Big Data Mining

Course Orientation for Big Data Mining

1071BDM01
TLVXM1A (M2244) (8619) (Fall 2018)
(MBA, DBETKU) (3 Credits, Required) [Full English Course]
(Master’s Program in Digital Business and Economics)
Mon, 9, 10, 11, (16:10-19:00) (B206)

Min-Yuh Day, Ph.D.
Assistant Professor

Department of Information Management
Tamkang University

http://mail.tku.edu.tw/myday

2018-09-10
Course Syllabus
Tamkang University
Academic Year 107, 1st Semester (Fall, 2018)

• Course Title: Big Data Mining
• Instructor: Min-Yuh Day
• Course Class: TLVXM1A (MBA DBETKU)
  – Master’s Program in Digital Business and Economics, 1A
• Details
  – Required
  – One Semester
  – 3 Credits
• Time & Place: Mon, 9, 10, 11, (16:10-19:00) (B206)
Department Teaching Objectives

• Train students not only to acquire knowledge from economics, finance, and industrial developments but also to apply information technology and analytical and quantitative skills to various situations.

• Students can enhance their competitiveness in facing rapid changes in world economy.
Department Core Competences

1. Cultivating students the ability of computer programming.
2. Training students the ability of website design for starting up a business.
3. Training students the ability of analyzing various situations in the financial market.
4. Helping students to acquire the knowledge of financial technology.
Course Introduction

• This course introduces the fundamental concepts and research issues of Big Data Mining.

• Topics include
  – ABC: AI, Big Data, Cloud Computing,
  – Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data,
  – Fundamental Big Data: MapReduce Paradigm, Hadoop and Spark Ecosystem,
  – Foundations of Big Data Mining in Python,
  – Supervised Learning: Classification and Prediction,
  – Unsupervised Learning: Cluster Analysis,
  – Unsupervised Learning: Association Analysis,
  – Machine Learning with Scikit-Learn in Python,
  – Deep Learning for Finance Big Data with TensorFlow,
  – Convolutional Neural Networks (CNN)
  – Recurrent Neural Networks (RNN)
  – Reinforcement Learning (RL)
  – Social Network Analysis (SNA)
Teaching Objectives

1. Understand and apply the fundamental concepts and research issues of big data mining.

2. Conduct information systems research in the context of big data mining.
Teaching Methods

• Lecture
• Discussion
• Simulation
• Practicum
• Problem Solving
Assessment

• Practicum
• Report
• Participation
<table>
<thead>
<tr>
<th>Week</th>
<th>Date</th>
<th>Subject/Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2018/09/10</td>
<td>Course Orientation for Big Data Mining</td>
</tr>
<tr>
<td>2</td>
<td>2018/09/17</td>
<td>ABC: AI, Big Data, Cloud Computing</td>
</tr>
<tr>
<td>3</td>
<td>2018/09/24</td>
<td>Mid-Autumn Festival (Day off)</td>
</tr>
<tr>
<td>4</td>
<td>2018/10/01</td>
<td>Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data</td>
</tr>
<tr>
<td>5</td>
<td>2018/10/08</td>
<td>Fundamental Big Data: MapReduce Paradigm, Hadoop and Spark Ecosystem</td>
</tr>
<tr>
<td>6</td>
<td>2018/10/15</td>
<td>Foundations of Big Data Mining in Python</td>
</tr>
<tr>
<td>7</td>
<td>2018/10/22</td>
<td>Supervised Learning: Classification and Prediction</td>
</tr>
<tr>
<td>8</td>
<td>2018/10/29</td>
<td>Unsupervised Learning: Cluster Analysis</td>
</tr>
<tr>
<td>9</td>
<td>2018/11/05</td>
<td>Unsupervised Learning: Association Analysis</td>
</tr>
</tbody>
</table>
# Course Schedule (2/2)

<table>
<thead>
<tr>
<th>Week</th>
<th>Date</th>
<th>Subject/Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>2018/11/12</td>
<td>Midterm Project Report</td>
</tr>
<tr>
<td>11</td>
<td>2018/11/19</td>
<td>Machine Learning with Scikit-Learn in Python</td>
</tr>
<tr>
<td>12</td>
<td>2018/11/26</td>
<td>Deep Learning for Finance Big Data with TensorFlow</td>
</tr>
<tr>
<td>13</td>
<td>2018/12/03</td>
<td>Convolutional Neural Networks (CNN)</td>
</tr>
<tr>
<td>14</td>
<td>2018/12/10</td>
<td>Recurrent Neural Networks (RNN)</td>
</tr>
<tr>
<td>15</td>
<td>2018/12/17</td>
<td>Reinforcement Learning (RL)</td>
</tr>
<tr>
<td>16</td>
<td>2018/12/24</td>
<td>Social Network Analysis (SNA)</td>
</tr>
<tr>
<td>17</td>
<td>2018/12/31</td>
<td>Bridge Holiday (Extra Day Off)</td>
</tr>
<tr>
<td>18</td>
<td>2019/01/07</td>
<td>Final Project Presentation</td>
</tr>
</tbody>
</table>
Grading Policy

