APPLIED DATA SCIENCE (UCL SECU0057)

Department of Security and Crime Science, UCL

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

26/03/2020

Contents

Introduction

This MSc module introduces students to the field of data science and provides them with the conceptual understanding of the underlying principles. The students will be able to address crime and security problems in a computational manner (e.g., using data-driven crime analysis) and will become critical consumers of data science approaches. It will also allow students to understand the approaches which must be taken to translate these theoretical concepts into real-world applications used by the police and security services, as well in academia to influence and drive government policy.

The techniques covered in this module will be of relevance to students undertaking a data science-related 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, 2-4pm.
- 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: Felix Soldner, Doctoral researcher, felix.soldner@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:

- develop a holistic understanding of what Data Science encompasses and its pervasive influence in modern life, including its impact within law enforcement, crime analysis and security science
- understand of key mathematical, computational, and statistical analysis techniques employed in the field of security to determine trends and patterns in measured crime data
- retrieve data from unstructured sources through web-scraping techniques
- write the code needed for the analytical steps in R
- appreciate the fundamentals of supervised and unsupervised machine learning
- understand the methodologies used to extract features from data.
- use techniques used in natural language processing to gain insights from text data
- discuss the ethical issues concerned with deploying latest advances in AI to areas such as criminal. justice, profiling
- understand the role that open data and the open science movement is playing in society, from academic research and the private sector, to shaping government policy

Learning hours

This module is worth 15 UCL credits (= 7.5 ECTS) which equals to 150 hours of study, i.e. 150h/11 weeks (incl reading week) = 14 hours per week.

Note that 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	1h	10h
Tutorials/practicals	10	1h	10h
Assessment: class test	1	1.5h	1.5h
Assessment: project	1	40.5h	40.5h
Homework/revision/self-study	11	8h	88h
TOTAL	-	-	150

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 2h-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
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
31	11	26 Mar	**Class test**

Content

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

Week 0 - Module preparation and software setup

Please ensure that you follow the guides below to setup your computer with the necessary software. We also assume that you have a basic understanding of pragmatically solving coding problems. Do do this, please take the time to walk through the tutorial below.

- Tutorial guide: R for crime scientists in 12 steps
- Tutorial guide: Getting ready for R
- Tutorial guide: How to solve data science problems

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
  - MDN HTML basics, MDN HTML tables
- Slides: Week 1: Web data collection 1, PDF
- Tutorial + Homework: Tutorial, week 1: web data collection, Tutorial solutions

Week 2 - Web data collection II (23 Jan)

- Topics covered: HTML basics, web scraping in R
- Required preparation
  - MDN Javascript basics
  - MDN CSS first steps
  - Ignatow & Mihalcea, 2018: C6 Web crawling
  - Rvest package documentation
- Slides: Week 2: Web data collection 2, PDF
- Tutorial + Homework: Tutorial, week 2: web data collection 2, Tutorial solutions

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
  - Schoonvelde, M., Brosius, A., Schumacher, G., & Bakker, B. N. (2019). Liberals lecture, conservatives communicate: Analyzing complexity and ideology in 381,609 political speeches. PLOS ONE, 14(2), e0208450. https://doi.org/10.1371/journal.pone.0208450
- Recommended:
  - Regular expressions with stringr
- Slides: Week 3: Text Mining 1, PDF
- Tutorial + Homework: Tutorial, week 3: Text Mining 1, Tutorial solutions

Week 4 - Text mining II (6 Feb)

- Topics covered: POS tagging, ngrams, sentiment analysis, sentiment trajectories
- Required preparation:
  - 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
  - Kleinberg, B., Mozes, M., & van der Vegt, I. (2018). Identifying the sentiment styles of YouTube's vloggers. 3581–3590. https://www.aclweb.org/anthology/D18-1394
  - Gao, J., Jockers, M. L., Laudun, J., & Tangherlini, T. (2016). A multiscale theory for the dynamical evolution of sentiment in novels. Behavioral, Economic and Socio-Cultural Computing (BESC), 2016 International Conference On, 1–4.
  - Jockers, M. (2015). Revealing Sentiment and Plot Arcs with the Syuzhet Package. http://www.matthewjockers.net/2015/02/02/syuzhet/
- Slides: Week 4: Text mining 2, PDF
- Tutorial+ Homework: Tutorial: week 4: Text mining 2, Tutorial solutions

Week 5 - Text mining III (13 Feb)

- Topics covered: text similarity, word embeddings
- Required preparation
  - Introduction to word embeddings, Mozes (2019)
  - Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1532–1543. https://doi.org/10.3115/v1/D14-1162 from https://nlp.stanford.edu/pubs/glove.pdf
  - Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. ArXiv:1310.4546 [Cs, Stat]. http://arxiv.org/abs/1310.4546
  - Vector dot products from https://www.khanacademy.org/math/linear-algebra/vectors-and-spaces/dot-cross-products/v/vector-dot-product-and-vector-length
  - Matrix vector products from https://www.khanacademy.org/math/linear-algebra/vectors-and-spaces/null-column-space/v/matrix-vector-products
  - Limitations of word embeddings, Burdick, 2019

