Lab 7: Data Mining.

Analiza documentelor text (Text mining)

Analiza textului are ca scop extragerea de informatii din documente (documentele sunt interpretate ca secvente de cuvinte). Principalele tipuri de prelucrari sunt: clasificarea si gruparea documentelor pe baza continutului lor. based on their content. Cea mai simpla abordare se bazeaza pe aplicarea urmatoarelor etape:

- Pre-procesarea textului prin:
  - Eliminarea cuvintelor de legatura (*stop words)*. Liste cu cuvinte de legatura pt diferite limbi pot fi gasite la http://www.ranks.nl/stopwords
  - Transformarea cuvintelor prin *stemming* (i.e. reducerea la radacina cuvintului). Cel mai popular algoritm de stemming este cel propus de Porter (vezi http://tartarus.org/martin/PorterStemmer/). Un serviciu web pt stemming este disponibil la http://text-processing.com/demo/stem/
- Construirea pt fiecare document a vectorului de frecvente(*frequency vector):* nr de aparitii ale fiecarui cuvant din dictionary in cadrul documentului. D

Exercitiul 1.

a) Deschideţi fişierul movieReviews.arff (conţine recenzii ale unor filme clasificate în două categorii: pozitive şi negative)
b) Construiţi setul de date cu ocurenţele termenilor în colecţia de recenzii folosind Filters->Unsupervised->Attribute->StringToWordVector
c) Aplicaţi un clasificator (e.g. Naïve Bayes) asupra setului de date. Obs: e necesar ca primul atribut (@@class@@) să fie definit ca atribut de clasă (folosind Edit, click dreapta pe @@class@@ şi selectând Attribute as class)
d) Analizaţi impactul utilizării etapei de stemming asupra calităţii clasificării.