Welcome to The Carpentries Etherpad!

This pad is synchronized as you type, so that everyone viewing this page sees the same text. This allows you to collaborate seamlessly on documents.

Use of this service is restricted to members of The Carpentries community; this is not for general purpose use (for that, try https://etherpad.wikimedia.org).

Users are expected to follow our code of conduct: https://docs.carpentries.org/topic_folders/policies/code-of-conduct.html

---

Welcome to the workshop!

Links

- Lesson materials: https://carpentries-incubator.github.io/python-text-analysis/
- Slides
    - Intro: https://docs.google.com/presentation/d/1Rm-mqUFE0b5JqLinRViZpWRtL7Pemr22KPzTPdYZP_4/edit?usp=sharing
    - Preprocessing, Vector Space, TF-IDF: https://docs.google.com/presentation/d/1Eg8MxK_pDXMPoVWjZ3dasrH2NxWhaLke_VIn0sZh4qE/edit?usp=sharing
    - LSA: https://docs.google.com/presentation/d/164b06LD6Cfz8Tenm1cBKytOnls13mh3A30uN2Q7fl00/edit?usp=sharing
    - Word2Vec: https://docs.google.com/presentation/d/1TAOuCYz3osVJWRWhDW_Pjj61OVNjjcI9IfDB2PLPdQw/edit?usp=sharing
    - Wrapup: https://docs.google.com/presentation/d/1HxA97ic1xUIBZGKfQJussVB9YK90b5uSlh6h_xyQ2qw/edit?usp=sharing
- Daily feedback: https://forms.gle/xVmNYqX3K39qmk2x9

Day 1

Sign in

Name, department/program/affiliation, and one thing you're looking forward to this summer

- Chris Endemann (he/him), Data Science Hub - going to yellowstone
- Ryan Bemowski (he/him), Data Science Hub - yard work
- Jennifer Patiño (she/her) UW-Madison Libraries - figuring out how to grow things in my shady garden
- Karl Holten (he/him) UW-Milwaukee Libraries- Festival season in Milwaukee
- Ann Hanlon (She/her) UW-Milwaukee Libraries - peonies and lilacs blooming
- Pubudu Kumarage (He/him) Waisman Center/UW- Madison - Lake Superior
- Erwin Lares - he - Research Cyberinfrastructure at DoIT - all the arts festivals I'm attending!
- Hsuan-Wei Liu (he/him), ISyE, UW-Madison - trip to DC
- Chris Henson (he/him), SMPH Data Analytics - Summer festivals all around
- Dorothea Salo (she/her), Information School, UW-Madison - American Players Theatre season!
- Erika Zhu (she/her), Life sciences communication, UW-Madison - travel to CO and CA
- Soyeon Choi (she), psychology
- Hadley Rahrig (she/her), psychology, UW-Madison - vegetable garden
- Qiuyu Yang (she/her), department of surgery, trip to Alaska
- Jie Chen UW-Milwaukee Libraries - travel
- Sijan Thapa
- Khine Thant Su, Department of Medicine, travel to San Antonio for a six-day bootcamp on reproducible research methods

Notes

Setup walkthrough — REDO this if completed before workshop. Some files have updated.

1. Clone the lesson GitHub repo (main branch): https://github.com/carpentries-incubator/python-text-analysis

- a. Click the link below to open the Github page.
- b. Click the green “Code <>” button.
- c. Click “Download as zip.”
- d. Unzip the directory to your desktop or working directory.

2. We will be using Google Colab to run Python code in our browsers. Colab was chosen to ensure all learners have similar processing power (using Google’s servers), and to streamline the setup required for the workshop.

- a. If you’ve never opened a new Colab notebook, first visit the Google Colab website (https://colab.research.google.com/) and click “New notebook” from the pop-up that shows up. When you open your first script, a “Colab Notebooks” folder will automatically be created in Google Drive.
- b. Visit Google Drive and find a newly created “Colab Notebooks” folder stored under MyDrive, /My Drive/Colab Notebooks
- Create a folder named text-analysis in the Colab Notebooks folder on Google Drive. The path

should look like this: /My Drive/Colab Notebooks/text-analysis/.

