Analyse des données

Cours

TD1

énoncé

énoncé

Filtrage collaboratif : ressources web

Données : http://grouplens.org/datasets/movielens/

cours :

http://acsweb.ucsd.edu/~dklim/mf_presentation.pdf

http://clgiles.ist.psu.edu/IST441/materials/powerpoint/RC-week10/Tutorial_IJCAI_2013-modified.pptx

Travaux dirigés

Ces travaux dirigés sont des "notebooks" fonctionnant sur l'interpréteur "jupyter". Les notebooks permettent d'écrire et d'exécuter des scripts python à l'aide d'un simple navigateur web. Les résultats d'exécution sont conservés et peuvent être retrouvés d'une session à l'autre.

Téléchargez les fichiers `.ipynb` dans un dossier. Ouvrez un terminal dans ce même dossier et tapez :

```
jupyter-notebook
```

Ceci ouvre un onglet de l'interpréteur jupyter dans votre navigateur. Cliquez ensuite sur le notebook sur lequel vous souhaitez travailler.

Pour utiliser un notebook, voir :

- 1. What is the Jupyter notebook?
- 2. Notebook basics
- 3. Running code
- 4. Working with Markdown cells

- Une vidéo en anglais

La librairie numpy

Les exercices et problèmes seront réalisés à l'aide des librairies scientifiques de Python: numpy, scipy et matplotlib. Le lien suivant propose une introduction à la librairie numpy (manipulation de matrices et de vecteurs en Python) :

- Tutoriel Numpy (rapide) Tutoriel Numpy (nbviewer)
- Aide-mémoire Numpy |Aide-mémoire Numpy (nbviewer)
- Python scientifique (plus détaillé)

- **Matrices creuses** :
  - Format COO
  - Format CSR

La librairie scikit-learn

- Documentation
- Flow Chart
- An introduction to machine learning with scikit-learn
- Working With Text Data

TD2

Téléchargez le fichier suivant :

- Enoncé du TD2 (notebook)
- Enoncé du TD2 (nbviewer)

TD3

Filtrage collaboratif sur la base "MovieLens"

Téléchargez le fichier suivant :

- Enoncé du TD3 (notebook)
- Enoncé du TD3 (nbviewer)

TD4

Le but de ce TD est de développer un moteur de recherche dans une base de textes. Nous utilisons une base constitués d'un peu plus de 11.000 messages postés sur des forums de discussion (anglophones), fréquemment utilisée en analyse des données.

Il s'agit de la base 20newsgroups.

Le TD est constitué de plusieurs exercices ainsi que d'un problème. Les principales opérations sont réalisées à l'aide de la librairie scikit-learn.

- indexation des fichiers et des termes
- génération une matrice creuse comptant les fréquences d'occurrence des termes
- vectorisation des messages et des requêtes
- fichier inverse
- calcul de similarité et tri de la liste des réponses

Enoncé du TD1 (notebook) Enoncé du TD1 (nbviewer)

From:
https://wiki.centrale-med.fr/informatique/ - **WiKi informatique**

Permanent link:
https://wiki.centrale-med.fr/informatique/restricted:rec

Last update: **2023/01/12 13:42**