DATA SCIENCE FOR CRIME SCIENTISTS (UCL SECU0050)

Department of Security and Crime Science, UCL

Bennett Kleinberg (bennett.kleinberg@ucl.ac.uk)

27/09/2019

Contents

Introduction . . . 1
Dates & times . . . 1
Contact & resources . . . 1
Learning outcomes . . . 2
Learning hours . . . 2
Structure . . . 2
Timetable . . . 2
Content . . . 3
Materials . . . 6
Assessment . . . 7
Attendance requirement . . . 9

THIS MODULE HANDBOOK IS STILL IN PROGRESS

Introduction

This UG (Year 3) module builds on the quantitative modules of Year 1 (Probability, Statistics and Modelling I) and Year 2 (Probability, Statistics and Modelling II), and introduces data science techniques as means for more sophisticated quantitative data analysis. This module aims to introduce students to computational methods for crime science.

The techniques covered in this module will be of relevance to students undertaking their final year independent research project as well as to those who wish to pursue a career in a data science or analyst role.

Dates & times

The module is running in Term 2, 2019/2020, from 13 January 2020 - 27 March 2020.

- Time and date: Thursdays, 9am-12pm.
- Each session includes a lecture and a tutorial (practical session)

UCL timetable page: https://timetable.ucl.ac.uk/tt/createCustomTimet.do#

Contact & resources

- Dr Bennett Kleinberg, Assistant Professor in Data Science, bennett.kleinberg@ucl.ac.uk
- TA: Daniel Hammocks, Doctoral researcher, daniel.hammocks@ucl.ac.uk

The **moodle page** will accompany this module here.

Q&A forum: if you have a questions/problem related to the content of the lectures/tutorials, then please use the Q&A forum.

Learning outcomes

Upon successful completion of this module, you will be able to:

- demonstrate knowledge of a broader range of analytical techniques used in the field of Security and Crime Science
- understand the purpose, advantages and disadvantages of different forms of data science techniques
- perform data science analyses on crime and/or-security related issues
- apply the data science pipeline on crime and/or-security related issues
- interpret and effectively report the results of said techniques

Learning hours

This module is worth 0.5 UCL credits (= 7.5 ECTS) which equals to 188 hours of study, i.e. 188h/11 weeks (incl reading week) = 17 hours per week.

Note: the content structure and the assessment assumes that you spend on average that amount of time with this module.

Component	Amount	Duration	Total hours
Lectures	10	2h	20h
Tutorials/practicals	10	1h	10h
Assessment: class test	1	1h	1h
Assessment: project	1	47h	47h
Homework/revision/self-study	11	10h	110h
TOTAL	-	-	188h

Structure

The general structure of this module is as follows: there are five content blocks (web data collection, text mining, machine learning, advanced techniques) which each will be covered in weekly 3h-sessions consisting of lectures and tutorials. The lectures cover the approaches on a conceptual (what do they do?) and functional level (how do they work?). The tutorials are practical sessions in which you will learn how to implement the techniques in the R programming language. During the tutorials, we will be there to assist you and help you.

Each week is (roughly) structured as follows:

- Reading and comprehending the required literature (necessary preparation for the lecture)
- Lecture (weekly content)
- Tutorial (practical implementation of lecture content)
- Homework (helps you consolidate the concepts and build your R skills portfolio)

Timetable

UCL week	Module week	Date	Topic
21	1	16 Jan	Web data collection I

UCL week	Module week	Date	Topic
22	2	23 Jan	Web data collection II
23	3	30 Jan	Text mining I
24	4	6 Feb	Text mining II
25	5	13 Feb	Text mining III
26	6	-	READING WEEK
27	7	27 Feb	Machine learning I
28	8	5 Mar	Machine learning II
29	9	12 Mar	Machine learning III
30	10	19 Mar	Case studies I
31	11	26 Mar	Case studies II
32	12	30 Mar	**Class test**

Content

Note: slides and tutorials will be added as the module progresses.

