Chair for Algorithms
and Data Structures
Prof. Dr. Hannah Bast
Elmar Haußmann

Information Retrieval
WS 2015/2016

http://ad-wiki.informatik.uni-freiburg.de/teaching

Exercise Sheet 9

Submit until Tuesday, December 22 at **2:00pm**

Exercise 1 (14 points)

Copy your code from the last exercise sheet (or from the master solutions if you prefer) to a new folder *sheet-08*. Remove the methods concerning LSI and query processing (you don't need that for this sheet), and keep only the methods needed to build the term-document matrix.

Then add the following methods, specified in the TIP file on the Wiki: *initialize_centroids* (2 points), *compute_distances* (4 points), *compute_assignment* (4 points), *compute_centroids* (4 points).

Make sure to write a unit test for each method (except *intialize_centroids* if you use random numbers).

The methods for normalization are already fully coded, see the remarks on slide 24 of the lecture.

Exercise 2 (6 points)

Implement k-means using the code and methods from Exercise 1. Run it on the movie dataset from Exercise Sheet 2 (file *movies2.txt*) using $k = 50$. Write the top-10 terms from each of the final centroids to a file *centroids.txt* (format: one line per centroid, with the 10 words separated by spaces). Report on the Wiki: creation time for the term-document matrix, number of iterations, final RSS, and total running for your *k_means* method.

Commit your file *centroids.txt* to the SVN and briefly discuss it (along with the usual feedback) in your *experiences.txt* for this sheet.

[turn over at christmas]

```
                                               zr
                                            Yt$$$.
                                         .,e$$$$$F’
                       4e r               $$$$$$$.
                       d$$br            _z$$$$$$$F‘
                        ?$$b._          ^?$$$$$$$
                         4$$$"     -eec.  ""JP" ..eee$%..
                         -**N #c   -^***.eE  ^z$P$$$$$$$$$r-
                .ze$$$$$eu?$eu ’$$$$b $=^*$$ .$$$$$$$$$$"
             --."?$$$$$$$$$c"$$c .""" e$K  =""*?$$$P""""
 ueee. ‘:‘  $E !!h ?$$$$$$$$b R$N’~!! *$$F J"""C.  ‘
J  ‘"$$eu‘!h !!!‘4!!<?$$$$$$$P ?".eee-z.ee" ~$$e.br
’j$$Ne‘?$$c‘4!~‘-e-:!:‘$$$$$$$ $$**"z $^R$P  3 "$$$bJ
 4$$$F".‘?$$c‘!! \).!!!‘?$$$$F.$$$# $u$% ee*"^ :4‘"$"?$q
  ""‘,!!!:‘$$N.4!!~~.~~4 ?$$F’$$F.@.* -L.e@$$$$ec.      "
  "Rr‘!!!!h ?$$c‘h: ‘# !! $F,r4$L***  e$$$$$$$$$$$$hc
    #e’4!!!!L‘$$b’!.:!h‘~~ .$F’"    d$$$$$$$$$$$$$$$$$h,
     ^$.‘!!!!h $$b‘!. -    $P /’   .$$$$$$$$$$$$$$$$$$$$$c
       "$c‘!!!h‘$$.4~      $$$r’   <$$$$$$$$$$$$$$$$$$$P"""
         ^te.‘~ $$b        ‘Fue-   ‘$$$$$$$$$$$$$$P".:  !! "<
            ^"=4$P"     .,,,. -^.   ?$$$$$$$$$$"?:. !! :!!~ ,,ec..
                  ..z$$$$$$$$$h,    ‘$$$$$$P"..‘!f :!f ~)Lze$$$P""""?i
                ud$$$$$$$$$$$$$$h    ‘?$$F <!!’<!>:~)ue$$P*"..:!!!!! J
              .K$$$$$$$$$$$$$$$$$,     P.>e’!f !~ ed$$P".:!!!!!!!!‘.d"
             z$$$$$$$$$$$$$$$$$$$$      4!!~\e$$$P‘:!!!!!!!!!!’.eP’
            -*". . "??$$$$$$$$$$$$       ~ ‘z$$$F".‘!!!!!!!!!!’,dP"
          ." ):!!h i‘!- ("?$$$$$$f        ,$$P":! ). ‘’!!!!‘,d$F’
     .ueeeu.J‘-^.!h <-  ~‘.. ??$$’       ,$$ :!!‘e$$$$e ‘,e$F’
  e$$$$$$$$$$$$$eeiC ")?:-<:%’:^?        ?$f !!! ?$$$$",F"
 P"....‘‘‘""?$$$$$$$$$euL^.!..‘ .         "Tu._.,‘‘""
 $ !!!!!!!!!!::.""??$$$$$$eJ~^=.            ‘‘‘‘
 ?$.‘!!!!!!!!!!!!!!:."??$$$$$c’.
  "?b.‘!!!!!!!!!!!!!!!!>."?$$$$c
    ^?$c‘’!!!!!!!!!!!’,eeb, "$$$k
       "?$e.‘’!!!!!!! $$$$$ ;.?$$
          "?$ee,‘‘’’!."?$P‘i!! 3P
              ""??$bec,,.,ceeeP"
                     ‘""""""
```