MOHAMMED VI POLYTECHNIC UNIVERSITY | FACULTÉ DE GOUVERNANCE, SCIENCES ECONOMIQUES ET SOCIALES

Syllabus :

Analysis in Data Science (DAT201P)

1. Course:

Analysis in Data Science (DAT201P)

2. Instructor:

Amir ABDUL REDA

I am an Assistant Professor in computational political science at Mohammed VI Polytechnic University (Rabat, Morocco). My research focuses on public opinion, social movements, and migration using big data and Natural Language Processing. In the past, I did my PhD at the University of Toronto and my work is published (or forthcoming) with Sociological Method and Research, Comparative Sociology, and Middle Eastern Studies.

3. Learning Objectives:

This course largely focuses on teaching introductory level techniques of data science—with a particular focus on a subfield of machine learning, Natural Language Processing (NLP)—but also offers a much-needed overview of the ethics of these techniques today. By the end of this course, students should have a solid understanding of the techniques of data science and the ethical debates that surround them. To be precise:

- Students should have a good understanding of the tools used by social scientists to collect and clean data using Python in the information age—which is what the Data Analysis & Cleaning Section of this class will focus on. They will be exposed to the codes used to this end but will not be expected to master them.
- Students should have a good understanding of the various methods available to conduct computer assisted analysis of texts—which is what the Natural Language Processing segment of this class will focus on. Here again, they will be exposed to the codes used to this end but will not be expected to master them.
- Students should have an introductory understanding of the ethics involved with data science today—which is what the ethics portion of the course will focus on.

4. Evaluation:

4.1 Midterm Exam (30 %):

The midterm exam will cover material from the first half of the course. Both readings and classroom discussions will be considered in the elaboration of the midterm exam, which will count for 30% of the final grade. Its date and time will be fixed by the faculty and duly communicated to students in advance.

4.2 Final Exam (30%):

The final exam will cover material from the entire course. It will make up 30% of the final grade. Its date and time will be fixed by the faculty and duly communicated to students in advance.

4.3 Coding Assignments (15%; 7.5% per homework)

The Python coding component of this course will be evaluated with problem sets that serve to reinforce the students' understandings of the codes they will be exposed to throughout this course. A final problem set will be delivered to students later in the semester to test their general understanding of all the codes introduced in this class. To prepare for this final problem set, students are encouraged to duly complete all the previous ones. This assignment will be graded on effort and completion, rather than accuracy.

4.4 Oral Presentations (15%)

Students will be given a grade out of 10% for an oral presentation of the readings of their choice for one of the classes. Students can decide to fulfil this assignment in groups in case there are enough students in the class to cover all classes in groups of two to four students maximum. A brief, 5-minute presentation is expected per group whereby each student is expected to present an equivalent part of the readings in question.

4.5 Student participation (10%)

Course attendance is mandatory and to encourage both assiduity and in-class participation 10% of the final grade for the course will be allocated to student's active participation in the course. Engaging in class discussions and attending the course on time will ensure a good evaluation on this component of the final grade.

5. Textbooks:

This course will largely rely on the following textbooks—they are freely available online and links to them will be provided throughout the syllabus. None of them exhaustively explore the topic at hand but they remain key in the literature—as such, journal articles are provided to complement them, when necessary, and links to these journals are also provided in the syllabus.

- Bird, Steven, Ewan Klein and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media.
  - This book is a very accessible introduction to NLP in Python. It is freely available online.
- Manning, Christopher D., Prabhakar Raghavan and Hinrich Schütze. 2009. *An Introduction to Information Retrieval*. Cambridge: Cambridge University Press.
  - This book also covers many of the topics of this course. An online version is available here.
- Jurafsky, Daniel and James H. Martin. 2008. *Speech and Language Processing*. New Jersey: Prentice Hall.
  - This important textbook also covers complementary topics. An online version is available here.
- Steinert-Threlkeld, Zachary C. 2018. *Twitter as Data*. Cambridge: Cambridge University Press.
  - This is a very useful technical manual for those who would like to go deeper into the use of Twitter as data. It is very approachable and very useful for learning the codes and concepts associated with the use of Twitter as data. A free, online version is available here.

