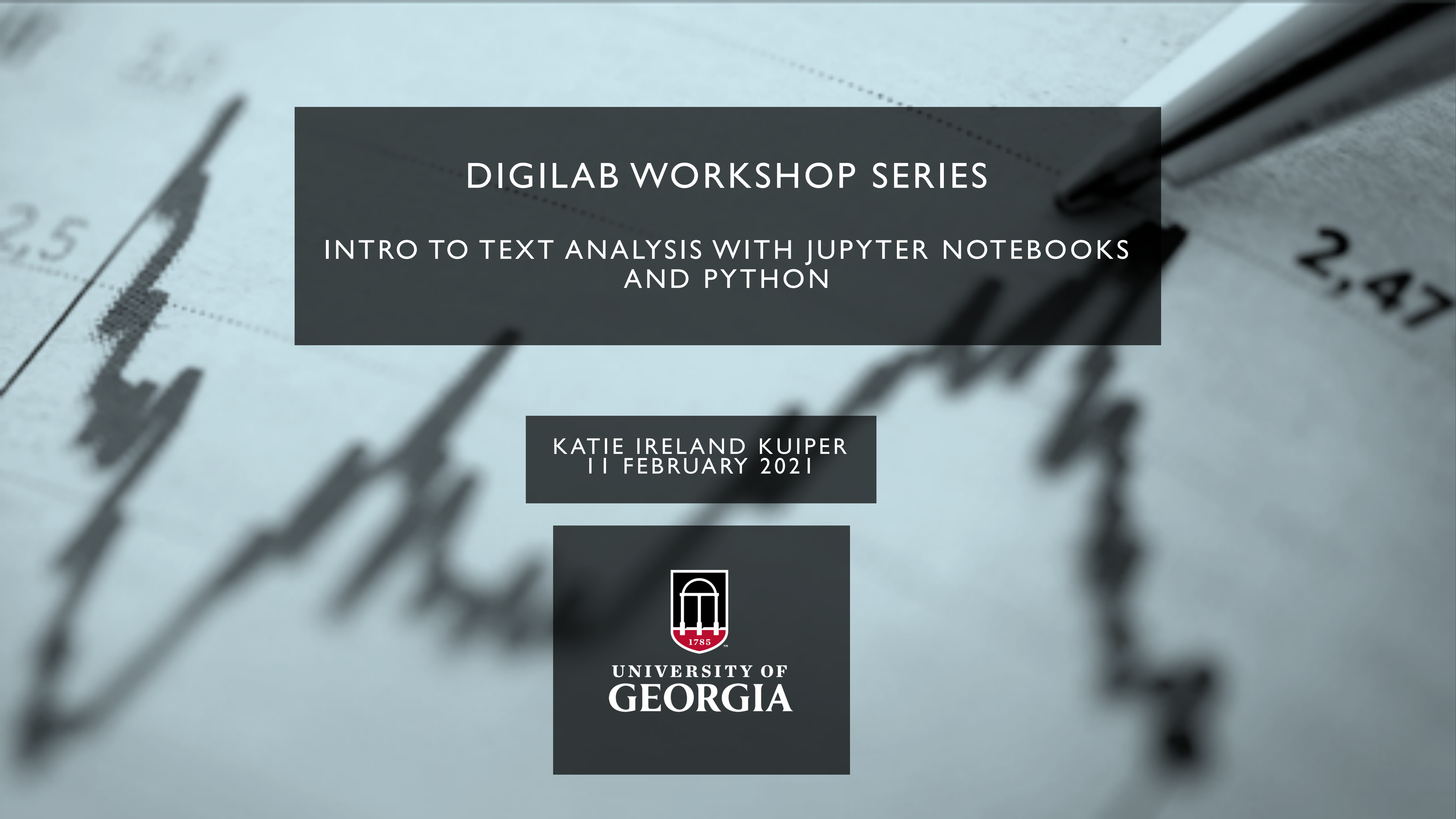


DIGILAB WORKSHOP SERIES

INTRO TO TEXT ANALYSIS WITH JUPYTER NOTEBOOKS AND PYTHON

KATIE IRELAND KUIPER
11 FEBRUARY 2021

1785
UNIVERSITY OF
GEORGIA

PYTHON

- Extremely useful programming language; it is highly intuitive and easy to use once you have some experience with it.
- Includes many built-in functions and useful libraries for text analysis and processing.

- Download and install:
- Anaconda

LIBRARIES

- NLTK
- pandas
- matplotlib
- gensim

Photo by Thomas Millot on Unsplash

Data types in python

- **Strings** are a sequence of characters; Strings are immutable.
- Textual data in Python is handled with *strings*. Strings are immutable sequences of Unicode code points. String literals are written in a variety of ways:
  - Single quotes:
  - 'allows embedded "double" quotes'
  - Double quotes:
  - "allows embedded 'single' quotes"
  - Triple quotes:
    - '''Three single quotes''', """Three double quotes"""
- **Lists:** Lists are mutable sequences, typically used to store collections; they allow you to store information.