Week 1 - Web data collection I (16 Jan)

- Topics covered: web data collection with APIs
- Required preparation
  - Pfeffer, J., Mayer, K., & Morstatter, F. (2018). Tampering with Twitter's Sample API. EPJ Data Science, 7(1), 50. https://doi.org/10.1140/epjds/s13688-018-0178-0
  - Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. ArXiv:1306.5204 [Physics]. Retrieved from http://arxiv.org/abs/1306.5204
  - Soldner, F., Ho, J. C., Makhortykh, M., van der Vegt, I. W. J., Mozes, M., & Kleinberg, B. (2019). Uphill from here: Sentiment patterns in videos from left- and right-wing YouTube news channels. Proceedings of the Third Workshop on Natural Language Processing and Computational Social Science, 84–93. https://doi.org/10.18653/v1/W19-2110
  - *tuber* - Access YouTube API via R. https://soodoku.github.io/tuber/
  - Jeong, J. (2019). What Is API and How To Use Youtube API? https://towardsdatascience.com/what-is-api-and-how-to-use-youtube-api-65525744f520
  - YouTube. Obtaining authorization credentials
- Slides
- Tutorial
- Homework

Week 2 - Web data collection II (23 Jan)

- Topics covered: HTML basics, web scraping in R
- Required preparation
  - MDN HTML basics, MDN HTML tables
  - MDN Javascript basics
  - MDN CSS first steps
  - Ignatow & Mihalcea, 2018: C6 Web crawling
  - Rvest package documentation
- Slides
- Tutorial
- Homework
  - Rvest tutorial 1

  - Rvest tutorial 2
  - Rvest tutorial 3

Week 3 - Text mining I (30 Jan)

- Topics covered: quantification problem, TF-IDF, text metrics
- Required preparation
  - Zipf's Mystery (YouTube)
  - Ignatow & Mihalcea, 2018: C7 Lexical resources
  - Ignatow & Mihalcea, 2018: C8 Basic text properties
  - Grolemund & Wickham, 2016: C14 Strings
  - Quanteda quick start guide
  - Regular expressions with stringr
- Slides
- Tutorial
- Homework
  - Quanteda replication Jockers
  - Quanteda social media analysis

Week 4 - Text mining II (6 Feb)

- Topics covered: POS tagging, named entities, ngrams, sentiment analysis, sentiment trajectories
- Required preparation
- Slides
- Tutorial
- Homework

Week 5 - Text mining III (13 Feb)

- Topics covered: word embeddings, text similarity
- Required preparation
  - Introduction to word embeddings
  - Pennington et al., 2014
- Slides
- Tutorial
- Homework

Week 6 - Reading week

Use this week to catch up on any literature or homework/tutorials.

Week 7 - Machine learning I (27 Feb)

- Topics covered: supervised machine learning (classification + regression), core algorithms, performance metrics
- Required preparation
  - Ignatow & Mihalcea, 2018: C13 Text classification
  - Ignatow & Mihalcea, 2018: C9 Supervised learning

  - Gatto, 2019: Supervised learning
  - Deisenroth et al., 2019: C12 Classification with Support Vector Machines
  - Kuhn, 2019: C5 Model training
  - Kuhn, 2019: C17 Measuring performance
- Slides
- Tutorial
- Homework

Week 8 - Machine learning II (5 Mar)

- Topics covered: unsupervised machine learning, core algorithms, pitfalls
- Required preparation
  - Gatto, 2019: C4 Unsupervised learning
  - Datacamp unsupervised learning
- Slides
- Tutorial
- Homework

Week 9 - Machine learning III (12 Mar)

Guest lecture: Josh Kamps (Doctoral researcher)

- Topics covered: neural networs, deep learning
- Required preparation
- Slides
- Tutorial
- Homework

Week 10 - Case studies I (19 Mar)

This lecture includes a guest talk:

Guest lecture: Daniel Hammocks (Doctoral researcher) - "Facial recognition systems"

- Required preparation
  - Metropolitan Police (2019). Live Facial Recognition. https://www.met.police.uk/live-facial-recognition-trial/
  - Big Brother Watch (2018). The lawless growth of facial recognition in UK policing. https://bigbrotherwatch.org.uk/wp-content/uploads/2018/05/Face-Off-final-digital-1.pdf
  - Machine Learning From Scratch: Part 3 - Arrays and representations. https://towardsdatascience.com/machine-learning-from-scratch-part-3-ed572330367d
  - Biometrics and Forensics Ethics Group (2019). Ethical issues arising from the police use of live facial recognition technology. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/781745/Facial_Recognition_Briefing_BFEG_February_2019.pdf
  - Recommended: Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, 1, I–I. https://doi.org/10.1109/CVPR.2001.990517
- Slides
- Tutorial
- Homework