• **Mark of Usual:** 40%

• **Final Project:** 60%
  – Midterm Project Report
  – Final Project Report
Textbooks and References

• Textbook: Slides
  – http://mail.tku.edu.tw/myday/teaching.htm#1071BDM

• References
  – Data Mining: Concepts and Techniques, Third Edition, Jiawei Han, MichelineKamber and Jian Pei, Morgan Kaufmann, 2011.

Source: https://www.amazon.com/Learning-Data-Mining-Python-Second/dp/1787126781


Source: https://www.amazon.com/Practical-Machine-Learning-Python-Problem-Solvers/dp/1484232062


Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners,
Jared Dean,

Data Mining: Concepts and Techniques, Third Edition, Jiawei Han, Micheline Kamber and Jian Pei, Morgan Kaufmann, 2011


Source: https://www.amazon.com/Network-Analysis-Applications-Lecture-Networks/dp/3319781952
Welcome to Colaboratory!

Colaboratory is a free Jupyter notebook environment that requires no setup and runs entirely in the cloud. See our FAQ for more info.

Getting Started

- Overview of Colaboratory
- Loading and saving data: Local files, Drive, Sheets, Google Cloud Storage
- Importing libraries and installing dependencies
- Using Google Cloud BigQuery
- Forms, Charts, Markdown, & Widgets
- TensorFlow with GPU
- Machine Learning Crash Course: Intro to Pandas & First Steps with TensorFlow

Highlighted Features

Seedbank
Looking for Colab notebooks to learn from? Check out Seedbank, a place to discover interactive machine learning examples.

TensorFlow execution

Colaboratory allows you to execute TensorFlow code in your browser with a single click. The example below adds two matrices.

\[
\begin{bmatrix}
1. & 1. & 1.
\end{bmatrix} +
\begin{bmatrix}
1. & 2. & 3.
\end{bmatrix} =
\begin{bmatrix}
2. & 3. & 4.
\end{bmatrix}
\]
Big Data Analytics and Data Mining
Big Data 4 V

Volume

- 40 ZETTABYTES (43 TRILLION GIGABYTES) of data will be created by 2020, an increase of 300 times from 2005
- 6 BILLION PEOPLE have cell phones
- WORLD POPULATION: 7 BILLION

The FOUR V’s of Big Data

- Volume
- Velocity
- Variety
- Veracity

Volume

- It’s estimated that 2.5 QUINTILLION BYTES (23 TRILLION GIGABYTES) of data are created each day
- Most companies in the U.S. have at least 100 TERABYTES (100 BILLION GIGABYTES) of data stored

Velocity

- The New York Stock Exchange captures 1 TB OF TRADE INFORMATION during each trading session
- Modern cars have close to 100 SENSORS that monitor items such as fuel level and tire pressure

Variety

- As of 2011, the global size of data in healthcare was estimated to be 150 EXABYTES (160 BILLION GIGABYTES)
- By 2014, it’s anticipated there will be 420 MILLION WEARABLE, WIRELESS HEALTH MONITORS
- 4 BILLION+ HOURS OF VIDEO are watched on YouTube each month
- 400 MILLION TWEETS are sent per day by about 200 million monthly active users

Veracity

- 1 IN 3 BUSINESS LEADERS don’t trust the information they use to make decisions
- Poor data quality costs the US economy around $3.1 TRILLION A YEAR
- 27% OF RESPONDENTS in one survey were unsure of how much of their data was inaccurate

Veracity

- 27% OF RESPONDENTS in one survey were unsure of how much of their data was inaccurate

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, ENC, SAS, IBM, MEPTEC, QAS

Source: https://www-01.ibm.com/software/data/bigdata/
Value
Artificial Intelligence

Machine Learning & Deep Learning

Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.
Stephan Kudyba (2014),
Big Data, Mining, and Analytics:
Components of Strategic Decision Making, Auerbach Publications

