Web Mining

M2 Statistics and Econometrics

2017 – 2018

Yoann Pitarch
pitarch@irit.fr

General Information

- Who?
    - Y. Pitarch (pitarch@irit.fr)
    - Associate professor in CS
- Where to find information?
    - *www.irit.fr/~Yoann.Pitarch* > ‘’*teachings*’’ section (Web Mining)
- Evaluation
    - Project
    - Kaggle competition

Web Mining

- Web is a collection of inter-related files on one or more Web servers.
- Web mining is

  The application of data mining techniques to extract knowledge from Web data

- Web data is
  - ***Web content*** – text, image, records, etc.
  - ***Web structure*** – hyperlinks, tags, etc.
  - ***Web usage*** – http logs, app server logs, etc.

Web Mining – History

- Term first used in 1996, defined in a ‘task oriented’ manner
- Alternate ‘data oriented’ definition given in 1997
- 1st panel discussion at ICTAI 1997
- Continuing forum
  - WebKDD workshops with ACM SIGKDD, 1999, 2000, 2001, 2002,
  - SIAM Web analytics workshop 2001, 2002,
- Special issues of DMKD journal, SIGKDD Explorations
- Papers in various data mining conferences & journals

Web Mining taxonomy

Web Structure Mining (WSM)

- Search Result Mining
- Capturing Web’s structure using link interconnections

Web Content Mining (WCM)

- Web Page Content Mining

Web Usage Mining (WUM)

- General Access Pattern Mining
- Customized Usage Tracking

Pre-processing Web Data

Web Content

Extract “snippets” from a Web document that represents the Web Document

Web Structure

Identifying interesting graph patterns or pre-processing the whole web graph to come up with metrics such as PageRank

Web Usage

User identification, session creation, robot detection and filtering, and extracting usage path patterns

A Few Themes in Web Mining

Some interesting problems on Web mining

- Mining what Web search engine finds
- Identification of authoritative Web pages
- Identification of Web communities
- Web document classification
- Weblog mining (usage, access, and evolution)
- Intelligent query answering in Web search

Schedule

- ***Session 1.*** A Python upgrading lecture
- ***Session 2.*** WSM: generalities, complex network properties and node-centric metrics
- ***Session 3.*** WSM: communities and link prediction
- ***Session 4.*** WCM: text representation and preprocessing
- ***Session 5.*** WCM: text clustering and classification
- ***Session 6.*** WUM: overview
- ***Session 7.*** Q&A about the project

— Session 1 —

- No theory, only practice
- Objective: to check how comfortable you feel with Python
- Instructions :
    1. Visit the teaching section of my website (www.irit.fr/~Yoann.Pitarch)
    2. Download the exercices (Web Mining section)
    3. Start coding