

MSc Data Science for Business

Partnership with Capgemini Consulting

June 2018



We give students the opportunity to work on a practical case study, enhancing both their business and data science skills



MSc Data Science for
Business

Practical case study

Text mining case (**2017**: detecting and analyzing public transportation fraud through social media; **2018**: identifying the most common smartphone defaults, shared by customers on phone websites, forums, ...) using state-of-the art methodologies to collect, format and analyze text data. The course spread over **8 weekly sessions** (7 in 2017).

Multidisciplinary course

Focus both on the **data science methodologies** (scraping, preparation & text analyses) and **business aspects** (project management & business case), through the **roll out of a complete case study**, concluded with a hackathon.

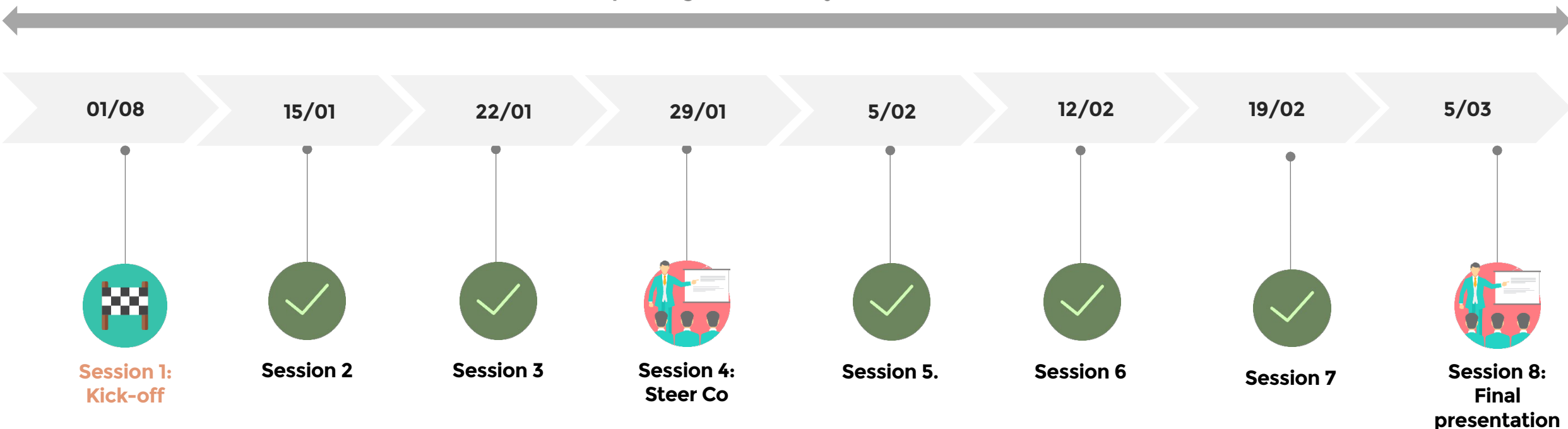
Dynamic animation

Facilitation alternating between **plenary sessions** (to upskill students on technical & business contents), **group hands-on** sessions and **brainstorming** (to build the different analyses & present results).



Planning & key steps of the course



8 weeks spanning from January to mid-March 2019



Légende :

-  **Regular sessions:** Synthetic status update on the progress made, plenary sessions and coding hands-on sessions
-  **Steering Committees:** Complete pres. of the progress made, results; difficulties and foreseen next steps to be prepared

Syllabus for the 2019 session



Evaluation modalities: a mid-term oral presentation and a final oral presentation per group

Course 1

- Course main objectives
- What is a data use case?
- Presentation of the Use Case on which the students will work through the data camp
- Text mining: plenary and hands-on

Course 2

- Text cleaning/processing: plenary & hands-on

Course 3

- Dimensionality reduction
- Topic extraction with LDA / Graph of Words

Course 4

- Sharing of topics found
- Introduction to word2Vec/Doc2VEc
- Sentiment analysis: plenary and hands-on

