Data Mining 2015 - Homework 5

Ludovico Fabbri 1197400

July 11, 2015

1 Problem 1

Here we ask you to implement Problem 4 of Homework 1, in Hadoop. In particular, write a MapReduce program, executable in Hadoop, to find the top-10 beers with the highest average overall score among the beers that have had at least 100 reviews. [..]

1.1

Source code files for hadoop:

- problem1/hadoop/makefile
- problem1/hadoop/TopTenBeers.java

Output:

- problem1/hadoop/topTenBeers-Out.txt

```
18.22621359223301   Westvleteren 12
17.862595419847327  Three Floyds Oak Aged Dark Lord Russian
    Imperial Stout
17.71153846153846   Bells Bourbon Barrel Double Cream/Expedition
    Stout
17.708520179372197  Goose Island Rare Bourbon County Stout
17.676470588235293  Lost Abbey Yellow Bus
17.64021164021164   De Dolle Stille Nacht Reserva 2000
17.62121212121212   Russian River Pliny the Younger
17.61904761904762   AleSmith Barrel Aged Speedway Stout
17.516393442622952  Cigar City Zhukovs Final Push
17.513698630136986  AleSmith Speedway Stout
```

2 Problem 2

In this problem we will see how we can download data using an API, and practice some of the text processing steps and MapReduce. The goal is to create an inverted index for the abstracts of some articles from the New York Times newspaper. [..]

2.1

Source code files for documents download and pre-processing in python(NyTimes API articles download, punctuation, numbers and stopwrods removal, stems computing):

- problem2/python/esercizio2.py
- problem2/python/ex2preProcessing.py

- problem2/python/program.py

Source code files for inverted index and tfidf preprocessing in hadoop(input file is the output of the previous pre-processing step done in python):

- problem2/hadoop/Makefile
- problem2/hadoop/Esercizio2.java

3 Problem 3

In this assignment we will implement the k-means algorithm in Hadoop. The specifi- cations of the program are to accept a file where each line corresponds to a document and a value k of the number of clusters, then select k random initial centers, run the k-means algorithm, and return a file where each line corresponds to a cluster. [..]

3.1

First i've used a python program to parse *authors.html* file using Beautifulsoup package to find all the .txt english documents, next in the same program i've did some preprocessing (in python) on those documents, removing punctuation, numbers, header, stopwords and computing stems. The output of this first step of preprocessing is written on *ex3_preProcessed_step1.txt* file.

Next step of preprocessing is done in Hadoop (taking as input the output of the previous step) and here we compute tfidf and normalization, as well as formatting the data output for the next stage(k-means). Output is written on *ex3_ preProcessed_ step2.txt* file.

About k-means implementation, i've used both euclidean distance and cosine-similarity (in different source files) for the assignment of the vector-documents to the nearest cluster. Since we have normalized each document-vector in the pre-processing stage, cosine-similarity is particuarly easy and fast to compute. Also i've noticed that usually with cosine-similarity convergence of k-means require less iterations than using euclidean distance (about 7 iterations vs 14/more iterations).

I've performed several clusterings of Gutenberg cd using both euclidean distance or cosine-similarity for K=10/20/50. I've used a small python program (*compare.py* located at *problem3/python/compare.py*) to check intersection between outputs: especially with k=50 there are clusters that are the same, but usually since k-means ends up with local optimum (not global) clusters could be different at convergence for different iterations of the algorithm. About combining clusterings, we can run k-means many times and take the best result among all the outputs (as said, it is a local optimum so it is likely to change for each repetition of k-means).