I AM THE TOKENIZER
I'VE COME TO TEAR YOUR STRINGS APART
generator.net

METHODS

String methods, uploading your own corpus

topic modeling

sentiment analysis

Using NLTK corpora

PYTHON (WITH JUPYTER NOTEBOOKS)

- **spaCy**: pos tagging, tokenization, dependency parsing, etc. Check out this tutorial for more about NLP with spaCy
- **CoreNLP**: lemmatization, pos tagging, tokenization, named entity recognition
- **NLTK**: Natural Language ToolKit; contains over 100 corpora, includes options for tokenization, tagging, parsing, document classification
- **Gensim:** useful for various types of topic modeling
- **PyNLPI:** open-source NLP library; great for of tasks ranging from building simplistic models and extraction of n-grams and frequency lists, with support for complex data types and algorithms
- **Pattern**: useful for web-crawling (webscraping) for creating your own corpora; includes options for tokenizing, pos tagging, etc
- **Polyglot**: very useful library for other languages than English
- **TextBlob:** includes options for pos-tagging, noun phrase extraction, classification, translation and sentiment analysis

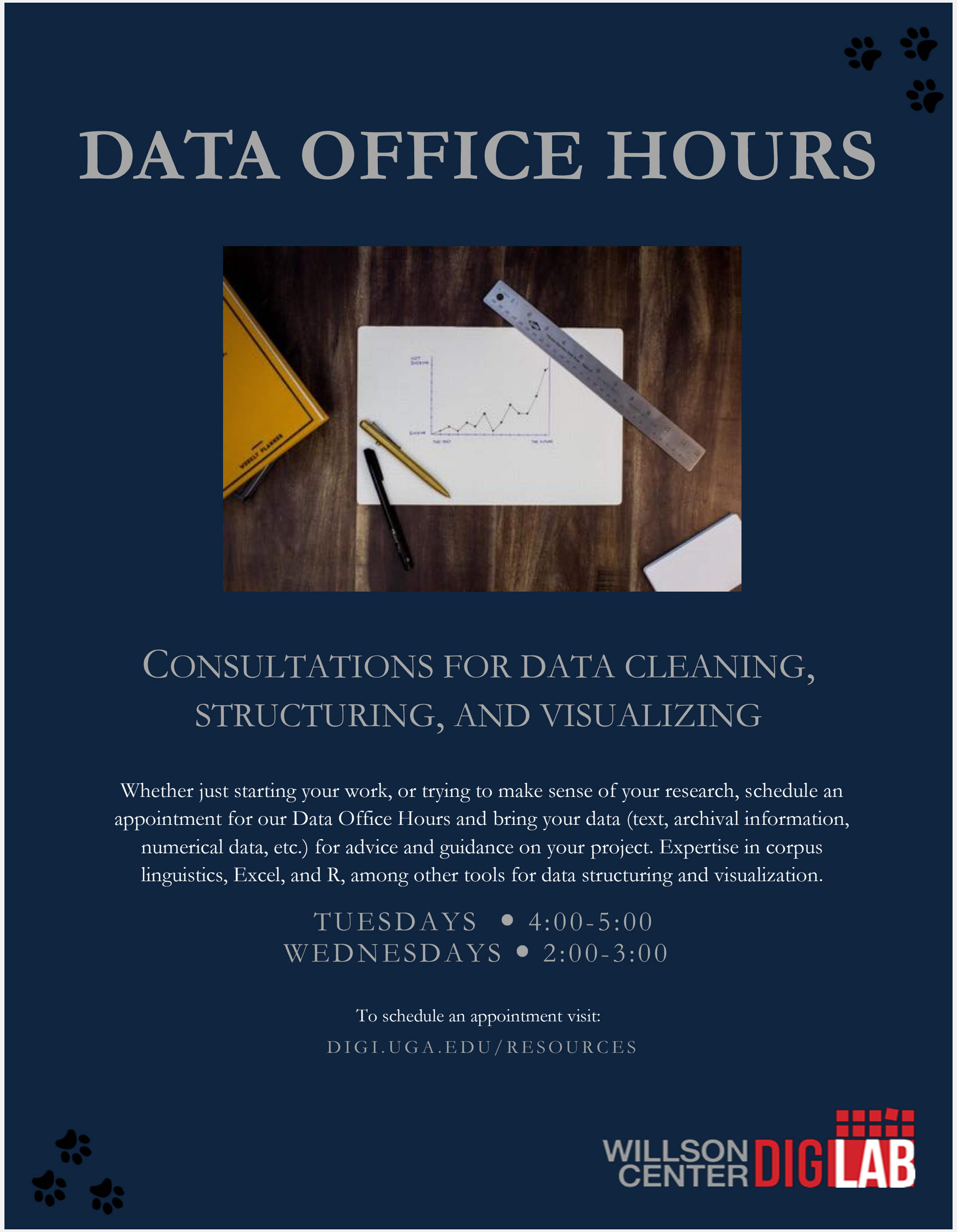
RESOURCES AT UGA

- Corpus Server
- Upcoming Courses
- Digilab Resources
- Data Office Hours

COURSES AT UGA

- This Fall 2021:
- Natural Language Processing: LING 4570/6570
- Style: ENGL/LING 4826/6826
- American English: ENGL/LING 4010/6010
- Note: These all count toward the Digital Humanities Undergraduate certificate!

RECOMMENDED RESOURCES



DATA OFFICE HOURS

CONSULTATIONS FOR DATA CLEANING, STRUCTURING, AND VISUALIZING

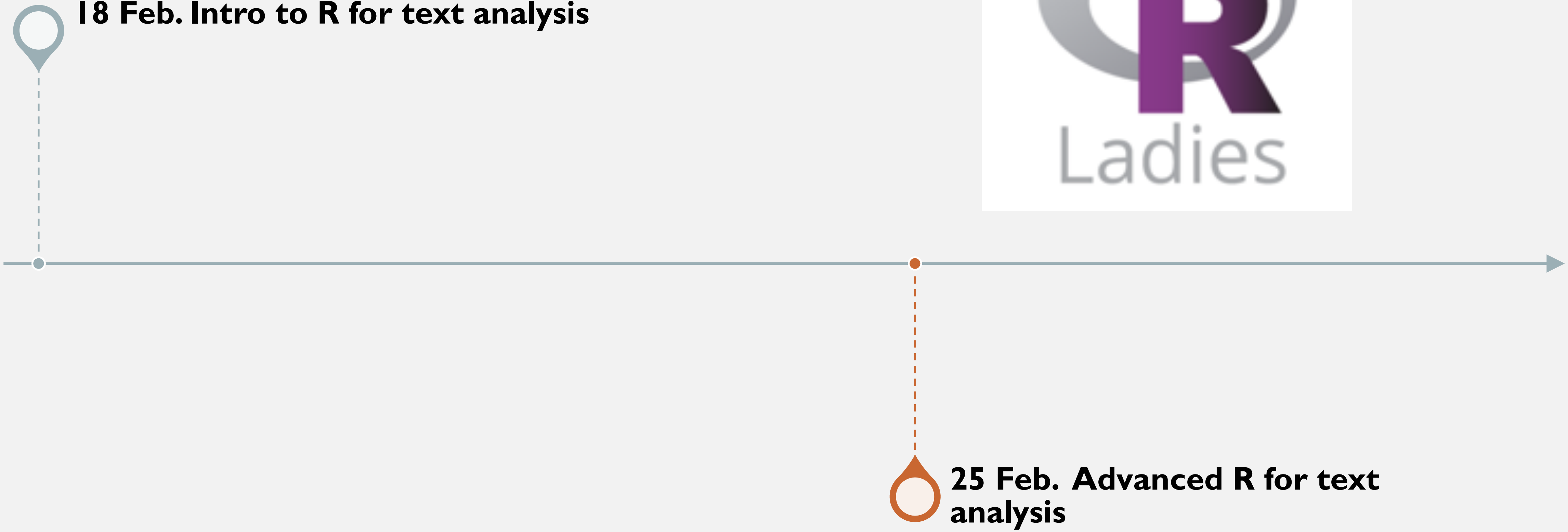
Whether just starting your work, or trying to make sense of your research, schedule an appointment for our Data Office Hours and bring your data (text, archival information, numerical data, etc.) for advice and guidance on your project. Expertise in corpus linguistics, Excel, and R, among other tools for data structuring and visualization.

TUESDAYS • 4:00-5:00
WEDNESDAYS • 2:00-3:00

To schedule an appointment visit:
DIGI.UGA.EDU/RESOURCES

- Natural Language Processing with Python by Bird et al.; Na-Rae Han's python tutorials
- Take NLP this fall!! Natural Language Processing: LING 4570/6570
- Data office hours!