Course 5

- Each group of students has to make a 20-minute presentation of type “Steering Committee” to present their work

Course 6

- Feedback on steering committee
- Semi supervised learning : plenary and hands-on session

Course 7

- Business Case
- Launch of the Hackathon

Course 8

- Final restitution by the students
- Hackathon results

Reading list



Text mining – General

- Feature Engineering and Selection
CS 294: Practical Machine Learning. October 1st, 2009.
Alexandre Bouchard-Côté
<https://people.eecs.berkeley.edu/~jordan/courses/294-fall09/lectures/feature/slides.pdf>
- Stemming and lemmatization
An Introduction to Information Retrieval, 2009.
Christopher D. Manning, Prabhakar Raghavan & Hinrich Schütze
<http://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html>

Topic extraction with LDA

- Latent Dirichlet Allocation, David M. Blei, Andrew Y. Ng, Michael I. Jordan
(<http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf>)

Naïve Bayes for text classification

- Tackling the Poor Assumptions of Naive Bayes Text Classifiers, Jason D. M. Rennie, Lawrence Shih, Jaime Teevan, David R. Karger (<http://www.aaai.org/Papers/ICML/2003/ICML03-081.pdf>)
- A Comparison of Event Models for Naive Bayes Text Classification, Andrew McCallum, Kamal Nigam ([link to pdf](#))

Word2Vec/ Doc2Vec

- Word2vec Parameter Learning Explained, Xin Rong (<https://arxiv.org/pdf/1411.2738.pdf>)
- An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation, Jey Han Lau and Timothy Baldwin (<https://arxiv.org/pdf/1607.05368.pdf>)

Graph of Words

- Graph-of-word and TW-IDF: New Approach to Ad Hoc IR, François Rousseau, Michalis Vazirgiannis (<https://frncsrss.github.io/papers/rousseau-cikm2013.pdf>)
- Text Categorization as a Graph Classification Problem, François Rousseau Emmanouil Kiagias LIX, Michalis Vazirgiannis (<http://www.aclweb.org/anthology/P15-1164>)
- Lecture notes by Michalis Vazirgiannis ([link to pdf](#))

Semi supervised learning

- Semi-Supervised Learning: Literature Survey, Xiaojin Zhu ([link to pdf](#))
- Word representations: A simple and general method for semi-supervised learning, Joseph Turian, Lev Ratinov, Yoshua Bengio ([link to pdf](#))

Appendix



A syllabus designed to teach students how to lead a text mining project



2017 syllabus

Course 1

- Course main objectives
- What is a data use case?
- Web scraping: plenary and hands-on

Course 2

- Review on progress made (scraping)
- Text cleaning/processing: plenary and hands-on

Course 3

- Review on progress made (text cleaning)
- Dimensionality reduction
- Unsupervised classification for topic extraction: k-means and spherical k-means

Course 4

- Each group of students made a 20-minute presentation of type “Steering Committee” to present their work
- Semi supervised learning: plenary and hands-on sessions

Course 5

- Business case

Course 6

- Intervention by Google: How to measure KPIs

Course 7

- Introduction to Word2Vec and Graph of Words
- Hackathon results

2018 syllabus

Course 1

- Course main objectives
- What is a data use case?
- Text mining: plenary and hands-on

Course 2

- Text cleaning/processing: plenary & hands-on

Course 3

- Dimensionality reduction
- Topic extraction with LDA

Course 4

- Sharing of topics found
- Topic analysis: Graph of Words
- Sentiment analysis: plenary and hands-on

Course 5

- Each group of students made a 20-minute presentation of type “Steering Committee” to present their work

Course 6

- Feedback on steering committee
- Semi supervised learning : plenary and hands-on session

Course 7

- Business Case

Course 8

- Final restitution by the students
- Hackathon results