Web Mining

Student Projects

Prof. Dr. Christian Bizer
Cäcilia Zirn
Oliver Lehmberg

FFS 2015
1. Some Advertisements for Competitions
2. Requirements for Student Projects
3. Requirements for Project Summaries
4. Final Exam
ENSAE ParisTech and ParisTech invite all data science students from Universities all around the world to participate in the 1st edition of the “Data Science Game”. By solving a data driven issue, students will be able to enlighten their data science expertise in a both competitive and friendly spirit.

**Data Science Game**

*A joint Initiative of ParisTech and ENSAE ParisTech.*

The competition is supported by two major partners: Google Inc., who will provide the scope and material of the competition, and Capgemini, who will provide an amazing setting for the competition in Paris.

**A two-phase competition:**

- An online non-eliminatory phase from mid-May to mid-June 2015.
- A two-day competition in Paris, the 20th and 21st of June 2015.
Goal:
Participants combine technology and innovative ideas to design prototypes of digital products that tackle today's educational challenges.
Student Projects

**Goals**
- Gain more practical experience with
  - Web Usage Mining,
  - Web Structure Mining or
  - Web Content Mining
- Get to know additional problem-specific
  - preprocessing methods
  - Web mining methods

**Expectation**
- Select an interesting Web mining problem of your choice
- Solve the problem using
  - the Web mining methods that we have learned so far
  - plus some advanced problem-specific data pre-processing
  - other Web mining methods which might be helpful for solving the problem and build on what we learned in class
Procedure

Teams of four students

1. realize a Web mining project
2. write 12 page report about the project and the methods employed in the project
3. present the project results to the other students (12 minutes presentation + 8 minutes discussion)

Final mark for the course

- 50 % written exam
- 30 % written summary about the project
- 20 % project presentation
## Schedule

<table>
<thead>
<tr>
<th>Week</th>
<th>Thursday</th>
<th>Friday</th>
</tr>
</thead>
<tbody>
<tr>
<td>16.04.2015</td>
<td>Introduction to Student Projects</td>
<td>Prepararation of Project Outlines</td>
</tr>
<tr>
<td>20.04.2015, 23:59</td>
<td>Submission of Project Outlines</td>
<td></td>
</tr>
<tr>
<td>23.04.2015</td>
<td>Feedback about Project Outlines</td>
<td>Project work</td>
</tr>
<tr>
<td>30.04.2015</td>
<td>Coaching</td>
<td>- Holiday -</td>
</tr>
<tr>
<td>07.05.2015</td>
<td>Project work</td>
<td>Coaching</td>
</tr>
<tr>
<td>14.05.2015</td>
<td>- Holiday -</td>
<td>Coaching</td>
</tr>
<tr>
<td>21.05.2015</td>
<td>Project work</td>
<td>Coaching</td>
</tr>
<tr>
<td>25.05.2015, 23:59</td>
<td>Submission of Project Reports</td>
<td></td>
</tr>
<tr>
<td>28.05.2015</td>
<td>Presentation of project results</td>
<td>Presentation of project results</td>
</tr>
</tbody>
</table>
Where to find Web Usage Data Sets?

- Netflix
  - provided by commercial movie rental website for Netflix competition ($1,000,000 for 10% better RMSE)
  - 480,000 users rated 18,000 movies, 100M ratings

- Web Data Commons
  - http://www.webdatacommons.org/structureddata/
  - Product/hotel/restaurant ratings as part of Microdata dataset

- Stanford Large Network Dataset Collection
  - Amazon product co-purchasing data

- Amazon Public Datasets
  - https://aws.amazon.com/datasets
  - Wikipedia Traffic Statistics, ....

- Web 2.0 Platforms offer plenty of additional rating data
  - e.g. LastFM, delicious
Where to find Web Structure Data Sets?