Week 11 - Case studies II (26 Mar)

This lecture includes two guest talks:

- Maximilian Mozes - "On the robustness of intelligent systems"
- Isabelle van der Vegt - "Data Science for Threat Assessment"
- Required preparation
  - Gu, T., Dolan-Gavitt, B., & Garg, S. (2019). BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain. ArXiv:1708.06733 [Cs]. Retrieved from http://arxiv.org/abs/1708.06733
  - Kurakin, A., Goodfellow, I., & Bengio, S. (2017). Adversarial examples in the physical world. ArXiv:1607.02533 [Cs, Stat]. Retrieved from http://arxiv.org/abs/1607.02533
  - Azucar, D., Marengo, D., & Settanni, M. (2018). Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis. Personality and Individual Differences, 124, 150–159. https://doi.org/10.1016/j.paid.2017.12.018
  - van der Vegt, I., Mozes, M., Gill, P., & Kleinberg, B. (2019). Online influence, offline violence: Linguistic responses to the "Unite the Right" rally. ArXiv:1908.11599 [Cs]. Retrieved from http://arxiv.org/abs/1908.11599
- Slides
- Tutorial: work on project.
- Homework

Materials

Software

We will use the R programming language. All packages, required resources and tools needed are openly available and free to download to any computer. We strongly encourage students to bring their own laptops to the tutorials so they can customise their work environment. However, we will have a computer cluster available where you can use the UCL computers.

Literature

You can find the required reading (as well as suggested further reading) in the table above and on the moodle page.)

For each week, the required literature includes one reading on techniques/tools that are important for data science projects but are not covered in the lectures or tutorials. You will likely need these techniques in the project and we assume that you have read and prepared the reading so that you are able to use the techniques.

Key resources for this module are:

- An Introduction to Text Mining (Ignatow & Mihalcea, 2018) - SAGE
- Text mining with R (Silge & Robinson, 2019) - interactive book version freely available at https://www.tidytextmining.com/
- Mathematics for Machine Learning (Deisenroth et al., 2019) - freely available at https://mml-book.github.io/book/mml-book.pdf
- R for Data Science (Grolemund & Wickham, 2016) - interactive book version freely available at https://r4ds.had.co.nz/

- The Caret package (Kuhn, 2019) - interactive book version freely available at https://topepo.github.io/caret/index.html
- Machine learning with R (Gatto, 2019) - interactive book version freely available at https://lgatto.github.io/IntroMachineLearningWithR/index.html
- Applied Predictive Modelling (Kuhn & Johnson, 2013) - this book is freely available to you as UCL student through UCL Library's free e-book access.

Other useful resources:

- Data Visualization with R (Kabacoff, 2018) - interactive book version freely available at https://rkabacoff.github.io/datavis/, especially:
  - Kabacoff, 2018: C2 Intro to ggplot2
  - Kabacoff, 2018: C3 Univariate graphs
  - Kabacoff, 2018: C4 Bivariate graphs
  - Kabacoff, 2018: C5 Multivariate graphs
- Quanteda text visualisation –>

Data

All datasets used are open-source and available without restrictions.

Assessment

Class test

- Weight for final grade: 30%
- Learning outcomes tested: (1) demonstrating knowledge of a broader range of analytical techniques used in the field of Security and Crime Science, (2) understanding the purpose, advantages and disadvantages of different forms of data science techniques, (3) interpreting the results of data science techniques.
- Date: 30 March 2020, 10am-12pm, IOE 20 Bedford Way, Room 305 – Clarke Hall.

Details: This 1-hour closed-book test covers theoretical, technical and conceptual aspects of the lectures, tutorials and the required literature. You will be given 10 multiple-choice questions. For each question, there will be five possible answer options. At least one option is correct and you are asked to select all options that apply (i.e. this could be one, two, three, four, or all five options). **If you select the correct - *and only the correct* - combination of answer options, you will be given ten points for a question.** In total, you can achieve 100 points in this test.

