Data Science on Unstructured Text Data

Instructor: Eduardo Arino de la Rubia (earino@gmail.com)
Credits: 1 (3x200min)
Dates: Feb 13, Feb 16 and Feb 19, starting at 17.30-21.00
Term: Winter 2018
Course level: [MA/MSc]

Course description

Text is ubiquitous. Humans have been storing information in written form for over 5000 years, and unfortunately the information in this information has defied principled quantitative analysis for much of that time. Unlocking skills and techniques to take text and derive sense and sentiment enables exploratory analysis and modeling on human communication. In this course we will use R packages such as ggraph, tidytext, dplyr and topicmodels to manipulate and understand a number of document collections. We will also learn to import textual data into R using twitteR, docxtractr and other packages.

Learning outcomes

By successfully completing the course the students will be able to:

- Learn how to transform text into the tidytext format for NLP
- Extract emotion and tone from text using sentiment analysis
- Understand what makes a document unique in a collection
- Understand how words and tokens and visualize them
- Import and export textual data into R
- Classify documents into groups using topic modeling
- Build models which take as input textual features

Reading list

Text Mining with R - ISBN 978-1-491-98165-9 by Julia Silge & David Robinson

Assessment

3 Daily Quizes (30%)
End of Course Assignment (60%)
Intellectual Presence (10%)

Note on Intellectual Presence

To be counted as intellectually present, you must demonstrate an intellectual presence, which means you are engaged in all classroom activities. An intellectual absence (including reading non-course related material, playing/texting on phone, using a laptop for non-class related activities) will be counted as an absence. Students who anticipate the need to be absent should be aware that this course is very compressed, and any absence will make it very challenging to complete this course.

Contacting Me

Email is a great way to contact me, it is basically a 24-hour link to my brain. However it is important that you communicate very clearly. All emails to me should have a subject which identifies that your email is regarding this course. The body of the email should have a clear point, and if you need a response, it should be very clearly stated. Clear communication is a virtue.

About The Instructor

Eduardo Arino de la Rubia is a data science leader at Facebook. He is a lifelong technologist with a passion for data science who thrives on effectively communicating data-driven insights throughout an organization. He was previously the Chief Data Scientist at Domino Data Lab, a company he joined as an advisor pre-seed, and took it through a successful series B led by Sequoia Capital. He spent 10 years at Ingram as the Principal Data Scientist. He is a graduate of the MTSU Computer Science department, and has a Masters in Negotiation, Conflict Resolution, and Peacebuilding from CSUDH.