6. Software:

The techniques introduced in this class will be taught using the Python 3 programming language. Python is the most popular programming language and freely available for download online—as a result, it has a very large support community onli

7. Final Grade Breakdown:

Student Participation: 10%
Oral Presentation: 15%
Coding Assignments: 15%
Midterm Exam: 30%
Final Exam: 30%

7.1 Late assignments:

Late assignments will receive a 2% lateness penalty for each 24 hours of lateness unless the lateness is justified with a doctor's note and/or discussed with the tacit approval of the professor. Students are warmly encouraged to approach the professor should they have extraordinary circumstances that prevent them from successfully submitting their assignment in time, in particular because of the extraordinary circumstances of our current times.

8. Detailed Course Outline:

Classes	Detailed Contents & Evaluations
Session 1 (2 hours)	Class Title: **Data Science & Big Data: an introduction (1.5 hours)**

	Class Objectives:
	• Ontology
	• Why data science
	  ○ For governments
	  ○ For the industry
	  ○ For academia
	• Opportunities & challenges
	• Outline of the Class:
	  ○ Data Mining/Information Retrieval
	  ○ NLP & Machine Learning
	  ○ Ethics of Data Science
	
	Mandatory Readings:
	• Syllabus
	• Philosophical (ontological) introduction to the logic of data science & natural language processing:
	  ○ Chapter 12 of Bird, Steven, Ewan Klein and Edward Loper. 2009.
	• Gary King on the importance of the data we use in data science for social scientific research and the evolution of science. Link here.
	
	Class Title:
	"Information Retrieval": Introduction (0.5 hours)
	
	Class Objectives:
	• What is data mining/information retrieval?
	• How do we data mine?
	• Why do we use computers & programing languages?
	  ○ Introduction to Python, discussion of R
	  ▪ Why Python and not another language?
	  ▪ History of Python
	  ▪ Installing Anaconda Python and basics
	  • Link to the installer
	
	Mandatory Readings:
	• For data mining/information retrieval
	  ○ Read Chapter 1 of Manning, Christopher D., Prabhakar Raghavan and Hinrich Schütze. 2009.
	• Read Preface "Why Python" of Bird, Steven, Ewan Klein and Edward Loper. 2009.
	Recommended Readings:
	• DO NOT install Pycharm, but watch this 45 minutes introductory video to Python
	• Online documentation for Python

Session 2 (2 hours)	Class Title:
	"Information Retrieval": Introduction (1 hour)
	
	Class Objectives:
	• What is data mining/information retrieval?
	• How do we data mine?
	• Why do we use computers & programing languages?
	  ○ Introduction to Python, discussion of R
	  ▪ Why Python and not another language?
	  ▪ History of Python
	  ▪ Installing Anaconda Python and basics
	  • Link to the installer
	
	Mandatory Readings:
	• For data mining/information retrieval

	○ Read Chapter 1 of Manning, Christopher D., Prabhakar Raghavan and Hinrich Schütze. 2009. • Read Preface "Why Python" of Bird, Steven, Ewan Klein and Edward Loper. 2009. Recommended Readings: • DO NOT install Pycharm, but watch this 45 minutes introductory video to Python • Online documentation for Python Class Title: **Data Structure (0.5 hour)** Class Objectives: • What does the data we use look like? ○ Parliamentary texts ○ Webpages ○ Social Media data ▪ Twitter, Facebook, Reddit ○ Open Ended Interviews • Python practice ○ Python's GUIs ▪ Jupyter Notebook & Spider ○ Python Packages/Libraries ▪ What are they? ▪ Installation of a range of basic libraries ▪ Numpy, scipy, matplotlib, sklearn, nltk etc. Mandatory Readings: • Read Preface "Python 3 and NLTK3" & "Natural Language Toolkit (NLTK)" of Bird, Steven, Ewan Klein and Edward Loper. 2009. • Chapter 1 of Bird, Steven, Ewan Klein and Edward Loper. 2009. ○ Sections 1.1 to 1.2
Session 3 (2 hours)	Class Title: **Data Structure (1 hour)** Class Objectives: • What does the data we use look like? ○ Parliamentary texts ○ Webpages ○ Social Media data ▪ Twitter, Facebook, Reddit ○ Open Ended Interviews • Python practice ○ Python's GUIs ▪ Jupyter Notebook & Spider ○ Python Packages/Libraries ▪ What are they? ▪ Installation of a range of basic libraries ▪ Numpy, scipy, matplotlib, sklearn, nltk etc. Mandatory Readings: • Read Preface "Python 3 and NLTK3" & "Natural Language Toolkit (NLTK)" of Bird, Steven, Ewan Klein and Edward Loper. 2009. • Chapter 1 of Bird, Steven, Ewan Klein and Edward Loper. 2009. ○ Sections 1.1 to 1.2 Class Title: **Data Analysis & Cleaning with Python I (1 hour)** Class Objectives:

	• Data types in Python
	  • Numeric & String Data Types
	• Lists & Dictionaries
	• Python practice:
	  ○ Loading datasets
	  ○ Intro to Pandas library
	  ○ Data encoding & format of various files
	• What is data cleaning?
	• How do we clean data with Python?
	  Basic steps & codes in data cleaning
	  Chapter 3 and 5 of Zelle 2016
	
	Mandatory Readings:
	• Chapter 1 of Bird, Steven, Ewan Klein and Edward Loper. 2009.
	  ○ Sections 1 to 3
	• Chapter 2 & Chapter 11 of Bird, Steven, Ewan Klein and Edward Loper. 2009.
	  ○ Sections 1 & 2 of Chapter 2
	  ○ Section 3 of Chapter 11—in particular 3.1-3.4. Glance over the codes, focus on the text for now.

Session 4
(3 hours)
	Class Title:
	Data Analysis & Cleaning with Python II: Applications (1.5 hours)
	
	Class Objectives:
	• Focus of the class
	  ○ Academic applications & loading the data
	• Parliamentary texts
	  ○ Literature
	• Web Scraping
	  ○ Literature
	• Streaming Twitter & Facebook
	  ○ Literature
	What it looks like with Python
	
	Mandatory Readings:
	• For streaming Twitter
	  ○ Have a brief look at Steinert-Threlkeld 2018 chapters 1-4.
	  • For Facebook
	  ○ CrowdTangle
	• Parliamentary texts
	  ○ Rheault et al. 2016. Measuring Emotion in Parliamentary Debates with Automated Textual Analysis.
	• Web Scraping
	  ○ Ren & Ren 2018. A Framework of Petroleum Information Retrieval System Based on Web Scraping with Python.
	Recommended Readings:
	• For streaming Twitter
	  ○ Gonzalez Bailon et al. 2011. The dynamics of Protest Recruitment through an Online Network
	  ○ Barbera & Steinert-Threlkeld. 2019. How to Use Social Media Data for Political Science Research.
	  ○ Steinert-Thelkeld & al. 2015. Online Social Networks and Offline Protest.
	  • For Facebook
	  ○ Berriche & Altay 2020
	• Parliamentary texts
	  ○ Spirling, Arthur. 2016. "Democratization and Linguistic Complexity: The Effect of Franchise Extension on Parliamentary Discourse, 1832–1915.

	○ Greene, Derek, and James P. Cross. 2017. "Exploring the Political Agenda of the European Parliament Using a Dynamic Topic Modeling Approach." • Web Scraping ○ Boeing & Waddell. 2016. New Insights into Rental Housing Markets across the United States: Web Scraping and Analyzing Craigslist Rental Listings. ○ Further reading in the aim of coding web scraping: Mitchell, Ryan. 2015. Web Scraping with Python. Sebastopol: O'Reilly Media. Class Title: **Natural Language Processing & Machine Learning I (1.5 hours)** Class Objectives: • What is machine learning? ○ Definition & History ▪ Knowledge base approach ▪ Machine learning ○ Key concepts & drawbacks ▪ Logistic regression & Naïve Bayes ▪ Supervised & Unsupervised learning ▪ Fitting and overfitting ▪ ML as a black box? • Annotating texts & intercoder reliability ○ The concept ○ Examples from teaching computers to detect/read written numbers • ML meets texts: the birth of NLP ○ Transition to next week's topic. ○ An example: Sentiment Analysis • Python Practice (if time allows): ○ Functions & conditional statements I ○ For arguments and for loops Mandatory Readings: ○ Introduction of Goodfellow et al. 2016. *Deep Learning.* ○ Chapter 5 & Chapter 7 of Jurafsky & Martin 2008 for logistic regression ○ For Python practice: Chapter 1 of Bird, Steven, Ewan Klein and Edward Loper. 2009. ▪ Section 4
Session 5 (2 hours)	Class Title: **Natural Language Processing Introduction 1: (1.5 hours)** Class Objectives: • Definition: What is NLP? • History of NLP • NLP models ○ Overview ○ Sentiment analysis ○ Topic Modeling: ▪ NMF & LDA ▪ Short Text Topic Modeling with GSDMM • Python Practice: ○ Functions & conditional statements II ○ If argument and using if with for loops Assignments: • Coding assignment #1 is due Mandatory Readings:

