UNIVERSIDADE FEDERAL DE SANTA CATARINA
Centro Tecnológico
Departamento de Informática e Estatística
Programa de Pós-Graduação em Ciência da Computação

COURSE PROGRAM

A critério do professor, não havendo alunos estrangeiros matriculados, a disciplina poderá ser ofertada em língua portuguesa.

1) Identification

Course: INE410136 - *Content Detection and Analysis on Big Web Data*
Credits: 30 hours/class - 2 credits
Semester: 2018/2
Professor: Carina Friedrich Dorneles

2) Level: Master and Doctorate

3) Prerequisites: None

4) Syllabus: Introduction to Big Data (a large amount of data view). Useful content detection. Document similarity and deduplication. Content extraction and feature selection. Web crawling and web scraping. Entity resolution. Understanding Web metadata. Named entity recognition. Web analytics.

5) Objectives

General: The general objective of this course is to introduce students to the area of content detection and analysis. This involves a general view of big data (considering non-structured data), understanding of data formats, their content detection and data extraction from them.

Specifics:

- Introducing big data in a non-structured data point of view (only as a large amount of data);
- Introducing analyzes and identification of useful data sources, using web crawling/scraping, detecting useful data from them including their text and metadata;
- Introducing data similarity and deduplication, entity resolution and entity named recognition;
- Introducing Web analytics, characterizing Web data science, understanding Web metadata, getting the data, interpreting it, organizing it, and learning from it.

5) Outline:

1. Introduction to Big Data (a large amount of data view).
   - 1.1 What does big data means? Where is it?
   - 1.2 Big data tools and techniques
   - 1.3 Basic Data Manipulation and Analysis
2. Content detection.
   - 2.1 Dark Data
   - 2.2 Data Extraction
   - 2.3 Web Crawling
   - 2.4 Web Scraping
   - 2.5 Useful versus noise content

2.6 Content extraction
2.7 Feature selection
2.8 Named entity recognition.

3. Similarity and deduplication.
   3.1 Document similarity
   3.2 Entity resolution.
4. Web analytics.
   4.1 Web Data science
   4.2 Understanding Web metadata.
   4.3 Getting the data, interpreting it, organizing it, and learning from it

6) Bibliography:

- Xin Luna Dong , Divesh Srivastava, Big data integration, Proceedings of the VLDB Endowment, v.6 n.11, p.1188-1189, August 2013.
- Aisha Siddiqa et. al. A survey of big data management: Taxonomy and state-of-the-art, Journal of Network and Computer Applications, Vol 71, 2016, Pages 151-166.
- Mattmann, Chris. A vision for data science. Nature, Vol. 493, No. 7433, pp. 473-475, January 24, 2013.
- Manku, Gurmeet Singh, Arvind Jain, and Anish Das Sarma. "Detecting near-duplicates for web crawling." Proceedings of the 16th international conference on World Wide Web. ACM, 2007
- Heymann, Paul and Paepcke, Andreas and Garcia-Molina, Hector (2010) Tagging Human Knowledge. In: Third ACM International Conference on Web Search and Data Mining (WSDM2010), February 3-6, 2010, New York City, NY, USA.
- Whang, Steven Euijong and Garcia-Molina, Hector Managing Information Leakage. In: CIDR 2011.

7) Complementary Bibliography:

- De Mauro, Andrea, Marco Greco, and Michele Grimaldi. "What is big data? A consensual definition and a review of key research topics." AIP conference proceedings. Vol. 1644. No. 1. AIP, 2015.
- Challenges and Opportunities with Big Data, A community white paper developed by leading researchers across the US, 2012. http://cra.org/ccc/docs/init/bigdatawhitepaper.pdf
- http://rtw.ml.cmu.edu/rtw/overview
- Up to date journal articles and events papers.