Derek Stearns

Tools for Data Science writeup

A lesson in boiling oceans

**Introduction**

The goal of my project was to create a class that would be useful at collecting tweets in the final two weeks of the project. I wanted to learn more about Spacy and how it could be used to normalize a dataset of tweets for purposes of the Wagner Group project, also called “Project 2”. The class would have a method to collect a number of RSS fed URLs, API credentials, search string, and options for storage so that a schedule could be utilized to maximize and automate collection of data over time. The output would be a dataframe that could be stored for later analysis.

Analysis of plan execution

Looking back at the proposal for this project, it could have been better managed. The proposal submitted suggested to first search for possible news sites by week 2 and to use weeks 3 through 6 to process NLP entries. To a degree this process was followed and found success. Unfortunately, the graph theory course I took this semester was more intense than anticipated. Additionally, the homework in this class as well as my professional and personal duties took unprecedented time away. With each break in task switching, it became increasingly harder to remember where I had left off in coding my project, making perceived forward progress challenging.

For this project I understand there is a trade-off balance between multiple spectrums: time, amount of data, complexity of code, usability of code, cleanliness of data and usefulness of data. Being a perfectionist at heart, I didn’t prioritize saving messy, raw data as I probably could have done; but rather opted toward the complexity and usability of code part of that balance. Seeing reminders in discord chats regarding a requirement for execution of the code in any environment also created additional perceived pressure to ensure the project code submitted was more hardened than the data. The breadth of application to multiple types of data streams also likely contributed to possible bugs in the code.

I understood that many other students would likely be using Google News and/or twitter as their source, so naively, I opted to start digging into alternatives in RSS feeds as a breadth first search strategy for possible novel data. As I started on working in Spacy, I took time to pause one working session during November to request the Twitter API credentials. It took about 24 hours for them to approve me. I also noted year’s past API projects I had created were no longer functioning with the upgrade to v2. The breadth first search strategy took me down a path working more closely with Spacy on RSS feed entries than I had in any previous work. At some conscious level I realized I was taking a bit too long to find and clean fringe data, however, working with spacy linguistics and entities had me convinced the code I was generating would be useful for future use.

**Successes**

As mentioned in the introduction, there was success. I found RSS feeds that contained “Wagner Group” references and, in the Jupyter notebook I continued working in, I was curating the single, empirical dataset with two additional xml feeds in queue. I was able to get the first mostly complete. However, working in a single set will eventually lead one into what could be considered a depth first strategy. Not continuing to evaluate additional RSS feeds in weeks 4 – 6 would lead toward a lack of data stored locally over time as well as last minute frantic coding.

Additional success was found when I was able to find a blocks of hours on December 11th and 12th. I found that all of my previous code written would be useless as I tried to assemble the data and complete the writeup and presentation. I had forethought to setup the twitter credentials while working on the Spacy code as mentioned. I was able to reference my bearer token and pick up where I had left off parsing tweets. I was able to write a quick iterative function that would use the entities after reading through Twitter’s V2 API that could save data in an acceptable format to the project plan.

When considering the definition of “learned”, it might be similar to say what “one remembers” going forward in perpetuity. However, I understand that knowledge is not only based on perception and is individualized, but knowledge will be lost if not practiced. And while this terminal mindset is certain, we must try to create new knowledge from our perception of causality in experience. Concretely, I learned and remember how Spacy treats its entities and the differences in wording when identifying the types of an object in the sentence. One less-specific learned concept was finding how *not* to go about creating a class.

**Learning experiences**

Writing stories and code could draw similarities to tree traversal. As one topic is opened, the depth of the idea can be explored, branched into and finally recursed once the appropriate depth has been determined; also called the “definition of done” in agile terminology. As a recognized pattern of human behavior, sometimes it can be difficult to define “done” and to segment stories or chapters away from a linear or stream of consciousness way of thought; and this project was not an exception in that sense for me. It reminds me of when you ask a person to cut a perfect circle from paper. It is possible a person will continue to cut it, for perfection’s sake, until the circle is too small to continue or nothing remains.

I have worked with programmers who seemingly have the ability to define their classes, methods and attributes with ease and in stream of thought. And in the context of class work, where object-oriented programming is practiced in a concrete way, that is generally easy for me. However, as I tried to define a class in this course, the consistent required task switching due to promises and liabilities outside this course created a difficult environment for consistent strategy to be applied to the class code, and as a result ended up with many issues in the code and a final code submission that with a higher probability to fail a check.

**Future learning experiences**

As I continue to mentor others on their python journey it’s important to remember that these journeys are not linear and that a singular experience does not define one’s competence. It certainly makes me wonder what a college professor’s experience could be like and how they approach evaluation of student learning and performance in this domain.

As I approach additional projects, I think it will be likely that I continue to practice code segmentation in a more objectively scoped and fractured way by using packages. This may help me keep rigor to structures already defined as well as manage scope of future projects.

**Conclusion**

There were some great successes in this course project and I learned practices I can avoid to be more effective in generating python class code. I also learned of the Wagner group’s existence, their recent headquarter destruction and more about Spacy and Twitter API. I continue to appreciate learning what size project I might be able to accomplish, or what size ocean, lake or pond might be boilable. I appreciate the opportunity and shared learning experience.