	○ For Python practice: Chapter 1 & Chapter 2 of Bird, Steven, Ewan Klein and Edward Loper. 2009. ▪ Section 4 of Chapter 1 & 3.2 of Chapter 2 ○ On Sentiment ▪ Nasukawa and Yi 2003 ○ On NMF ▪ Lee & Seung 1999 Recommended Readings: ○ On Sentiment ▪ Taboada & al. 2011 ▪ Bakshi et al. 2016 ▪ Lin and He 2009 ▪ Hutto and Gilbert 2014 for VADER ○ On NMF ▪ Arora et al 2012 ▪ O'Callaghan et al. 2015 ○ On LDA ▪ Scientific article: ▪ Blei, Ng and Jordan, "Latent Dirichlet Allocation", *Journal of Machine Learning Research* 3(2003): 993-1022. ▪ Look at this article and this article. ○ On Short Text Topic Modeling ▪ Read this article. ▪ The original scientific article is here. Class Title: **Natural Language Processing Introduction II (0.5 hours)** Class Objectives: • Linguistic theory: ○ Part of speech tagging ○ Collocations: Unigrams, bigrams, *n* grams • Python Practice: Functions & conditional statements III Else arguments and using it with if and for loops Mandatory Readings: ○ Chapter 8 of Jurafsky & Martin 2008 for part of speech tagging ○ Chapter 5, Chapter 6, Chapter 8 of Bird, Steven, Ewan Klein and Edward Loper. 2009. ▪ Sections 1 & 2 of Chapter 5 ▪ Section 1.4 of Chapter 6 ▪ Section 2 of Chapter 8 for n grams & bigrams
Session 6 (3 hours)	Class Title: **Natural Language Processing Introduction II (1 hour)** Class Objectives: • Linguistic theory: ○ Part of speech tagging ○ Collocations: Unigrams, bigrams, *n* grams • Python Practice: Functions & conditional statements III Else arguments and using it with if and for loops Mandatory Readings: ○ Chapter 8 of Jurafsky & Martin 2008 for part of speech tagging ○ Chapter 5, Chapter 6, Chapter 8 of Bird, Steven, Ewan Klein and Edward Loper. 2009. ▪ Sections 1 & 2 of Chapter 5

	▪ Section 1.4 of Chapter 6 ▪ Section 2 of Chapter 8 for n grams & bigrams Class Title: **Natural Language Processing Introduction III (2 hours)** Class Objectives: • Text Normalization: o Stemming & Lemmatization o Tokenization & Grammar parsing o Stop words • Applications of NLP: o Named Entity Recognition • Python Practice: o Twitter data & the JSON format o Functions & conditional statements IV o While arguments and using it with else, if, and for. Defining functions and how to use them. Mandatory Readings: o Chapter 2 of Manning, Raghavan & Schutze 2008 for Lemmatization, Tokenization & stop words o Chapter 7 (section 5 & 7) of Bird, Steven, Ewan Klein and Edward Loper. 2009. For named entity recognition Recommended Readings: o Chapter 2 of Jurafsky & Martin 2008 for Lemmatization & grammar parsing (amongst others)
Session 7 (2 hours)	Class Title: **Natural Language Processing: Lexicons for Sentiment Analysis (1.5 hours)** Class Objectives: • What are Lexicons? o Examples from literature o How we use them: overview of literature • Creating & Using them o Practical examples Mandatory Readings: o Chapter 21 of Jurafsky & Martin 2008 for Lexicons Class Title: **Natural Language Processing: Vector Space Models (0.5 hours)** Class Objectives: • Parametric Indexes & Zone Indexes • Weighting • Vector Space Models & its applications Assignments: • Coding assignment #2 is due Mandatory Readings: o Chapter 6 of Manning, Raghavan & Schutze 2008 for parametric and zone indexes, weighting and vector space models.
Session 8 (3 hours)	Class Title: **Natural Language Processing: Vector Space Models (1 hour)**

