Proiect

Informaţii generale despre proiect

- Proiectul se va realiza individual!
- Aplicatie + referat. Aplicaţia trebuie să fie implementată în limbajul Python si pentru a avea nota de trecere la examen, aplicatia trebuie sa functioneze.
- Referatul reprezinta de fapt o documentatie a aplicatiei si include:
    - Titlul proiectului
    - Ideea proiectului / abstract
    - Descrierea datelor folosite
    - Descrierea tehnicilor de invatare automata folosite
    - Software-ul si pachetele folosite
    - Rezultate experimentale obtinute
    - Concluzii
    - Referinţe bibliografice
- Aproximativ 3-4 pagini dar textul să fie scris folosind cuvintele voastre, nu copiat de pe internet!
- Referatul trebuie sa conţină şi printscreen-uri ale aplicatiei.
- Obligatoriu: referatul trebuie sa contina la sfarsit o sectiune numita: „Rulare pe calculatorul personal” in care sa puneti printscreen-uri care sa ateste rularea aplicatiei pe calculatorul vostru!!!

Sugestii pentru proiect:

Linkuri utile:

- http://cs229.stanford.edu/projects2013.html
- http://archive.ics.uci.edu/ml/
- www.kaggle.com

Alte exemple:

Ex 1:

Precipitation data: http://www.jisao.washington.edu/data_sets/widmann/

This dataset has includes 45 years of daily precipitation data from the Northwest of the US:

Project ideas: Weather prediction: Learn a probabilistic model to predict rain levels

Ex 2:

This dataset contains webpages from 4 universities, labeled with whether they are professor, student, project, or other pages: http://www-2.cs.cmu.edu/~webkb/

Project ideas:

Learning classifiers to predict the type of webpage from the text

Can you improve accuracy by exploiting correlations between pages that point to each other using graphical models?

Papers:

http://www-2.cs.cmu.edu/~webkb/

http://www.cs.berkeley.edu/~taskar/pubs/rmn.ps

Ex 3:

The datasets provided below are sets of emails. The goal is to identify which parts of the email refer to a person name. This task is an example of the general problem area of Information Extraction.

http://www.cs.cmu.edu/~einat/datasets.html

Project Ideas: Model the task as a Sequential Labeling problem, where each email is a sequence of tokens, and each token can have either a label of "person-name" or "not-a-person-name".

Papers: http://www.cs.cmu.edu/~einat/email.pdf

Ex 4:

Enron E-mail Dataset: http://www.cs.cmu.edu/~enron/

The Enron E-mail data set contains about 500,000 e-mails from about 150 users.

Project ideas: Can you classify the text of an e-mail message to decide who sent it?