Source code for pre-processing in python (step1):

- problem3/python/ex3_preProcessing.py

Source code for pre-processing in hadoop (step2):

- problem3/hadoop/hadoop/preProcessing/Makefile
- problem3/hadoop/hadoop/preProcessing/src/Esercizio3_preProcess.java

Source code and output files for k-means in hadoop:

- problem3/hadoop/hadoop/kmeans/Makefile
- problem3/hadoop/hadoop/kmeans/src/Esercizio3_kmeans.java
- problem3/hadoop/hadoop/kmeans/src/Esercizio3_kmeans_cosine.java
- problem3/hadoop/hadoop/kmeans/cd outputs/*.txt

This is an example of output file for k=10 using euclidean distance:

```
1 hlnty10.txt oleng10.txt wwend10.txt 1lotj10.txt 2ws0210.txt
    bgita10.txt 3lotj10.txt fs40w10.txt 1hofh10.txt 2hofh10.txt
    homer10.txt 2ws0110.txt niebl10.txt 8moor10.txt 2ws0810.txt
    tcosa10.txt mcrst11.txt anide10.txt wssnt10.txt 2rbnh10.txt 8
    grtr10.txt 8ntle10.txt 3drvb10.txt nblng10.txt rime10.txt
    bllfn10.txt esymn10.txt 2ws1410.txt iliad10b.txt holyw10.txt
    canpw10.txt mdvll10.txt 1drvb10.txt wsvns10.txt 2mart10.txt 1
    mart10.txt lgsp10.txt tmbn110.txt plrabn12.txt kjv10.txt
    pgbev10.txt bwulf11.txt idyll10a.txt marmn10a.txt 4lotj10.txt 1
    argn10.txt shlyc10.txt 8trsa10.txt tmbn210.txt llake10.txt
    sinex10.txt tmrcr10.txt 2ws4510.txt poe5v10.txt orfur10.txt
    ftroy10.txt 2lotj10.txt 2ws1510.txt 2ws4410.txt 2ws0410.txt
    pcwar10.txt fb10w11.txt 8purg10.txt kalec10.txt 2ws0310.txt 8
    augr10.txt 0ddcc10.txt 0ddcl10.txt rgain10.txt thdcm10.txt
    koran12a.txt 1mlyd10.txt
10  btsnl10.txt wizoz10.txt 8igjp10.txt ozland10.txt kswom10.txt
    rd2oz10.txt magoz10.txt 8hmvg10.txt doroz10.txt
2 legva12.txt lgtrd10.txt cs10w10.txt ithoa10.txt sioux10.txt
    roget15a.txt wnbrg11.txt oldno10.txt ponye10.txt rnpz810.txt
    cwgen11.txt mlcal10.txt strtt10.txt aaard10.txt bhawk10.txt
    callw10.txt 8hcal10.txt mtrcs10.txt mohic10.txt zambs10.txt 8
    year10.txt lcjnl10.txt 8read10.txt cfvrw10.txt dchla10.txt 8
    wsh110.txt
```

```
3 2ws1910.txt lstbw10.txt 2ws2310.txt phant12.txt liber11.txt 2
    ws2610.txt badge10a.txt sp85g10.txt drfst10.txt remus10.txt
    frank14.txt benhr10.txt rbddh10.txt cdben10.txt study10.txt 2
    ws2510.txt prjtr10.txt tarz210.txt gp37w10.txt phado10.txt
    jmlta10.txt prtrt10.txt rlchn10.txt 8josh10.txt memho11.txt
    poe3v11.txt 8fmtm10.txt chshr10.txt twrdn10.txt sorol10.txt 1
    donq10.txt 2ws3510.txt 8dubc10.txt hstwl10.txt 2ws1210.txt
    spzar10.txt 2ws4110.txt ee710.txt pengl10.txt andsj10.txt
    chacr10.txt sign410.txt rshft10.txt poe4v10.txt vlpnr10.txt
    shrhe10.txt 8swnn10.txt ttnic10.txt scarp10.txt vfear11a.txt
    btowe10.txt 2ws3310.txt mbova10.txt hphnc10.txt dtroy10.txt
    vampy10.txt 2city12.txt tbisp10.txt hgrkr10.txt emohh10.txt 2
    yb4m10.txt 8luth10.txt rholm10.txt 2ws2710.txt milnd11.txt
