PA164 Projects

Theme: Homeless here

Data:
Czech newspapers (idnes, lidovky, pravo, dennik, hn),
or any of your taste, not only blesk, aha or sip. This choice has to be confirmed by a teacher.

Format:
Plain text, UTF-8

<title> Title </title> <subtitle> </subtitle> <contents> </contents>
<date>2013-10-15</date> <source>Pravo</source>

from 15.10.2013 to 15.10.2012. If not enough article, continue in the past

Classified to a date and a source
100 documents per a student

If you are in doubts that the text, present it at classes.

Tasks:

pre-processing : 1x feature (attribute) selection
1x feature construction (e.g. N-gramy, phrases, multiword expressions, …)

Classification

Clustering (unsupervised learning)

Association rules

Outlier detection

Keyword extraction (no need for explicit feature selection and feature extraction)