Pipeline Creation

# **Dan Siegel Spring 2019 github.com/danjsiegel**

# Which Domain?

What domain is this data going to come from? Please list 10 references (with a brief annotation) to use to make sense of what you’re doing with these data.

1. <https://towardsdatascience.com/building-an-etl-pipeline-in-python-f96845089635> - good tutorial on how to make a pipeline. This one is based upon NLP and cleans text in advance of training
2. <https://scikit-learn.org/stable/modules/generated/sklearn.pipeline.Pipeline.html> pipeline documentation from the python library I'll be using
3. <https://medium.com/vickdata/a-simple-guide-to-scikit-learn-pipelines-4ac0d974bdcf> another guide to the workflow
4. <https://medium.com/@yanhann10/a-brief-view-of-machine-learning-pipeline-in-python-5f50b941fca8> an overview guide to pipelining
5. <https://www.kdnuggets.com/2017/12/managing-machine-learning-workflows-scikit-learn-pipelines-part-1.html> a guide to building 3 pipelines, Log reg, SVM, Detrees
6. <http://queirozf.com/entries/scikit-learn-pipeline-examples> another example of pipelines
7. <https://ramhiser.com/post/2018-04-16-building-scikit-learn-pipeline-with-pandas-dataframe/> again, another pipeliner
8. <https://medium.com/@datamonsters/text-preprocessing-in-python-steps-tools-and-examples-bf025f872908> preprocessing with NLTK
9. <https://towardsdatascience.com/productionizing-your-machine-learning-model-221468b0726d> article about making your model production ready with a restful API
10. <https://www.analyticsvidhya.com/blog/2017/09/machine-learning-models-as-apis-using-flask/> Flask to make a restful API to interact with a model

# Which Data?

What is the dataset you’ll be examining? Please provide a codebook if there is one or a link to the dataset as well as a detailed description.

<https://www.kaggle.com/c/spooky-author-identification/data> "The competition dataset contains text from works of fiction written by spooky authors of the public domain: Edgar Allan Poe, HP Lovecraft and Mary Shelley. The data was prepared by chunking larger texts into sentences using CoreNLP's MaxEnt sentence tokenizer, so you may notice the odd non-sentence here and there"

# Research Questions? Benefits? Why analyze these data?

How are you proposing to analyze this dataset? This is about your approach. Here, you’ll be proposing your research questions as well as justifications for why you’d offer these data in this way.

The research here is focused around productionizing code, and shipping something useable, not just a fun jupyter notebook.

# What Method?

What methods will you be using? What will those methods provide in terms of analysis? How is this useful?

I will be using SKLearn's pipeline features. I will try to productionize to a restful API if I am able to, but at the very least, I want to create code which can take input csv, processes it and spit out a prediction.

# Potential Issues?

What challenges do you anticipate having? What could cause this project to go off schedule?

Not super familiar with Flask. I took an Udemy course on it, but I have never published something to production with it.

# Concluding Remarks

Tie it all together. Think of this section as your final report’s abstract.

Productionizing text analysis through pipelines and potentially APIs.