    grexp10a.txt 80day11.txt 8ldvc10.txt mtdtl11.txt 8tomj10.txt
    wlwrk10.txt gltrv10.txt beheb10.txt 2musk10.txt gardn11.txt
    ggpnt10.txt mormon13.txt gn06v10.txt mollf10.txt 1ws4710.txt
    oroos10.txt 8phil10.txt pas8w10.txt sunzu10.txt hdark12a.txt
    cnfcs10.txt judsm10.txt duglas11.txt jj13b10.txt ironm11.txt
    arjpl10.txt wtfng10.txt mthts11.txt owlcr10.txt 19rus10.txt
    bough11.txt cyrus10.txt fkchp10.txt scrlt12.txt adrwn10.txt
    wrnpc10.txt rob3w10.txt 2ws3910.txt brtns10.txt crsto12.txt
    msprs10.txt 8rinc10.txt esycr10.txt thjwl10.txt crito10.txt
    bskrv11a.txt 8celt10.txt lchms10.txt ssklt10.txt lvbma11.txt
    nativ10.txt nplnb10.txt 8gtdr10.txt 1musk12.txt lfcpn10.txt
    plpwr10.txt mdmar10.txt 1ws5110.txt 2ws3710.txt notun11.txt
    dracu12.txt 8homr10.txt nc13v11.txt prppr11.txt 3musk11.txt
    andvl11.txt 2ws3410.txt 2000010.txt fevch10.txt onepi10.txt
    anthm10z.txt pimil10.txt moon10.txt tfgtv10.txt poe2v10.txt
    aesop11.txt truss10.txt thx1010.txt poeti10.txt 8recn10.txt
    dyssy10.txt 2ws0910.txt sleep11.txt fuman12.txt 2ws3610.txt 8
    rbaa10.txt poe1v10.txt 8moon10.txt tarzn10.txt jungl10.txt
    advsh12.txt 8mala10.txt cstwy11.txt 8clln10.txt skawe10.txt 8
    ldvn10.txt dscmn10.txt shkdd10.txt 8vnmm10.txt armyl10.txt
    pplgy10.txt lsttn10.txt nb17v11.txt sprvr11.txt 8jrc710.txt
    agino10.txt ovtop10.txt eotsw10.txt drthn11.txt mn20v11.txt 5
    wiab10.txt baron10.txt moby11.txt comet10.txt manif12.txt
    dmoro11.txt rcddv10.txt 8shkm10.txt cloud10.txt hpaot10.txt 2
    ws2410.txt tprnc10.txt plgrm11.txt 2ws1810.txt litbl10.txt 8
    rran10.txt timem11.txt resur10.txt mdprp10.txt 2mlyd10.txt
    lesms10.txt 8tjna10.txt 8rheb10.txt chldh10.txt gm00v11.txt 8
    saht10.txt fldct10.txt warw12.txt 1mrar10.txt
4 g1001108.txt 61001108.txt 41001108.txt tftaa10.txt 21001108.txt
    f1001108.txt 81001108.txt 31001108.txt 51001108.txt a1001108.
    txt c1001108.txt 71001108.txt e1001108.txt 11001108.txt
    d1001108.txt 91001108.txt samur10.txt b1001108.txt
5 jm00v10.txt 4spne10.txt whwar10.txt 6linc11.txt pprty10.txt
    bill11.txt 8aggr10.txt twtp110.txt civil11.txt utopi10.txt
    panic10.txt bribe10.txt dcart10.txt 8sced10.txt 2dfre11.txt
    thngl10.txt irish10.txt 8romn10.txt lflcn10.txt areop10.txt 1
    dfre10.txt warje10.txt 7linc11.txt sphjd10.txt 3spne10.txt
    slman10.txt st15w10.txt tgovt10.txt 8rtib10.txt 2cahe10.txt
    hcath10.txt 8csus10.txt 8rome10.txt tfdbt10.txt 3dfre10.txt 6
    dfre10.txt mdntp10.txt eduha10.txt 1spne10.txt 1linc11.txt
    ebacn10.txt 8ushx10.txt 4dfre10.txt bygdv10.txt mtlad10.txt
    inagu10.txt 1cahe10.txt cnstr10.txt 8elit10.txt 8hist10.txt
    phrlc10.txt 2spne10.txt jandc10.txt mrclt10.txt twtp210.txt
    pmisr10.txt 8wwrt10.txt tctgr10.txt comsn10.txt pmbrb10.txt