Applied Data Science Project

- Weight for final grade: 70%
- Learning outcomes tested: (1) demonstrating knowledge of a broader range of analytical techniques used in the field of Security and Crime Science, (2) performing data science analyses on crime and/or-security related issues, (3) applying the data science pipeline on crime and/or-security related issues, (4) interpreting and effectively reporting the results of said techniques
- Deadline: 16 April 2020, 4pm.
- Page limit: See below.

  This assessment is the capstone project of the module. It requires you to address a research problem in the full data science workflow (e.g., obtaining the data, processing the data, modelling the data, building predictive models, reporting on the findings, interpreting the outcomes). You will write a brief report on your project (a template will be provided) and you have to submit the R code needed to reproduce your findings. After passing this assessment, you will have the demonstrated the skills to solve a problem using data science techniques.

Feedback

A full project is a major step in your data science skills career. To help you in the process, you will receive feedback from us on a concept draft of your project. The deadline for the concept draft is **9 March 2020 (4pm)** via Turn-it-in on moodle. The requirements for the feedback submission are available on moodle. *The submission of the concept draft accounts for 10% of the Applied Data Science Project.*

Assessment topic

This year's topic is authorship attribution. Recently, an area of digital forensics closely connected to NLP that seeks to identify who wrote a piece of text/novel/song/etc. has received considerable attention (e.g., Why Molière most likely did write his plays, Researcher uses AI to unravel the mystery of Shakespeare's co-author, and Plecháč, 2019).

With more techniques at the researchers' disposal, this area is likely to impact on how we do digital forensic investigations (e.g. when trying to find who wrote a post). The project requires you to conduct your own authorship attribution project including (web) data collection, text processing and text mining, and approaches to identify authorship (e.g. through machine learning). *The exact nature of the project is up to you.* **The only restriction on the topic is that you must not use novels as the source of data**.

Further details:

- the project should be on authorship attribution (not author profiling)
- all steps in this project must be reproducible with your code supplement
- the project should have a large-scale focus (i.e. not a case study with one document)
- all three core areas of this module should be used: web data collection, text mining, machine learning

The code supplement

Submit your R code in the form of a commented R notebook. To ensure that no code is lost and that we can review all code equally, submit your code as an anonymised view-only version on the Open Science Framework. Create a private repository, upload your code as an R notebook and create an anonymised, view-only link that you include in your report (for a guide on creating that link, see here). For details on reporting your code as an R notebook, you can consult these guides: guide 1, guide 2.

The report

For this assignment, you are asked to report your findings in the form of a short paper. Specifically, you should use the template of the ACL conference proceedings (these can be downloaded here for Latex and Word or can be imported into Overleaf here).

Additional requirements for the report:

- use the ACL style guidelines (easiest through the templates)
- the **page limit is 5 content pages** + unlimited pages for the reference list
- use the ACL referencing style (this is available in reference managers like Zotero or handled directly in Overleaf) - i.e. adhere to their font type, font size and heading guidelines.
- use the anonymous submission version which contains line numbers
- the paper must contain only your examination number in the author line
- include a footnote with an anonymised view-only link to your code on the OSF (see above).
- submit the report via moodle as a pdf file using the following file name: *SECU0050_12345.pdf* (replace 12345 with your examination number)

Deliverables

- Concept draft (9 Mar 2020)

- Project report (16 April 2020)
- Code supplement (16 April 2020)

Grading criteria

Criterion	Meaning	Weight
Originality of project	The degree to which the student demonstrates insight and is able to use an innovative approach to address the question of authorship attribution.	20%
Quality of data science techniques	The degree to which the techniques used in this project are appropriate to answer the research question and are utilised and interpreted properly.	40%
Quality of the R code	The degree to which the R code is well-documented, reproducible (with provided data if needed), and correct.	20%
Quality of the report	The clarity, layout, formatting and overall quality of the written report.	10%
Concept draft submission	Whether or not the concept draft was submitted in the required format.	10%

Attendance requirement

We are obliged to record the attendance at all sessions (lectures and tutorials) and each student will have to attend at least 70% of the sessions to be able to pass the module. If you cannot attend for a good reason, please let the TA know about this well in advance. *We strongly advise you to attend all sessions as this eases the assessment for you and will help you get the most out of this module.*