	Class Objectives: • Parametric Indexes & Zone Indexes • Weighting • Vector Space Models & its applications Assignments: • Coding assignment #2 is due Mandatory Readings: ○ Chapter 6 of Manning, Raghavan & Schutze 2008 for parametric and zone indexes, weighting and vector space models. Class Title: **Natural Language Processing & Machine Learning II (2 hours)** Class Objectives: • Reminder: ○ Sentiment Analysis ○ The logic of sentiment analysis • Applications of sentiment analysis ○ In industry ○ In government ○ In finance: ▪ The importance of sentiment in finance ▪ Malkiel 1973 and the random walk hypothesis ▪ Emotions & Finance ▪ CNN Fear & Greed Index • In academic literature ○ Overview of some articles that use it • Python practice ○ Introducing some sentiment analysis libraries with focus on VADER • Sentiment analysis & different languages—does it work as well? Machine Translations Mandatory Readings: • Sentiment analysis & social sciences ○ Young, Lori, and Stuart Soroka. 2012. "Affective News: The Automated Coding of Sentiment in Political Texts." • Readings on Machine translations ○ Mohammad, Salameh & Kiritchenko 2016 Recommended Readings: • Sentiment analysis & social sciences ○ Bollen, Johan, Huina Mao, and Alberto Pepe. 2011. Modeling Public Mood and Emotion: Twitter Sentiment and Socio-Economic Phenomena. • Readings on VADER ○ The package ○ Finance applications. ○ Applications in research: ▪ Hutto & Gilbert 2014 ▪ Some other web resources here and here. • Readings on Machine translations ○ Refaee & Rieser 2015 ○ Balahur & Turchi 2013
Session 9 (2 hours)	Class Title: **Natural Language Processing & Supervised Learning**

	Class Objectives:
	• Reminder on Supervised vs Unsupervised Learning
	• Naïve Bayes classifiers
	  ○ "Bag of words" approach
	What can algos tell us?
	
	Mandatory Readings:
	• Chapter 13 of Manning, Raghavan & Schutze 2008 for Naïve Bayes & text classification
	Recommended Readings:
	• Chapter 4 of Jurafsky & Martin 2008 for Naïve Bayes
	• Chapter 6 of of Bird, Steven, Ewan Klein and Edward Loper. 2009.
	  ○ Read through the models for general understanding, focus in particular on section 7.1 "What models tell us" & 8 "Summary"
	Class Title:
	Natural Language Processing & Unsupervised Learning
	
	Class Objectives:
	• Reminder on Unsupervised vs Supervised Learning
	• Hierarchical Agglomerative Clustering
	• Divisive Clustering
	
	Mandatory Readings:
	Chapter 17 of Manning, Raghavan & Schutze 2008 for Hierarchical Clustering

Session 10 (3 hours)	Class Title:
	Natural Language Processing & Ethics I
	
	Class Objectives:
	• Social media data & privacy
	  ○ The sector of business intelligence, NLP, and social media
	  ○ Cambridge Analytica Scandal
	  ○ Crimson Hexagon
	• If possible, invited speaker from the business intelligence field
	
	Mandatory Readings:
	• Chapter 6 of Salganik. 2018. Bit by Bit.
	  ○ First half
	
	Class Title:
	Machine Learning & Ethics II
	
	Class Objectives:
	• Video & photo data & privacy
	• If possible, invited speaker from that field
	
	Mandatory Readings:
	• Chapter 6 of Salganik. 2018. Bit by Bit.
	  ○ Second half

10. TDs:

Tutorial Types	Detailed Contents & evaluations
Tutorial 1	**Objective(s):** The objective of this activity is to offer a formal timeframe through which students can sit down in groups—on MS Teams—to go through the readings for the next planned class(es). This formal timeframe should help

	students better organize their reading schedule for the class. **Evaluations:** Some of these tutorials will require reading notes upon completion.
Tutorial 2	**Objective(s):** The objective of this activity is to offer a formal timeframe through which students can sit down in groups—on MS Teams or in person—in order to go through the Python codes for the previous class(es) and start working on the coding assignments. This formal timeframe should help students better organize their Python learning for the class. **Detailed Plan:** Students will be required to log in to the tutorial on MS Teams or be present physically to go through the codes for the last class(es), along with the upcoming coding assignment. **Teaching Method:** The students will be required to engage with the codes personally (either individually or in group) to absorb and get used to learning new coding material by themselves. The professor will be available for questions and any other help during this tutorial. **Evaluations:** Some of these tutorials will require coding notes upon completion.