

# Advanced Data Modeling

Summer Semester 2009

- Exercises IX -

*To be handed in before **2009-07-22, 14:00** via e-mail to [bercovici@uni-koblenz.de](mailto:bercovici@uni-koblenz.de) And [dividino@uni-koblenz.de](mailto:dividino@uni-koblenz.de), subject line: [ADM] ...*

Please send your eclipse project and your output file.

## 1. De.li.ci.ous Dataset

In the HDFS you can find 1.000 txt files about de.li.ci.ous users. Each file represents an user and you can information about the bookmarks (url). For each bookmark you can find a timestand representing the date when the user has bookmarket this url and the respectives tags he has associated to it.

The files are organized as following:

File: user01.txt

[url01] [timestamp] [tag011|tag012 | ....]

[url02] [timestamp] [tag021|tag022 | ....]

[url03] [timestamp] [tag031|tag032 | ....]

...

## 2. Spammers: Find the spammers.

We define a spammer a user that more than 50% of his URL belongs to the same domain and have been bookmarkt in the same day. For example, the user 0182910 has url

14.03.2008 www.cadetic.de

14.03.2008 www.re.cadeitc.de

14.03.2008 www.tu.cadetic.de

14.03.2008 www.saj.cadetic.de

Write a Java code that return a list of spammers and the urls that this users has bookmarked and which has characterized him as a spammer.