Web Data Mining

Lecture 1: Introduction and Course Overview

Milan Dojčinovski
milan.dojchinovski@fit.cvut.cz

Czech Technical University in Prague - Faculty of Information Technologies - Software and Web Engineering

Summer semester 2019/2020
Humla v0.3

Overview

- Introduction
- Web Data Mining
- Course at a Glance
- Communication and Resources
- Python

Hellos

- Milan Dojčinovski
  - *Assistant professor, researcher*
  - *Czech Technical Uninversity in Prague, Czech Republic*
  - *Institute of Applied Informatics at Leipzig University, Germany*
  - *Research interests*
    → *Semantic web, Linked Data, Web services, Information Extraction*

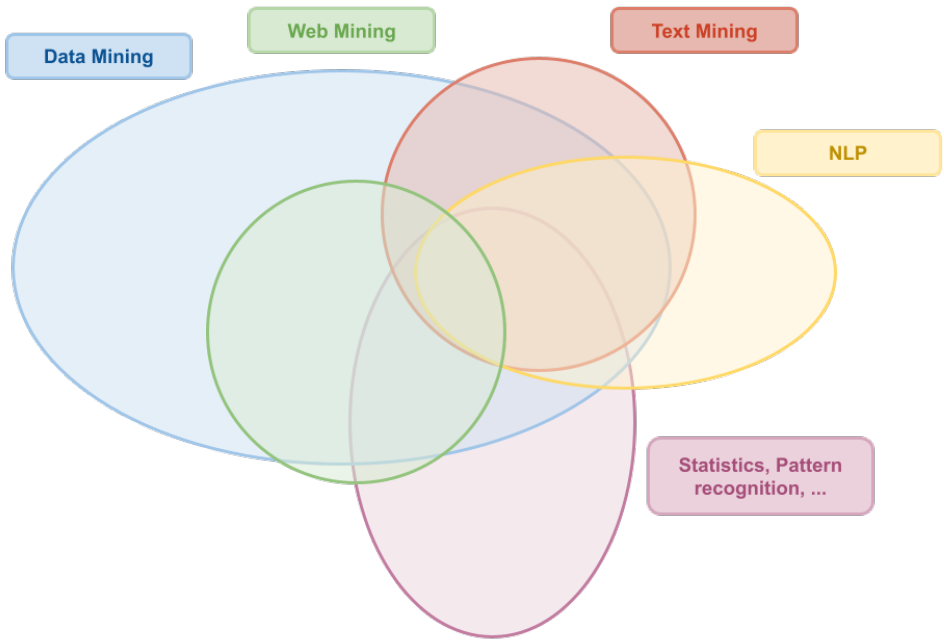
Overview

- Introduction
- Web Data Mining
- Course at a Glance
- Communication and Resources
- Python

Web Data Mining

- Web
  - *A huge, widely-distributed, diverse & heterogeneous, semi-structured, linked, redundant and dynamic information repository.*
- Mining
  - *Extracting something useful or valuable from a baser substance.*
- Data Mining
  - *A process of analyzing data from different perspectives and summarizing it into useful information.*
- Web Data Mining
  - *Is an application of the data mining techniques to find interesting and potentially useful knowledge from web data and services.*