- Upload the “**data**” , “**code**” , and "**notebooks**" folders that were downloaded from git (inside the “python-text-analysis-main” folder) to the “text-analysis” folder you created in Google Drive:
    - /My Drive/Colab Notebooks/text-analysis/data
    - /My Drive/Colab Notebooks/text-analysis/code
    - /My Drive/Colab Notebooks/text-analysis/notebooks

Episode 1 - Intro to NLP

https://carpentries-incubator.github.io/python-text-analysis/01-basicConcepts/index.html

Episode 2 - Building a Corpus

https://carpentries-incubator.github.io/python-text-analysis/02-apis/index.html

PyTesseract is one way of doing Optical Character Recognition in Python: https://uw-madison-datascience.github.io/ML-X-Nexus/Toolbox/Libraries/PyTesseract.html

Building Legal Literacies for Text Data Mining: https://berkeley.pressbooks.pub/buildinglltdm/

Model cards: https://huggingface.co/docs/hub/en/model-cards

https://www.gutenberg.org/cache/epub/158/pg158.txt
https://chroniclingamerica.loc.gov/lccn/sn83030431/1917-07-15/ed-1/seq-47/

Gutenberg also has audio versions! Which is exciting for "multimodal" (text+image) applications

Episode 3: Preparing and Preprocessing Your Data

https://carpentries-incubator.github.io/python-text-analysis/03-preprocessing/index.html

any order for these preprocess?
- Tokenization happens first, the order for the other steps isn't as important

Do we need to do all these preprocessing steps before we put the texts into huggingface models?
- The huggingface models have preprocessing and tokenizers. So they've already built the pipelines for a lot of the models. But it is good to know what those pipelines are doing.

Episode 4: Vector space

https://carpentries-incubator.github.io/python-text-analysis/04-vectorSpace/index.html

Episode 5: TF-IDF

https://carpentries-incubator.github.io/python-text-analysis/05-tf-idf-documentEmbeddings/index.html

Day 2

Sign in

Name, department/program/affiliation, and any questions you have about yesterday's material

- Chris Endemann (he/him), Data Science Hub
- Karl Holten (he/him) UW Milwaukee Libraries
- Jennifer Patiño (she/her) UW-Madison Libraries
- Pubudu Kumarage (he/him) Waisman Center UW
- Hsuan-Wei Liu (he/him) UW Madison
- Dorothea Salo (she/her), UW-Madison iSchool
    - Oh! I did have a question. Broadly speaking, how much text is needed for topic modeling to yield useful results? E.g. would it work on a corpus of newspaper headlines, or is that too sparse and having the full news stories would be better?
- Ann Hanlon (she/her), UW-Milwaukee Libraries
- Erwin Lares - he - Research Cyberinfrastructure at DoIT
- Ryan Bemowski (he/him), Data Science Hub - Helper
- Hadley Rahrig (she/her), psychology, UW-Madison
- Qiuyu Yang (she/her), department of surgery
- Chris Henson (he/him) SMPH Data Analytics
- Soyeon Choi (she), psychology
- Sijan Thapa (she/her), AAE
- Khine Thant Su (she/her), Department of Medicine. When doing unsupervised learning like topic modeling, are there guidelines to determine the optimal number of topic groups (like the value for k in k-means clustering?)

Review of libraries used so far

- **sklearn (https://scikit-learn.org/stable/):** general purpose machine learning (ML) library that supports many models, embeddings, and preprocessing methods
    - TF-IDF embedding, distance metric
- **spacy (https://spacy.io/):** spaCy is a fast, production-ready NLP library in Python that provides tools for tokenization, part-of-speech tagging, parsing, named entity recognition, and more, with support for pretrained language models.
    - spacyt = spacy.load("en_core_web_sm") # is loading a pretrained small English NLP pipeline from spaCy, and assigning it to the variable spacyt. Gives you tokenizer, part-of-speech tagger, named entity recognition, and lemmatizer

Day 3

Sign in

Name, department/program/affiliation, and any questions you have about yesterday's material

- Chris Endemann (he/him), Data Science Hub
- Karl Holten (he/him) UWM Libraries
- Ryan Bemowski (he/him), Data Science Hub
- Hsuan-Wei Liu (he/him), ISyE
- Erwin Lares - he - RCI
- Ann Hanlon (she/her) UWM Libraries
- Soyeon Choi (she), psychology
- Pubudu Kumarage, Waisman Center, UW
- Hadley Rahrig (she/her), psychology, UW-Madison
- Qiuyu Yang, department of surgery