Postgraduate Course
Large-scale Media Analytics (MSc)

Instructor Information

Federico Alvarez

E-mail: federico.alvarez@upm.es
Work Phone: +34 91 549 57 00 Ext: 8073

Course Information

Course Description

Current data analysis applications require the management of extremely large collections of heterogeneous multimedia data. The extraction of knowledge from these huge datasets is a difficult problem with a broad scope.

This subject aims at presenting the most relevant techniques and methodologies for large scale multimedia analysis. In particular, we will discuss the application of widely used machine learning techniques (dimensionality reduction, classification, clustering) to textual, image and spatio-temporal data. Heterogeneous information networks and suitable data mining techniques will also be described.

Big data technologies will be introduced, including efficient acquisition, storage and processing of huge amounts of heterogeneous data.

Some of the described techniques will be applied to relevant use cases, such as web search and recommendation problems.

Practical sessions will be proposed in which students will apply these tools to real datasets and become familiar with powerful analysis frameworks.

Prerequisites

- Statistical modelling

- Notions of Python and Java will be valuable

Related courses

- Predictive and descriptive learning

- Machine learning lab

- Data science foundations and applications

Course Goal

To develop an understanding of the concepts and main techniques for multimedia analysis in large-scale environments.

Course objectives:

By the end of the course, students should:

- Be able to select and apply adequate machine learning techniques to large-scale multimedia datasets and evaluate their performance.
- Be familiar with Big Data technologies and their application to multimedia content.
- Be able to develop basic applications in relevant current use cases in the industry (web search, recommendation, etc).

Program

1. ***Introduction to multimedia analytics (4 hours):***
   - *Multimedia content analysis and applications.*
   - *Data wrangling and exploratory data analysis.*
2. ***Machine learning for multimedia content (12 hours):***
   - *Text and speech data (natural language processing): topic extraction, summarization, translation, etc.*
   - *Image and video (machine vision): feature extraction, pattern recognition, content analysis, etc.*
   - *Spatio-temporal data: regression, clustering, classification.*
   - *Heterogeneous information networks (internet, social media): similarity search, recommendation, prediction, etc.*
   - *Exercises in Python (4 hours).*
3. ***Big data engineering (8 hours):***
   - *Data acquisition and storage: relational, graph databases, file systems.*
   - *Centralized vs parallel computing.*
   - *Technologies and functionalities: MapReduce, Hadoop, Spark, Scala.*
   - *Exercise in Java Mahout (2 hours)*
4. ***Use case 1: search and ranking (6 hours)***
   - *Information retrieval*
   - *Web search: PageRank algorithm (Google)*
   - *Exercise in Python (2 hours).*
5. ***Use case 2: recommendation (6 hours):***
   - *Collaborative filtering (KNN)*
   - *Content-based filtering*

- *Exercises in Java or Python (2 hours).*

6. ***Presentations (4 hours):***
   - *10 papers overview (20 minutes each)*

Textbooks:

1. Segaran, T. (2007). *Programming collective intelligence: building smart web 2.0 applications*. " O'Reilly Media, Inc.".
2. Aggarwal, C. C., & Zhai, C. (2012). *Mining text data*. Springer Science & Business Media.
3. Schutt, R., & O'Neil, C. (2013). *Doing data science: Straight talk from the frontline*. " O'Reilly Media, Inc.".
4. McKinney, W. (2012). *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. " O'Reilly Media, Inc.".

Papers:

[1] Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank citation ranking: bringing order to the web.

[2] Bennett, J., & Lanning, S. (2007, August). The netflix prize. In *Proceedings of KDD cup and workshop* (Vol. 2007, p. 35).

[3] Shvachko, K., Kuang, H., Radia, S., & Chansler, R. (2010, May). The hadoop distributed file system. In *Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium on* (pp. 1-10). IEEE.

[4] Schafer, J. B., Konstan, J., & Riedl, J. (1999, November). Recommender systems in e-commerce. In *Proceedings of the 1st ACM conference on Electronic commerce* (pp. 158-166). ACM. (1526 refs)

[5] Resnick, P., & Varian, H. R. (1997). Recommender systems. *Communications of the ACM*, *40*(3), 56-58. (3500 refs)

[6] Miller, B. N., Albert, I., Lam, S. K., Konstan, J. A., & Riedl, J. (2003, January). MovieLens unplugged: experiences with an occasionally connected recommender system. In *Proceedings of the 8th international conference on Intelligent user interfaces* (pp. 263-266). ACM. (500 refs)

[7] Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. *Communications of the ACM*, *51*(1), 107-113.

[8] Ghemawat, S., Gobioff, H., & Leung, S. T. (2003, October). The Google file system. In *ACM SIGOPS operating systems review* (Vol. 37, No. 5, pp. 29-43). ACM.

[9] DeCandia, G., Hastorun, D., Jampani, M., Kakulapati, G., Lakshman, A., Pilchin, A., ... & Vogels, W. (2007, October). Dynamo: amazon's highly available key-value store. In *ACM SIGOPS Operating Systems Review* (Vol. 41, No. 6, pp. 205-220). ACM.
[10] Chang, F., Dean, J., Ghemawat, S., Hsieh, W. C., Wallach, D. A., Burrows, M., ... & Gruber, R. E. (2008). Bigtable: A distributed storage system for structured data. *ACM Transactions on Computer Systems (TOCS)*, *26*(2), 4.

[11] Lakshman, A., & Malik, P. (2010). Cassandra: a decentralized structured storage system. *ACM SIGOPS Operating Systems Review*, *44*(2), 35-40.

[12] Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, *55*(10), 78-87.

[13] Dean, J., & Ghemawat, S. (2010). MapReduce: a flexible data processing tool. *Communications of the ACM*, *53*(1), 72-77.

[14] Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., ... & Zhou, Z. H. (2008). Top 10 algorithms in data mining. *Knowledge and information systems*, *14*(1), 1-37.

[15] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, *12*, 2825-2830.

Student Assessment Criteria

Evaluation:

- Attend and follow the theory and practical sessions and hand in a report for each lab session.
- Read and extract the main ideas from a relevant paper in the field, and make a presentation to the group (15 minutes plus 5 minutes discussion).
- Test: 30 questions on the theoretical content of the course, selected papers and code.

Lab sessions	50%
Overview of a selected paper	20%
Test	30%

Since it is often difficult to develop simple closed-form solution, graded projects will be assigned throughout the semester involving the development of computer programs to simulate and study the presented applications.

SW libraries in different languages, and BigData architectures will be used to guide the student to develop solutions for industrial applications of the subject presented techniques.