  - Wendlandt, L., Kummerfeld, J. K., & Mihalcea, R. (2018). Factors Influencing the Surprising Instability of Word Embeddings. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2092–2102. https://doi.org/10.18653/v1/N18-1190
- Recommended:
  - Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. Proceedings of the National Academy of Sciences, 115(16), E3635–E3644. https://doi.org/10.1073/pnas.1720347115
  - Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. ArXiv:1607.06520 [Cs, Stat]. http://arxiv.org/abs/1607.06520
  - Vylomova, E., Rimell, L., Cohn, T., & Baldwin, T. (2016). Take and Took, Gaggle and Goose, Book and Read: Evaluating the Utility of Vector Differences for Lexical Relation Learning. ArXiv:1509.01692 [Cs]. http://arxiv.org/abs/1509.01692
- Slides: Week 5: Text mining 3, PDF
- Tutorial + Homework: Tutorial, week 5: Text mining 3, Tutorial solutions

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
- Recommended: YT, 3Blue1Brown: Bayes theorem, and making probability intuitive
- Slides: Week 7: Machine Learning 1, PDF
- Tutorial + Homework: Tutorial, week 7: Machine Learning 1, Tutorial solutions

Week 8 - Machine learning II (5 Mar)

- Topics covered: unsupervised machine learning, core algorithm, problems
- Required preparation
  - Gatto, 2019: C4 Unsupervised learning
  - Denny, M. J., & Spirling, A. (2018). Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It. Political Analysis, 26(2), 168–189. https://doi.org/10.1017/pan.2017.44
  - Datacamp unsupervised learning
- Slides: Week 8: Machine Learning 2, PDF
- Tutorial + Homework: Tutorial, week 8: Machine Learning 2, Tutorial solutions

Week 9 - Machine learning III (12 Mar)

Guest lecture: Josh Kamps (Doctoral researcher)

- Topics covered: neural networks in R, intro to deep learning
- Required preparation:
  - Domingos, P. (2012). A few useful things to know about machine learning. Communications of the ACM, 55(10), 78. https://doi.org/10.1145/2347736.2347755
  - Nielsen, M. A. (2015). Neural Networks and Deep Learning. http://neuralnetworksanddeeplearning.com
- Slides: Week 9 : Machine Learning 3
- Tutorial + homework: Tutoriall, week 9: Machine Learning 3

Week 10 - Case studies (19 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: *video-recorded lectures due to the Corona virus situation*
- Tutorial + Homework: *You can spend this tutorial working on your project under our guidance.*

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
  - *UCL library services have ordered a few copies of this book*
  - You can buy the book at a UCL student discount via this link https://studysites.uk.sagepub.com/dts-studentdeal/
- 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

Coursework

- 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.
- Deadline: 30 April 2020, 4pm

You receive 4 questions that you have to answer using methods and techniques covered in this module.

Grading criteria

Criterion	Meaning	Weight
Quality of the answer to Q1	The quality of the solution provided to question 1 (incl. R code).	25%
Quality of the answer to Q2	The quality of the solution provided to question 2 (incl. R code).	25%
Quality of the answer to Q3	The quality of the solution provided to question 3 (incl. R code).	25%

Criterion	Meaning	Weight
Quality of the answer to Q4	The quality of the solution provided to question 4 (incl. R code).	25%

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: 30 April 2020, 4pm.
- Page limit: See below.
- Assessment details as PDF.

  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.*

- Feedback form template as PDF.

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 restrictions on the topic and scope are: you must not use novels or Twitter as the source of data, and you must have at least 500 data points (e.g. texts/paragraphs/etc.) for each of at least 10 authors.**.

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 8 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: *SECU0057_12345.pdf* (replace 12345 with your examination number)

Deliverables

- Concept draft (9 Mar 2020)
- Project report (30 April 2020, 4pm)
- Code supplement (30 April 2020, 4pm)

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.	15%
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.	15%
Quality of the R code	The degree to which the R code is well-documented, reproducible (with provided data if needed), and correct.	20%
Report "Introduction" section	The quality of the review of related works and the logical flow of the argument in the introduction section.	10%
Report "Method" section	The clarity of the method section that details the steps of data acquisition, preprocessing and analysis.	10%
Report "Results" section	The suitability and clarity of the results section.	10%
Report "Discussion" section	The quality of the overall interpretation as well as limitations and suggestions for future work.	15%
Concept draft submission	Whether or not the concept draft was submitted in the required format.	5%

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.*
