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


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 mean? 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.

7) Complementary Bibliography:

- 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://cra.org/ccc/docs/init/bigdatawhitepaper.pdf)
- [http://rtw.ml.cmu.edu/rtw/overview](http://rtw.ml.cmu.edu/rtw/overview)
- Up to date journal articles and events papers.