Introduction to Machine Learning in Digital Humanities

Digital Humanities Summer Institute
University of Victoria
June 12-16, 2017

Instructors: Paul Barrett (paul.barrett@utoronto.ca) and Nathan Taback (nathan.taback@utoronto.ca)

The learning objective of this course is to become familiar with machine learning (ML) techniques used in the digital humanities (DH).

By the end of the course students will be able to:

- Describe how ML algorithms are used in the digital humanities.
- Describe and use standard tools in supervised ML including linear and logistic regression and trees.
- Critique specific applications in DH that make use of supervised ML techniques.
- Describe and use standard tools in unsupervised ML including cluster analysis.
- Critique specific applications in DH that make use of unsupervised ML techniques.
- Describe and use basic topic models to discover hidden topics/themes in documents.

Software for Machine Learning

We will be using R, an open source statistical language, to implement the ML algorithms that will be discussed in class. R Studio is an integrated development environment for R. Students should install R and R Studio on their laptops that they will bring to class.

Step #1: Download and install R from https://cran.rstudio.com

Step #2: Download and install R Studio Desktop (Free version) from https://www.rstudio.com/products/rstudio/download/

Daily Schedule

Morning – 9:00 -12:00 (except 10:15 on day 1), afternoon – 1:30 – 4:00.

Students should bring a laptop to class with R installed.

Afternoon case studies will involve small group work on a case study and class presentations.

HG WELLS / BRONTE SISTERS EXAMPLE
JOCKERS examples: need one supervised and one unsupervised

Day 1 (Morning)

- Introduction to machine learning
- Introduction to digital humanities
- Introduction to GitHub
- Introduction to R using R Studio (students should install R and R Studio before the first class)
- Nathan and Paul introduce concepts in ML and DH

Day 1 (Afternoon)

Bob Dylan: "Gotta Serve Somebody" – Data Visualization
-Raise questions to the class
-Get them to work on a new text. Create a new notebook

Day 2 (Morning)

- Introduction to supervised machine learning techniques used in DH.
- Sentiment analysis on twitter using tidytext
- Interact with class to show the change from ulimited to unlimited & get them to score the tweets themselves
- #DHSI2017 – is the conference going well?
- tf-idf / n grams
- Different lexicons

Day 2 (Afternoon)

Supervised ML using R: logistic regression
Students do Shakespeare and Jonson supervised learning example
Plan B if Shakespeare doesn't work: Email version with modified code

Day 3 (Morning)

- Introduction to unsupervised machine learning techniques used in DH.
- Clustering: HG Wells / Austin Clarke
- Topic Modeling

Day 3 (Afternoon)

- Example of topic modeling – what to use? CanLit?
- Take out French language stuff?
- Editorial decisions – how do they affect this?

Day 4 (Morning)

- Student teams work on data set with questions that we provide.
- A few options: Shakespeare, CanLit, Reuters, Jane Austen, Twitter: Choose a hash tag and figure out what the topics are.
- OR we carry over what we haven't gotten into

Day 4 (Afternoon)

More work time

Day 5 (Morning)

- Putting it all together.