Architecture of Big Data Analytics

Big Data Sources
- * Internal
- * External
- * Multiple formats
- * Multiple locations
- * Multiple applications

Big Data Transformation
- Middleware
- Extract Transform Load
- Data Warehouse
- Traditional Format CSV, Tables
- Raw Data
- Transformed Data

Big Data Platforms & Tools
- Hadoop
- MapReduce
- Pig
- Hive
- Jaql
- Zookeeper
- Hbase
- Cassandra
- Oozie
- Avro
- Mahout
- Others

Big Data Analytics Applications
- Queries
- Reports
- OLAP
- Data Mining

Source: Stephan Kudyba (2014), Big Data, Mining, and Analytics: Components of Strategic Decision Making, Auerbach Publications
Architecture of Big Data Analytics

Data Mining
Big Data Analytics Applications

Big Data Sources
- Internal
- External
- Multiple formats
- Multiple locations
- Multiple applications

Big Data Transformation

Big Data Platforms & Tools
- Queries
- Reports
- OLAP
- Data Mining

Big Data Analytics Applications

Source: Stephan Kudyba (2014), Big Data, Mining, and Analytics: Components of Strategic Decision Making, Auerbach Publications
Social Big Data Mining
(Hiroshi Ishikawa, 2015)

Architecture for Social Big Data Mining
(Hiroshi Ishikawa, 2015)

Enabling Technologies
- Integrated analysis model
- Natural Language Processing
- Information Extraction
- Anomaly Detection
- Discovery of relationships among heterogeneous data
- Large-scale visualization
- Parallel distrusted processing

Analysts
- Model Construction
- Explanation by Model
- Construction and confirmation of individual hypothesis
- Description and execution of application-specific task

Source: Hiroshi Ishikawa (2015), Social Big Data Mining, CRC Press
Data Warehouse

Data Mining and Business Intelligence

Increasing potential to support business decisions

Decision Making

Data Presentation
Visualization Techniques

Data Mining
Information Discovery

Data Exploration
Statistical Summary, Querying, and Reporting

Data Preprocessing/Integration, Data Warehouses

Data Sources
Paper, Files, Web documents, Scientific experiments, Database Systems

End User

Business Analyst

Data Analyst

DBA

Source: Jiawei Han and Micheline Kamber (2006), Data Mining: Concepts and Techniques, Second Edition, Elsevier
The Evolution of BI Capabilities

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Data Mining at the Intersection of Many Disciplines

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Data Mining:
The KDD Process for Extracting Useful Knowledge from Volumes of Data

Data Mining

Knowledge Discovery in Databases (KDD) Process

(Fayyad et al., 1996)

Knowledge Discovery (KDD) Process

Data mining: core of knowledge discovery process

- Data Cleaning
- Data Integration
- Data Warehouse
- Task-relevant Data
- Selection
- Data Mining
- Pattern Evaluation

Source: Han & Kamber (2006)
Data Mining Processing Pipeline
(Charu Aggarwal, 2015)

Data Collection → Data Preprocessing
- Feature Extraction
- Cleaning and Integration
→ Analytical Processing
- Building Block 1
- Building Block 2
→ Output for Analyst

Feedback (Optional)

Source: Charu Aggarwal (2015), Data Mining: The Textbook Hardcover, Springer
# A Taxonomy for Data Mining Tasks

## Learning Method | Popular Algorithms
--- | ---
Supervised | Classification and Regression Trees, ANN, SVM, Genetic Algorithms
Supervised | Decision trees, ANN/MLP, SVM, Rough sets, Genetic Algorithms
Supervised | Linear/Nonlinear Regression, Regression trees, ANN/MLP, SVM
Unsupervised | Apriori, OneR, ZeroR, Eclat
Unsupervised | Expectation Maximization, Apriori Algorithm, Graph-based Matching
Unsupervised | Apriori Algorithm, FP-Growth technique
Unsupervised | K-means, ANN/SOM
Unsupervised | K-means, Expectation Maximization (EM)

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Business Insights with Social Analytics
Analyzing the Social Web: Social Network Analysis
Jennifer Golbeck (2013), Analyzing the Social Web, Morgan Kaufmann
The 14th NTCIR (2018 - 2019)

Evaluation of Information Access Technologies

January 2018 - June 2019

What's New

- February 1, 2018: Call for participation to the NTCIR-14 Kick-Off Event released.
- February 1, 2018: Call for participation to the NTCIR-14 QALab-PoliInfo Kick-Off Event released.

