WELCOME

If you have a remote teammate, please use headphones

DAT 562 TEXT MINING

PROF. YULIA NEVSKAYA

- Lab 2: Text Pre-Processing and Feature Engineering

The Plan for Today

- We'll be doing **text normalization and vectorization** in Python
  - **What is text normalization (pre-processing)?**
  - **What is text vectorization?**

Plan for Today

- We'll be doing **text normalization and vectorization** in Python

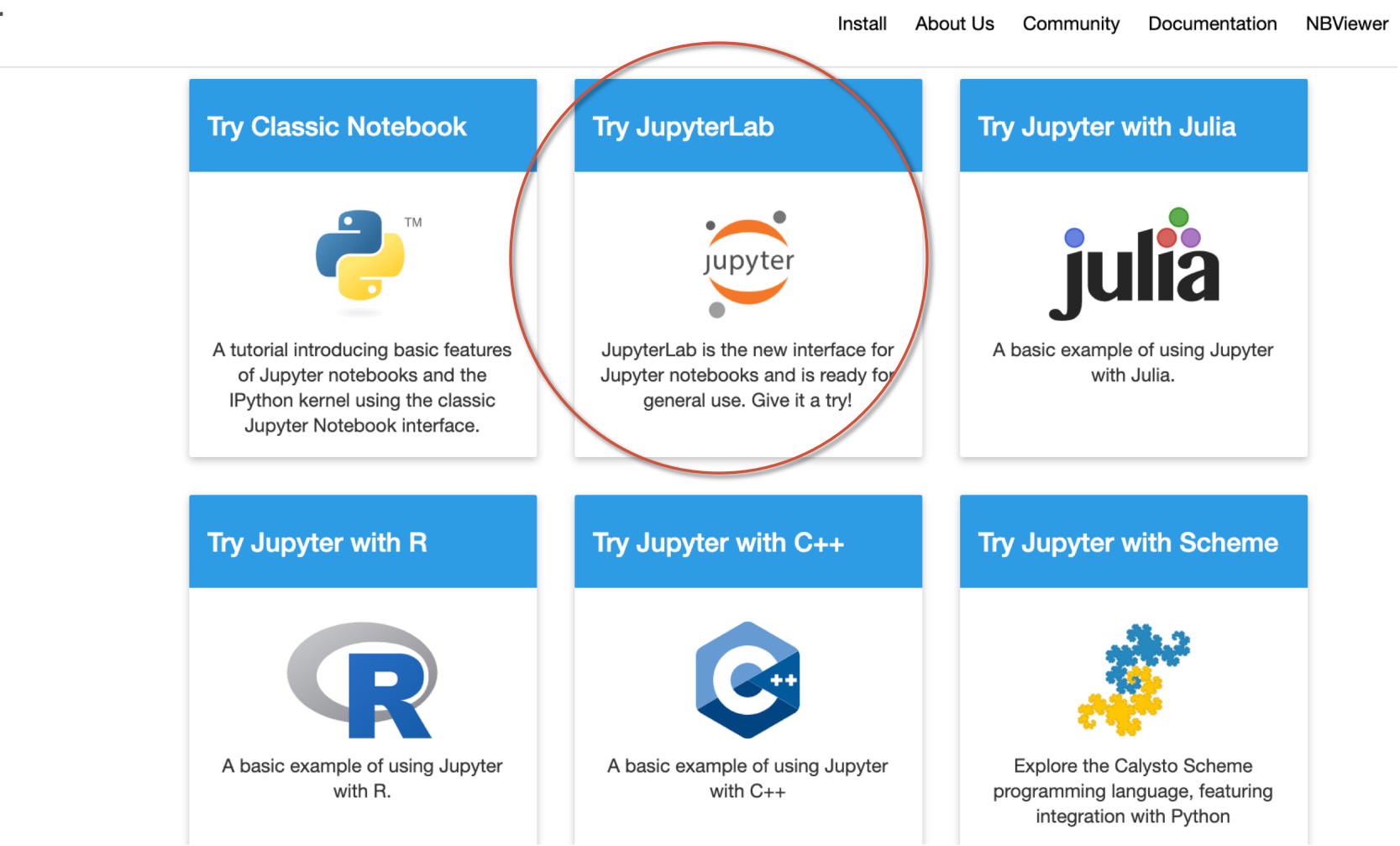
  - **What is text normalization (pre-processing)?**

    Preparation of text for feature extraction (vectorization): tokenization, stemming, stopwords removal and so on

  - **What is text vectorization?**

    Feature extraction: engineering of variables that represent text data in a format suitable for analysis using major machine learning algorithms (methods: "bag-of-words", TF-IDF, n-grams)

Jupyter.org/try (Try JupyterLab)



The Lab Process

- Your assignment is on Canvas:
  - You **can collaborate** on the assignment with your team
  - Upload your answer **individually** at the end of the lab (recommended) or by Wednesday, 11:59pm St. Louis time

- If you have a **remote** teammate:
  - Please log in to Zoom to see your teammate
  - Use your headphones if you are on Zoom

What You Need to Do Today

- Go to Canvas / Module 2 / During Lab to get the assignment
  - Download two (2) Jupyter notebooks and put them into the same directory(!) on your machine
  - The provided Jupyter notebooks will:
    - Guide you through text normalization and vectorization Exercises
- **YOUR TASKS FOR TODAY:**
  - **1) Do all the EXERCISES and answer QUESTIONS in the script**
  - **2) Submit individually the Jupyter notebook with your results and answers on Canvas**

Reminders

- Start work on the Group Project:
  - Explore available datasets
- Office Hours in-person and on Zoom (all times are CST):
  - Wednesday 9:30-10:30am
  - Friday 12:30-1:30pm