Friday 9:30 – Data Validation for Data Science

All the materials and notebooks needed can be found in this repository: <https://github.com/NatanMish/data_validation>

Friday 11:00 – Feature Engineering Made Simple

Participants will perform a hands-on, end-to-end, feature building exercise, with particular emphasis on feature engineering using Anovos (<https://anovos.ai/> or <https://github.com/anovos/anovos>)

Friday 13:30 – Introducing more of the standard library

Repository link: <https://github.com/simonwardjones/pydata-talk-2022>

Friday 15:30 – Document/sentence similarity solution

- A simple dataset of book titles sourced from: <https://raw.githubusercontent.com/noahjett/Movie-Goodreads-Analysis/master/books.csv>  
- The classic 20 News Group data sourced from Scikit-Learn dataset module: <https://scikit-learn.org/0.19/datasets/twenty_newsgroups.html>  
- STS benchmark dataset located here: <http://ixa2.si.ehu.es/stswiki/images/4/48/Stsbenchmark.tar.gz> further details on this benchmark can be found here: <https://ixa2.si.ehu.eus/stswiki/index.php/STSbenchmark>

Saturday 13:30 – Data Validation for Data Science

Suresh, H., Guttag, J., Kaiser, D., & Shah, J. (2021). Understanding Potential Sources of Harm throughout the Machine Learning Life Cycle. MIT Case Studies in Social and Ethical Responsibilities of Computing, (Summer 2021). <https://doi.org/10.21428/2c646de5.c16a07bb>