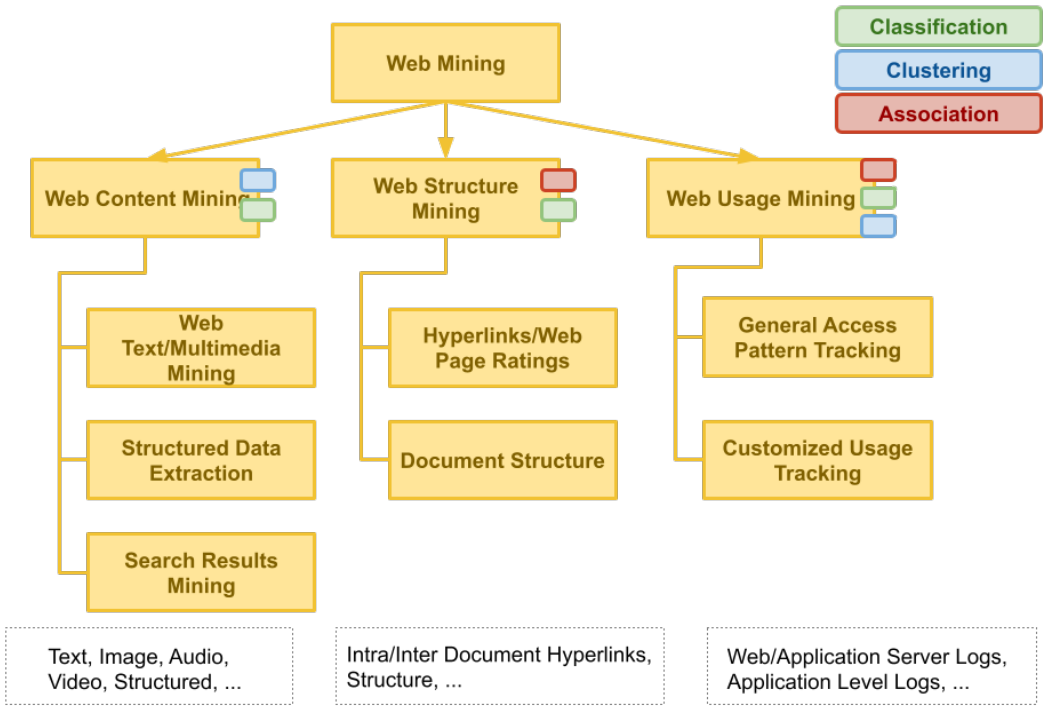
Areas Overlap



Web Mining

- Web Content Mining (WCM)
  - *Process of extracting information from the content of web documents.*
  - *Using of intelligent web agents - crawlers, robots, ...*
  - *Information Retrieval, Natural Language Processing, Computer Vision.*
- Web Structure Mining (WSM)
  - *Web pages as nodes, hyperlinks as edges.*
  - *Process of discovering structure information on the web.*
  - *Intra-page vs Inter-page.*
  - *Hyperlink analysis.*
- Web Usage Mining (WUM)
  - *Extracting useful information from server logs.*
  - *Process of finding out what users are looking on Internet.*
  - *User identification, sessionization, pattern discovery.*
- WCM → WSM → WUM

Web Mining Taxonomy



Web Mining Applications

- Web Content Mining (WCM)
  - *document clustering or categorization*
  - *topic identification/tracking*
  - *concept discovery*
  - *focused crawling*
  - *content-based personalization*
  - *intelligent search tools*
  - *search engines, ...*
- Web Structure Mining (WSM)
  - *document retrieval and ranking*
  - *discovery of hubs and authorities*
  - *discovery of web communities*
  - *social network analysis*
  - *search engines, SEO, ...*
- Web Usage Mining (WUM)
  - *user and customer behavior modeling*
  - *web sites optimization*
  - *web marketing/advertising*
  - *recommender systems*
  - *web analytics, ...*

Overview

- Introduction
- Web Data Mining
- Course at a Glance
- Communication and Resources
- Python

Motivation in Brief

- Rapid data growth
  - *huge amounts of freely available data on the Web*
  - *largest publicly available dataset*
  - *wide and diverse coverage of information*
  - *dynamic changes of the content*
  - *about 62 billion webpages (estimated size of Google's index)*
- Social Web
  - *users generate huge amount of data*
    → connections, comments, likes
  - *for the people and by the people*
  - *users are prosumers*
    → at the same time generate and consume information
  - *Facebook, Twitter, YouTube, Instagram, etc.*

Motivation in Brief (cont.)

