Text Analysis with NLTK

Helpful Resources

- Pythex (for practice making regular expressions in Python): https://pythex.org
- Using Markdown cells in Jupyter Notebooks: https://medium.com/analytics-vidhya/the-ultimate-markdown-guide-for-jupyter-notebook-d5e5abf728fd
- Git documentation: https://git-scm.com/docs

Week 2: The Natural Language Toolkit (NLTK)

Assignment

1. **Find a dataset of text that you would like to analyze with NLTK.**
   If you're not sure where to start, here are a few websites where you may find text data you'd like to analyze:

   - Historical publications from the National Library of Scotland: https://data.nls.uk/data/digitised-collections/
     Note: Some of the datasets can be quite large, so you can always download the sample data instead of the full text if you're interested in one of the larger datasets but don't want to deal with every single file it contains!
   - A book or speech provided in the NLTK library (see the section using " from nltk.book import * "): https://www.nltk.org/book/ch01.html
   - Project Gutenberg – you could download the "Plain Text UTF-8" format of one or more books from this website, such as: https://www.gutenberg.org/ebooks/1279
   - Save any Word document (or several Word documents!) you've created as a Plain Text file (ending in ".txt" instead of ".doc")

2. **Read "An Introduction to the Data Biography" and complete one for your dataset.**
   The Introduction: https://weallcount.com/2019/01/21/an-introduction-to-the-data-biography/
   The data biography tool: https://wac-survey-rails.herokuapp.com

3. **Write down 3 questions you'd like to ask about the text you chose in #1.**

4. **For 1 or more of the questions you wrote in #2, write down the tools NLTK provides that you'd need to use to try to answer that question, and number those tools in the order that you'd apply them.** Consider pre-processing activities in addition to analysis activities!

5. **In a Jupyter Notebook, try to answer one of your questions from #3 using NLTK, following the steps you outlined in #4**

Go Further

- Try another text analysis tool, **AntConc**: https://programminghistorian.org/en/lessons/corpus-analysis-with-antconc
- Take CodeAcademy's **"Learn Git"** course: https://www.codecademy.com/learn/learn-git
- Build a chatbot with another text analysis tool, **SpaCy**: https://www.codecademy.com/learn/paths/build-chatbots-with-python
- Try the **Edinburgh Geoparser**: https://programminghistorian.org/en/lessons/geoparsing-text-with-edinburgh