Intro to Text Analysis with Python

November 3, 2021

Brooke Luetgert

THE UNIVERSITY OF CHICAGO
Office of Research and National Laboratories Research Computing Center

Contact Information

Brooke Luetgert, PhD. Senior Computational Scientist at Research and Computing Center (RCC)

Email: **luetgert@uchicago.edu**
Office: TAAC 2, 5607 South Drexel
Telephone: (773)-834-5313

RCC Help Desk: Reg. 216, Mon-Fri 9 AM-5 PM

Materials on GitHub- use search bar, enter **user:luetgert**
Our folder is **luetgert/TextAnalysis**

Before we begin

1) Please download and install Anaconda (anaconda.com/products/individual)
2) Please navigate to Github for the class materials: (github.com/luetgert/TextAnalysis) You will click on the green code button to download all materials as a zip file. Unzip the file and place it on your desktop.
3) You will now create a virtual environment using the .yml file provided (pre-install all libraries). We will activate this environment and launch our Jupyter Notebook from within this environment.

Resources

Installing a virtual environment with Anaconda Jupyter Notebooks (Mac and PC examples):

https://medium.com/swlh/installing-jupyter-notebook-and-using-your-own-environment-on-mac-fa41efd4639d (look specifically at part 2)

https://www.geeksforgeeks.org/using-jupyter-notebook-in-virtual-environment/

1) At terminal $ conda env create --file pathToFile --name textanalysis (note that pathToFile will be replaced with where the .yml is located) The creation process takes time. When the dollar prompt returns, the process is complete and the environment can be activated. ONLY DO THIS ONCE.

2) In order to activate the environment, at terminal please type $ conda activate textanalysis (note that you will see the environment active before the $ prompt)

3) From here, we will load a Jupyter notebook in our environment. At terminal please type $ jupyter notebook pathToFile (note that pathToFile will be replaced with path to Jupyter notebook)

Proper Housekeeping

When we use virtual environments, we must remember to exit out of Jupyter correctly. We will exit from our session. In terminal, we will (1) use ctrl+c to close the Jupyter notebook and (2) deactivate the conda environment to revert to the base environment before we close the terminal window.

```
$ conda deactivate
```

Text as Data

Textual analysis is a methodology that **involves using language (here understood as written English) to gain information regarding how people make sense of and communicate life and life experiences**. We will use written text passages, available data from online sources and data that we scrape and assemble to provide cues on communication and the content/ sentiment conveyed in that source.

Things to remember

We want to write Python programs that can analyze vast amounts of textual data, but:

- Information contained in text has a richness and depth unlike numerical data. Quantification of qualitative data is hard and imperfect.
- Computers struggle with text. Grammar rules, contextual meaning, and negation. (Think of a word such as “like”. Think of statements such as “A robin is a _____” vs “A robin is not a _____” (bird vs. tree)
- Advances in machine learning algorithms are helping us to extract ever more nuance- and therefore value- in textual data.

Goals

We want to achieve Natural Language Understanding, but we must start with Natural Language Processing (NLP). Our goals include:

- topic classification (automatically tagging texts by topic),
- feature extraction (identifying specific characteristics in a text), and
- sentiment analysis (recognizing the emotions that underlie a given text).

Plan for this series

We will explore several libraries within Python for computer assisted data collection, text analysis and content analysis.

1. **Collect and prepare** unstructured data and transform it into information we can analysis (Tokenization, parsing and stopwords)- **November 3**
2. **Text Classification**- rule-based, machine learning alternatives- **November 10**
3. **Evaluation**- How can we measure accuracy, precision and recall in our results? **November 17**