- Data processing is difficult
  - *Big Data!*
  - *very large and complex datasets*
  - *existing database management tools do not meet the needs*
- Open challenges
  - *How to capture data*
  - *How to store data*
  - *How to **process and analyze** data*
  - *How to search for data*
  - *....in a reasonable time!*

Scope

- Accessing Data
  - *Crawling and information extraction*
- Storage
  - *Indexing principles*
- Data Analysis methods
  - *Text mining*
  - *Social network analysis*
  - *Web usage mining*
  - *Recommender systems*
- Applications
  - *Use case examples*
  - *Visualizations*
- Algorithms

Prerequisites

- Web Architecture
  - *Basics of HTTP, XML, XPath, HTML, URI*
- Programming skills
  - *Object-oriented programming*
  - *Principles*
    → *class, object, inheritance, encapsulation, ...*
    → *basis for service concepts*
- Others
  - *Graph theory and basic algorithms.*

Organization of Lectures

- 12 lectures
  - *English: Milan Dojčinovski*
  - *Czech: Jaroslav Kuchař*
- Plan
  1. *Motivation and Course Overview (html)*
  2. *Data Access and Acquisition Methods (html)*
  3. *Data Access and Acquisition Methods 2 (html)*
  4. *Text Mining 1 (html)*
  5. *Text Mining 2 (html)*
  6. *Social Network Analysis 1 (html)*
  7. *Social Network Analysis 2 (html)*
  8. *Page Rank and HITS (html)*
  9. *Web Usage Mining/Web Analytics (html)*
  10. *Recommender Systems (html)*
  11. *Mining Data Streams (html)*
  12. *Reserve (html)*

Organization of Tutorials

- Labs every second week
  - *individual work (no teams!)*
  - *be prepared for the lab!*
  - *work alone, ask others for advices*
- Number of sessions: 6
  1. *Data acquisition/crawling*
  2. *Text mining*
  3. *Social Network Analysis*
  4. *Structure mining*
  5. *Web usage mining/Pattern recognition*
  6. *Recommender systems/Data streams*

Methodology for Individual Work

- Methodology:
  1. *Data crawling*
     – *Non-trivial data acquisition from a web resource*
  2. *Data (pre)processing/Information extraction*
     – *Automatic/semi-automatic extractions/transformations*
  3. *Storing and indexing data*
     – *Dealing with issues how to properly store/index data*
  4. *Application of appropriate algorithms*
     – *Text mining, social network analysis, web usage mining, recommender systems, …*
- Results:
  – *work alone, ask others for advices*
  – *documentation*
- (Optionally individual work)

Assessment

- Practicals
  – *Tutorials (on-site/online)*
  – *Homeworks*
    → *Every task gives you some amount of points*
    → *Total maximal points: 40, to pass:* ***20***
- Labs
- Final Exam
  – *Mandatory written test: 2-3 exercises, ~1 hour*
    → *each gives you a max. of 20-30 points, the total is 60 points*
- Final score:
  – *practical (max 40) + final exam (max 60) = 100 max points*
  – *discussion may adjust your points freely*

Final Marks

Grade	Points	Verbal (Czech)	Verbal (English)
A	100-90	výborně	excellent
B	89-80	velmi dobře	very good
C	79-70	dobře	good
D	69-60	uspokojivě	satisfactory
E	59-50	dostatečně	sufficient
F	<50	nedostatečně	failed

Source: Study and examination regulations

- Everything good and bad will count
  – *practicals, coding, (pro)activity, passiveness, hacking (bad and good), lectures, exam, cheating, ...*

Overview

- Introduction
- Web Data Mining
- Course at a Glance
- Communication and Resources
- Python

Communication

- Language
  - *Text: English (slides, tweets, posts, instructions, etc.)*
  - *Voice: English and Czech*
- Direct
  - *You can always contact me directly at*
    → English: milan.dojchinovski@fit.cvut.cz

Overview of Resources

- Lectures
- Course Page
- Books
  1. *Liu, B. Web Data Mining, Springer-Verlag Berlin Heidelberg, 2011. ISBN 978-3-642-19459-7.*
  2. *Charu C. Aggarwal. Machine Learning for Text. Springer Publishing Company, Incorporated, 2018. ISBN: 9783319735313.*
  3. *Easley, D., Kleinberg, J. Networks, Crowds, and Markets: Reasoning About a Highly Connected World, Cambridge University Press, 2010. ISBN 978-0521195331.*
  4. *Ricci, F., Rokach, L., Shapira, B., B. Kantor, P. Recommender Systems Handbook, Springer, 2010. ISBN 978-0387858197.*
  5. *Charu C. Aggarwal. Recommender Systems: The Textbook. Springer Publishing Company, Incorporated, 2016. ISBN: 9783319296579.*

Overview of Resources (cont.)