- **Stanford Large Network Dataset Collection**
  - Social networks: Facebook, Google+
  - Citation networks: Arxiv, US Patents
  - Product co-purchasing network: Amazon

- **Web Data Commons Hyperlink Networks**
  - aggregated by website (40 million vertices)
  - [http://webdatacommons.org/hyperlinkgraph/](http://webdatacommons.org/hyperlinkgraph/)

- **The Koblenz Network Collection**
  - 235 networks about various topics
  - [http://konect.uni-koblenz.de/](http://konect.uni-koblenz.de/)

- **Billion Triples Challenge Dataset**
  - Linked Data crawled from the public Web

- **Archive.org**
  - Friendster social network (2011)
Where to find Web Content Data Sets?

- **Multiple-Aspect Restaurant Reviews**
  - Reviews taken from we8there, ratings for 5 main aspects of restaurants
  - [http://people.csail.mit.edu/bsnyder/naacl07](http://people.csail.mit.edu/bsnyder/naacl07)

- **The J.D. Power and Associates Sentiment Corpus**
  - Blog entries about cars and cameras, manually labeled with product features (including labeled part-of relations), opinion phrases and opinion targets
  - [http://verbs.colorado.edu/jdpacorpus/](http://verbs.colorado.edu/jdpacorpus/)

- **Web Data Commons**
  - Product/hotel/restaurant reviews as part of Microdata dataset
  - [http://www.webdatacommons.org/structureddata/](http://www.webdatacommons.org/structureddata/)

- **Stanford Large Network Dataset Collection**
  - Amazon product metadata and review information about 548,552 different products

- **Academictorrents.com**
  - Various large data sets
  - e.g. Enron Email Bag of Words, Arizona State University Twitter Data Set

- **Programmable Web**
  - Website giving an overview about 4100 public Web APIs
Some Project Ideas (not binding)

- **Web Usage Log Mining**
  1. Learn a classifier for categorizing the visitors of your website
  2. Identify common navigation paths, drop-out pages

- **Recommender Systems**
  1. Defend a recommender system by identifying fake ratings
  2. Experiment with hybrid recommenders in specific application domain
  3. Linked Open Data-enabled Recommender Systems Challenge

- **Network Analysis**
  1. Common Crawl Hyperlink Graph (analyze by country or topical domain)
  2. Linked Data Cloud (analyze by country, topical domain)
  3. Analyze Graph Structure of Wikipedia or DBpedia (detect communities)

- **Sentiment Analysis**
  1. Extracting product features and opinions using advanced methods
  2. Generating opinion summaries (aggregating extracted information)
  3. Comparison of machine learning vs. lexicon-based approaches
Project Outlines

- 2-3 pages (sharp!) without title or toc pages, DWS master thesis layout
- due 20.04.2015, 23:59
- send by eMail to Chris, Cäcilia, and Oliver
- answer the following questions:
  1. What is the problem you are solving?
  2. What data will you use?
     - Where will you get it?
     - How will you gather it?
  3. How will you solve the problem?
     - What preprocessing steps will be required?
     - Which algorithms you plan to use?
     - Be as specific as you can!
  4. How will you evaluate, measure success?
Coaching Sessions

- We will give you tips and answer questions concerning your project.

- Please send us an email that you want to attend a coaching session.
  - until Tuesday night
  - including the questions that you like to discuss

- We will assign you a time slot afterwards and inform you about the slot via email.

- You are required to attend at least one coaching session for getting feedback on your initial results.
Project Summaries

- Max. 12 pages (sharp!) without title or toc pages.
- Every additional page (including appendixes) downgrades your mark by 0.3
- due 25.05.2015, 23:59
- send by email to Chris, Cäcilia, and Oliver

Outline for project summaries:
1. Application area and goals
2. Structure and size of the data set
3. Preprocessing
4. Actual Web Mining
   - including evaluation results
   - including different approaches that you tried
5. Discussion of Results

Requirements
- You must use the DWS master thesis layout.
- Please cite sources properly. Preferred citation style [Author, year].
- Also submit your code/Rapid Miner processes and (a subset) of your data.
Final Exam

- **Date:** 1.6.2015
- **Duration:** 60 minutes
- **Structure:** 5 - 6 open questions that
  - check whether you have understood the content of the lecture
  - require you to describe the ideas behind algorithms and methods
  - might require you to do some simple calculations