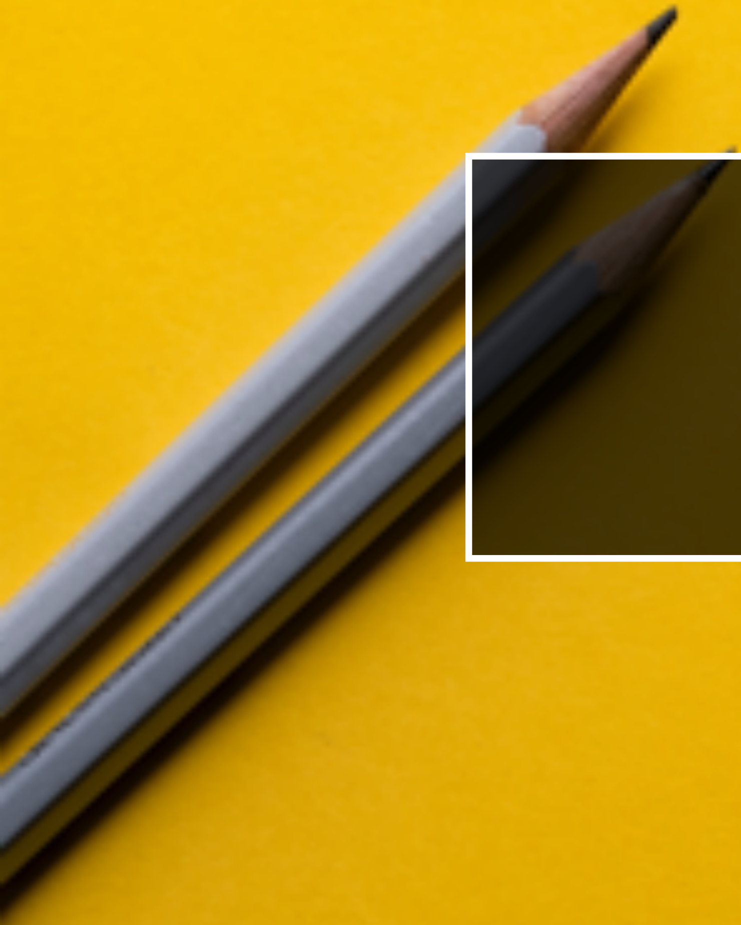
COMING UP NEXT…



18 Feb. Intro to R for text analysis

25 Feb. Advanced R for text analysis

INSTALL R AND R STUDIO FOR NEXT WEEK!



THANKS FOR LISTENING!

KATHERINE.KUIPER25@UGA.EDU

WORKS CITED

- Bansal, Shivam. 2016. Beginners Guide to Topic Modeling in Python. https://www.analyticsvidhya.com/blog/2016/08/beginners-guide-to-topic-modeling-in-python/
- Bird, Steven, Ewan Klein, and Edward Loper. 2019. Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit
- Brown, Simon. 2016. Tips for Computational Text Analysis. https://matrix.berkeley.edu/research/tips-computational-text-analysis
- Chalaguine, Lisa. 2020. Getting started with text analysis in Python. https://towardsdatascience.com/getting-started-with-text-analysis-in-python-ca13590eb4f7
- Gensim: 3.8.3 documentation. 2020. https://pypi.org/project/gensim/
- Evert, Stefan and Andrew Hardie. 2011. Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics 2011 conference,* University of Birmingham, UK.
- Evert, Stefan. 2003. The CQP Query Language Tutorial.
- Han, Na-Rae. Python 3 tutorials. http://www.pitt.edu/~naraehan/python3/.
- HathiTrust. https://www.hathitrust.org/about.
- Laudun, John. Text Analytics 101. https://johnlaudun.org/20130221-text-analytics-101/
- Loria, Steven. 2020. TextBlob: Simplified Text Processing. https://textblob.readthedocs.io/en/dev/

- https://monkeylearn.com/text-analysis/
- Matthes, Eric. 2016. *Python Crash Course: A Hands-on, Project-based introduction to programming.*
- Malik, Usman. 2021. Removing Stop Words from Strings in Python. https://stackabuse.com/removing-stop-words-from-strings-in-python/
- Munir, Samira. 2019. Basic Sentiment Analysis using NLTK. https://towardsdatascience.com/basic-binary-sentiment-analysis-using-nltk-c94ba17ae386
- Millot, Thomas. Photo. Unsplash
- NLTK 3.5 documentation. https://www.nltk.org
- Pandas documentation. 2021. V. 1.2.2. https://pandas.pydata.org/docs/
- Prabhakaran, Selva. Topic Modeling with Gensim (Python). https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/
- Project Gutenberg. https://www.gutenberg.org
- Saldaña, Zoë Wilkinson . 2018. Sentiment Analysis for Exploratory Data Analysis, *The Programming Historian.* https://doi.org/10.46430/phen0079.
- Sankoff, D. & Sankoff, G. Sample survey methods and computer-assisted analysis in the study of grammatical variation. In Darnell R. (ed.) *Canadian Languages in their Social Context* Edmonton: Linguistic Research Incorporated. 1973. 7–64.
- Shukla, et al., "Natural Language Processing (NLP) with Python — Tutorial", Towards AI, 2020
- Witten, Ian. 2004. Text mining. https://www.cms.waikato.ac.nz/~ihw/papers/04-IHW-Textmining.pdf