- Books
  1. *Kaushik, A. Web Analytics 2.0: The Art of Online Accountability and Science of Customer Centricity, Sybex, 2009. ISBN 978-0470529393.*
  2. *Marmanis, H., Babenko, D. Algorithms of the Intelligent Web, Manning Publications, 2009. ISBN 978-1933988665.*
  3. *A. Russel, M. Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More (3rd edition), O'Reilly Media, 2019. ISBN 978-1491985045.*
  4. *Chakrabarti, S. Mining the Web: Discovering Knowledge from Hypertext Data, Morgan Kaufmann, 2002. ISBN 1558607544.*
  5. *Mitchell, R. Web Scraping with Python: Collecting Data from the Modern Web, O'Reilly Media; 2 edition, 2018, ISBN 978-1491985571*

Code Examples

- Example code in various languages
  – *Python, Java, JavaScript, xml, command line (bash) and plain text (e.g., BNFlike syntax, regular expressions, etc.)*
- Code colors
  – *different colors of the listings' gutters based on the language*

```
1 public String test() {
2     // this is a Java code
3 }
```

```
1 <test><!-- this is a XML code --></test>
```

```
1 echo "this is a bash test"
```

Overview

- Introduction
- Web Data Mining
- Course at a Glance
- Communication and Resources
- Python

Python Introduction

- Python
  - *high-level*
  - *interpreted*
  - *interactive*
  - *object-oriented*
  - *easy-to-use syntax*
  - *general-purpose*

```
1 num1 = 3
2 num2 = 5
3 sum = num1+num2
4 print(sum)
```

```
1 def myFun(x, y=1):
2     return x+y
3 print(myFun(3))
```

Others

- Installations
  - *Available for many OS*
    - *→ Linux, Windows, macOS, ...*
- Python 2 vs Python 3
  - *better Unicode support*
  - *print and exec being statements, integers using floor division*
  - *worse library support*
  - *some OS still use 2.x as default*
  - *...*
- Python 2.7 has officially reached the end of life
  - *January 1st, 2020*

```
# Python 2
print 'Hello, World!'
print '3 / 2 =', 3 / 2 # 1
```

```
# Python 3
print('Hello, World!')
print('3 / 2 =', 3 / 2) # 1.5
```


Running Python

- Python interpreter

```
$ python
Python 3.6.3 |Anaconda, Inc.| (default, Oct  6 2017, 12:04:38)
[GCC 4.2.1 Compatible Clang 4.0.1 (tags/RELEASE_401/final)] on darwin
Type "help", "copyright", "credits" or "license" for more information.
>>> print('Hello, World!')
Hello, World!
>>>
```

- Executing scripts

```
$ cat hello.py
if __name__ == "__main__":
    print('Hello, World!')
```

```
$ python hello.py
Hello, World!
```

- package installation

```
pip install package-name
```

Python basics

- Numbers

```
>>> tax = 12.5 / 100
>>> price = 100.50
>>> price * tax
12.5625
```

- Strings

```
>>> word = 'Python'
>>> word[0]
'P'
>>> word[-1]
'n'
>>> word[0:2]
'Py'
>>> word[:2] + word[2:]
'Python'
```

- Lists

```
>>> squares = [1, 4, 9, 16, 25]
>>> squares[0]
1
```


Python basics (cont.)