    wltnt10.txt plivs10.txt 1onwr10.txt conra10.txt 5dfre11.txt
    trthn10.txt 3linc11.txt nqpmr10.txt repub13.txt sjv0410.txt
    otoos11.txt bfaut10.txt 5linc11.txt twtp410.txt 4linc11.txt 8
    euhs10.txt suall10.txt trabi10.txt vorow10.txt hbtht10.txt
    const11.txt pge0112.txt totlc10.txt cprrn10.txt soulb10.txt
    feder16.txt when12.txt hioaj10.txt prblm10.txt pscmg10.txt
    wwasw10.txt 5spne10.txt dscep10.txt
6 esper10.txt 2drvb10.txt cbtls12.txt emihh10a.txt 8cury10.txt 00
    ws110.txt spatr10.txt troic10.txt 2ws2910.txt
7 8tspv111.txt hhmms11.txt dnhst10.txt vstil10.txt hmjnc11.txt
    lndle10.txt 1jcfs10.txt njals10.txt ltswd10.txt 8ataw11.txt
    icfsh10.txt vlsng10.txt 8tspv211.txt svncl10.txt vbgle11a.txt
    nnsns10.txt smtlc10.txt btbst10.txt mohwk10.txt utrkj10.txt 8
    jrny10.txt
8 tturn10.txt nnchr10.txt 8frog10.txt alice30.txt 8crmp10.txt 2
    ws3010.txt shndy10.txt 8eftl10.txt potww10.txt 2ws1610.txt
    g138v10.txt 1vkip11.txt ulyss12.txt 8idol10.txt 2ws4010.txt
    mlfls10.txt nkrnn11.txt sstcq10.txt dmsnd11.txt cptcr11a.txt
    rplan10.txt mpolo10.txt tshak10.txt tess10.txt 2ws2810.txt
    mayrc10.txt bbeau10.txt tbyty10.txt jnglb10.txt idiot10.txt 2
    ws1710.txt 2ws3210.txt 2ws2110.txt wwrld10.txt siddh10.txt 2
    ws2210.txt ncklb10.txt h8ahc10.txt dkang10.txt cpogs10.txt
    prgob10.txt zenda10.txt ppikg10.txt 1ws4910.txt ylwlp10.txt
    agnsg10.txt vanya10.txt 2ws2010.txt mjbrb10.txt bnrwy10.txt
    humbn10.txt treas11.txt curio10.txt nsnvl10.txt 2ws3810.txt
    wflsh10.txt fgths10.txt cprfd10.txt 2ws1110.txt rosry11.txt 2
    ws4210.txt cndrl10.txt jude11.txt gbwlc10.txt avon10.txt
    pygml10.txt dblnr11.txt clckm10.txt hoend10.txt anne11.txt
    myant11.txt jpnft10.txt wuthr10.txt hfinn12.txt 8knck10.txt
    lglass18.txt carol13.txt 8ghst10.txt 1rbnh10.txt wwill10.txt 2
    ws1010.txt birds10.txt secad10.txt olivr10.txt hardt10.txt 2
    ws3110.txt utomc11.txt wmnlv10.txt janey11.txt swoop10.txt
    vccty10.txt mwktm10a.txt grimm10.txt pam1w10.txt scarr10.txt 2
    ws0610.txt tiobe10.txt mhyde10.txt 2ws4310.txt
9 pandp12.txt lwmen12.txt vfair12.txt blkhs10.txt 8hfld10.txt
    sawyr11.txt sense11.txt lacob11.txt b033w10.txt 8frml10.txt
```

3.2 Extra credit

Source code for the extra credit:

- problem3/hadoop/extraCredit/Makefile
- problem3/hadoop/extraCredit/src/Esercizio3_extra.java

First round of the hadoop program produces inverted index (in the hdfs /tmp folder). The second round produces formatted output to pass to k-means algorithm (hdfs /out).

k-means implementation is the same and is located at:

- problem3/hadoop/kmeans/*

Clustering output for k=10 after 20 iterations of k-means (stopped before convergence) is located at:

- problem3/hadoop/extraCredit/output/dvd_clusters_10.txt