Text Analysis with NLTK

March 30-April 8, 2022

Week 1

Centre for Data, Culture & Society

Instructor: Lucy Havens

Course Structure

Anticipate about ~7 hours/week

- 2 course meetings per week
  - 10 - 11 AM Wednesdays
  - 10 - 11 AM Fridays
- 1 assignment per week ~2 hours
- Office hours upon request
- Independent learning ~2 hours

Teams for introductions, meetings, office hours, questions, files

Course Topics

Text analysis - analyzing unstructured data

Python

Regular Expressions

Natural Language Toolkit (NLTK)

INTRODUCTIONS!

Why are you interested in text analysis?

Have you used Python before?

Have you used Jupyter Notebooks before?

Have you used Regular Expressions before?

Have you used NLTK before?

Participant Expectations

Wednesday and Friday classes are introductions to material

Assignments will be given on Wednesday

Classes are not recorded but all class materials will be uploaded to Teams

Please let me know in advance if you cannot attend!

Message me on Teams to schedule office hours for questions

Course Software

Jupyter Notebooks

- With Google Colab

  https://colab.research.google.com

- With Notable - “Language and Machine Learning” Notebook

  https://www.ed.ac.uk/information-services/learning-technology/noteable/accessing-noteable

  After logging into MyEd: https://noteable.edina.ac.uk/launch

- Locally (install with pip/pip3 or conda)

DEMO!

Further Resources

Noteable User Guide
https://noteable.edina.ac.uk/user_guide/#hide_ge_7

Jupyter Notebooks in Noteable
https://github.com/edina/Exemplars2020/blob/master/TeachingDocs/Tutorials/UsingNoteableBeginner.ipynb

Jupyter Notebooks
https://glam-workbench.github.io/getting-started/

Python
https://programminghistorian.org/en/lessons/introduction-and-installation

Note: these and more are provided in NLTK_Assignment1.pdf

NLTK

Natural Language Toolkit

Natural language = human language = “unstructured” data

NLTK

Examples of data sources for natural language:

- Books
- Newspapers
- Magazines
- Websites
- Transcriptions of audio (i.e. interview, movie dialogue)
- Social media

Always read the licensing/copyright information and terms of use!

Why use NLTK?

What kinds of questions can you ask when you can use a programming language to study hundreds, thousands, or even millions of pages of digital text?

"Distant reading"

Why NLTK isn't everything

What kinds of questions can you ask when you can physically hold and look at a printed text, be it an original publication or later edition of the text?

"Close reading"
Book history

NLTK Terminology

Tokens vs. words

Digitized vs. digital

Normalization (a.k.a. standardization)

Document vs. corpus vs. corpora

Getting Familiar with a Text

Built-in methods include:

```
.concordance("word", lines=20)
.similar("word")
.common_contexts(["list", "of", "words"])
.dispersion_plot(["list", "of", "words"])
```

Reference: https://www.nltk.org/book/ch01.html

DEMO!

Summarizing a Text

Built-in functions and methods include:

```
len(text)
sorted(vocabulary_of_text)
.count(word
```

Reference: https://www.nltk.org/book/ch01.html

DEMO!

Assignment

Prework

If you have not used/need a refresher on Python, Jupyter Notebooks

Steps 1-5: Independent learning

Step 6: A tutorial to complete in your own Jupyter Notebook

Go Further & Helpful Resources - optional, for now or future reference

See Files > Class Materials > NLTK_Assignment1.pdf

Thanks everyone!

Next class: Friday, 10-11 AM

Please message me on Teams for office hours!