- Flow controls

```
>>> if x < 0:
...     print('Less')
... elif x == 0:
...     print('Zero')
... else:
...     print('More')
```

```
>>> for s in squares:
...     print(s)
```

- Functions

```
def myFun(x, y=1):
    return x+y
print(myFun(3))
```

- Data Structures

```
# list
fruits = ['orange', 'apple', 'pear', 'banana', 'kiwi', 'apple', 'banana']
# set
basket = {'orange', 'banana', 'pear', 'apple'}
# dictionary
counts = {'orange': 1, 'apple': 2}
```

Python basics (cont.)

- Iterators

```
>>> mylist = [x*x for x in range(3)]
>>> for i in mylist:
...     print(i)
0
1
```

- Generators
  - *you can only iterate over them once*
  - *they generate the values on the fly*
  - *calculation on-demand, also called lazy evaluation*

```
>>> mygenerator = (x*x for x in range(3))
>>> for i in mygenerator:
...     print(i)
0
1
```

```
>>> def createGenerator():
...     mylist = range(3)
...     for i in mylist:
...         yield i*i
>>> mygenerator = createGenerator()
>>> print(mygenerator)
<generator object createGenerator at 0xb7555c34>
>>> for i in mygenerator:
...     print(i)
0
1
4
```


Tools

- Virtualenv
  - *a tool to create isolated Python environments*
  - *solving problems with*
    - *→ dependencies and versions*
    - *→ indirectly permissions*
- Workflow

```
# create new environment
virtualenv mi-ddw

# activate
source mi-ddw/bin/activate

# operations ...
pip install ...

# deactivate and remove
deactivate
rm -r ./mi-ddw
```

Tools (cont.)

- Anaconda
  - *data science and machine learning distribution*
  - *hundreds of popular data science packages*
    → a collection of 1,000+ open source packages
  - *open-source package manager, environment manager*

```
# package installation
conda install package-name

# update package
conda update package-name
```

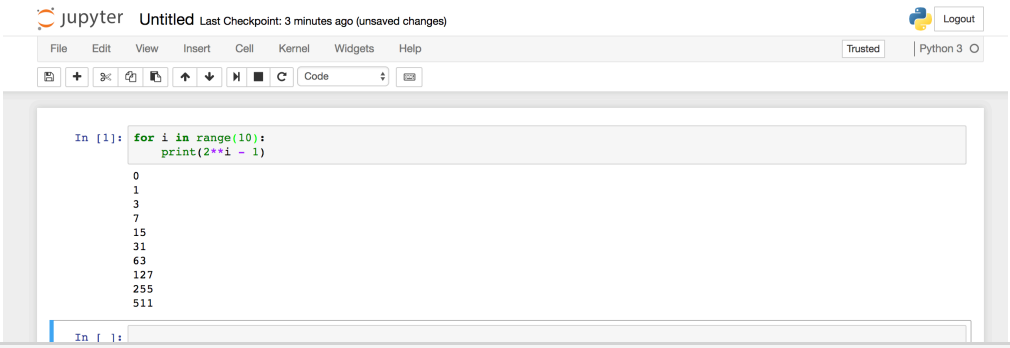
```
# environments
conda create --name my_env python=3
source activate my_env
...
source deactivate
```


Tools (cont.)

- Jupyter
  - *open-source web application allowing documents that contain live code, equations, visualizations and narrative text*
  - *data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more*

```
# part of the Anaconda or
pip3 install jupyter

# start notebook
jupyter notebook
```



Important packages

- NumPy
  - *N-dimensional array object, linear algebra, Fourier transform, and random number capabilities*
- Pandas
  - *Python Data Analysis Library, Great for data munging and preparation, fast and efficient DataFrame object for data manipulation with integrated indexing*
- Natural Language Toolkit - NLTK
  - *Python implementation of Text Mining, Natural Language Processing algorithms, ...*
- NetworkX
  - *Python language software package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks.*
- …

Python Resources

- Recommended tutorials
  - *Czech*
    → *http://naucse.python.cz/*
  - *English*
    → *https://docs.python.org/*
- Books
  - *https://wiki.python.org/moin/PythonBooks*
  - *https://www.amazon.com/Best-Sellers-Books-Python-Programming/zgbs/books/285856*
  - ...