Homework 5: Clustering of Books – Part Two

Fundamentals of Data Science, 2016/2017

Note: to complete this homework you should master the basics of Python and NumPy covered in previous lessons – especially list comprehension, function definition, control loops, and multidimensional arrays. If you are not comfortable with these subjects, review them now.

In the following, "ID" denotes your student's ID if you have one, else it denotes your last name (please, suppress accents if present).

Your goal is to submit a Python module named `libID.py` and a Python script named `ID.py` satisfying the requirements described below.

1 `libID.py`

This module should implement the functions described in the class syllabus available online:

```
charfreq(filename, filter)
euc(x, y)
cldist(c1, c2)
closest(L)
single_linkage(L, k)
```

Example.

```
>>> import libID
>>> files = ['da-1.txt', 'da-2.txt', 'en-1.txt', 'en-2.txt']
>>> D = [libID.charfreq(f) for f in files]
>>> D
[array([ 0.24,  0.43,  0.13,  0.14,  0.05]),
 array([ 0.23,  0.46,  0.13,  0.12,  0.05]),
 array([ 0.21,  0.34,  0.18,  0.2 ,  0.08]),
 array([ 0.22,  0.33,  0.17,  0.2 ,  0.07])]
>>> libID.single_linkage(D, 2)
[[0, 1], [2, 3]]
```

2 `ID.py`

`ID.py` should take as command line arguments: the path of n text files, a string of characters, and an integer k. For example with $n = 7$ and $k = 2$:

```
$ python ID.py book1.txt book2.txt ... book7.txt aeiou 2
```

(hint: `sys.argv`, list slicing). The script should then cluster the files into k clusters using single linkage, where each file is represented by a vector of frequencies (histogram), which tells how often each of the characters in the specified string occurs among all the charactes in the specified string (so, the same we have done in class). The script should then just output, for each cluster, one line that specifies which files it contains.

Example.

```
$ python ID.py da-1.txt da-2.txt en-1.txt en-2.txt fr-1.txt fr-2.txt it-1.txt it-2.txt aeiou 3
['da-1.txt', 'da-2.txt']
['en-1.txt', 'en-2.txt', 'fr-1.txt', 'fr-2.txt']
['it-1.txt', 'it-2.txt']
```

3 Submitting

You should submit the two Python files on or before Thursday November 24, 2016, 23:59 (Rome time), via email to the usual address, using the subject "ID hw5".