December 5, 2017: The NTCIR-14 Task Selection Committee has selected the following six Tasks.
Lifelig-3, OpenLiveQ-2, QA Lab-4, STC-3, WWW-2, CENTRE.

August 23, 2017: NTCIR-14 Call for Task Proposals released. (Closed.)

About Proceedings

After the NTCIR-14 conference, a post-proceedings of revised selected papers will be published in the Springer Lecture Notes on Computer Science (LNCS) series.

http://research.nii.ac.jp/ntcir/ntcir-14/index.html
NTCIR-14 Short Text Conversation Task (STC-3)

Welcome to the top page of STC-3@NTCIR-14!
STC-3 offers three subtasks:

- [Chinese Emotional Conversation Generation (CECG) Subtask](#)
- Dialogue Quality (DQ) Subtask (for Chinese and English)
- Nugget Detection (ND) Subtask (for Chinese and English)

**Key dates for DQ and ND Subtasks**

Feb-Mar 2018 Crashing Chinese test data from Weibo

Oct 2017-Jan 2018 Training data translation into English
Apr-Jun, 2018 Test data translation into English
Jul-Aug 2018 Training/test data annotation
Aug 31, 2018 STC-3 task registrations due (CECG, DQ, ND)
Sep 1, 2018 Training data with annotations released
Nov 1, 2018 Test data released
Nov 30, 2018 Run submissions due
Dec 20, 2018 Results and draft overview released to participants
Feb 1, 2019 Participant papers due
Mar 1, 2019 Acceptance notification
Mar 20, 2019 All camera-ready papers due
Jun 2019 NTCIR-14 Conference@NII
Short Text Conversation Task (STC-3)
Chinese Emotional Conversation Generation (CECG) Subtask

Call for Participation

In recent years, there has been a rising tendency in AI research to enhance Human-Computer Interaction by humanizing machines. However, to create a robot capable of acting and talking with a user at the human level requires the robot to understand human cognitive behaviors, while one of the most important human behaviors is expressing and understanding emotions and affects. As a vital part of human intelligence, emotional intelligence is defined as the ability to perceive, integrate, understand, and regulate emotions. Though a variety of models have been proposed for conversation generation from large-scale social data, it is still quite challenging (and yet to be addressed) to generate emotional responses.

In this challenge, participants are expected to generate Chinese responses that are not only appropriate in content but also adequate in emotion, which is quite important for building an empathic chatting machine. For instance, if user says “My cat died yesterday”, the most appropriate response may be “It’s so sad, so sorry to hear that” to express sadness, but also could be “Bad things always happen, I hope you will be happy soon” to express comfort.

Previous Evaluation Challenge at NLPCC 2017

Overview of the NLPCC 2017 Shared Task: Emotion Generation Challenge

http://www.aihuang.org/p/challenge.html
Short Text Conversation (NTCIR-13 STC2) Retrieval-based

retrieval-based method

Given a new post, can a coherent and useful comment be returned by searching a post-comment repository?

post-comment repository

Short Text Conversation
(NTCIR-13 STC2)
Generation-based

generation-based method

Given a new post, can a fluent, coherent and useful comment be generated?

The Trained Generator

Generated comment
Generated comment
Generated comment

post-comment repository

Summary

• This course introduces the fundamental concepts and research issues of Big Data Mining.

• Topics include
  – ABC: AI, Big Data, Cloud Computing,
  – Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data,
  – Fundamental Big Data: MapReduce Paradigm, Hadoop and Spark Ecosystem,
  – Foundations of Big Data Mining in Python,
  – Supervised Learning: Classification and Prediction,
  – Unsupervised Learning: Cluster Analysis,
  – Unsupervised Learning: Association Analysis,
  – Machine Learning with Scikit-Learn in Python,
  – Deep Learning for Finance Big Data with TensorFlow,
  – Convolutional Neural Networks (CNN)
  – Recurrent Neural Networks (RNN)
  – Reinforcement Learning (RL)
  – Social Network Analysis (SNA)
Big Data Mining

Contact

Min-Yuh Day, Ph.D.
Assistant Professor
Department of Information Management,
Tamkang University

Tel: 886-2-26215656 ext. 2846
Fax: 886-2-26209737
Office: B929
Address: No.151, Yingzhuan Rd., Danshui Dist.,
New Taipei City 25137, Taiwan (R.O.C.)
Email: myday@mail.tku.edu.tw
Web: http://mail.tku.edu.tw/myday/