victoria
  Kim
  Kim_Clijsters
Class 8:
  Bush
  Balkenende
  Verhofstadt
  Blair
  George_Bush
  SchrÃ¶der
  Sharon
  Berlusconi
  Kok
  Abbas
Class 9:
  Nederland
  BelgiÃ«
  Irak
  Vlaanderen
  VS
  Tom_Boonen
  Frankrijk
  Duitsland
  Europa
  ItaliÃ«
```

Exercise 4&5: Label classes

In this task, an approach to define the 10 classes by their containing entities is given. Especially the four tags *MISC*, *ORG*, *LOC* and *PER* are considered.

Class 0: shortcuts

Class 1: basically names → PER, but *Zalm* is a fish

Class 2: name → PER

Class 3: cities → LOC

Class 4: different abstract things → MISC

Class 5: names → PER

Class 6: mixed: letters and companies → partly ORG

Class 7: mixed: names and companies → partly PER and ORG

Class 8: names → PER

Class 9: mostly countries → LOC

It is not completely possible to assign perfect class names to the data, but the results are quite sophisticated.