Research Proposal

Jason Lunder

CPSC 325

In exploring my potential project, I will be researching a few key areas of data science and machine learning. My current project ideas relate to analysing the spread of misinformative, false, or “fake” information, specifically developing/improving on methods to track its spread through only using the linguistic content (though object detection in visual content as a supplement may be helpful) of the information itself. This will thus involve computational linguistics/natural language processing primarily, with possible excursions in computer vision. I plan to use python and python frameworks almost exclusively for this project, which I have extensive experience in. I may use Google collab if necessary as well.

In order to do this, I will need to research current and past attempts to track the spread of misinformation, including registries of sites, news outlets, and particularly publishers of scientific journals that have been deemed as “predatory” or otherwise identified as a source of false information, in order to narrow down my data collection process. In this, I will need to investigate various methods of webscraping and api access, as I may need to gather informotion from many sites and services. This, coupled with a need to display and organize data, will likely require the writing of an API and utilization of data visualization programs/platforms. Ideally, I will not need to host anything on cloud services. I have some exeperience with a couple frameworks for doing this that I know can be hosted locally and for free. These are Bentoml for the API and Grafana for visualization, though I will likely explore other data visualization programs, as Grafana was difficult to work with in my experience.

Also required will be utilization of pytorch and surrounding frameworks, which I have experience in. With this I may need to implement/use a number of algorithms/ML architectures that I do not have as much experience with, including transformers and GANs. Some of these, in conjunction with simpler, fully connected networks, LSTMs and CNNS, that I have more experience with, may be useful in my project, so I will research their uses, benefits, drawbacks, and implementations to see if they may apply to my project and determine how I can include them.

In order to research transformer models and GANs as they pertain to my use case, I’ll first study the primary papers associated with each, specifically, <https://arxiv.org/abs/1406.2661> and <https://arxiv.org/abs/1406.2661>. These should help me gain insight into the types of problems that these architectures are useful on and determine if and how these are applicable to my use case. This should take no more than a week, if I devote approximately 4-5 hours per paper. My notes on the theory and algorithms explained in the papers will be evidence of my completion of this research task.

I will then continue my research by looking for existing models that can make predictions that would be useful to my investigation, including semantic analyis models, models that predict membership of text in a larger corpus, and generative models focusing on linguistics. This will help me to gain insights into the application of these architectures as it may pertain to my project, and help me refine my project into something novel, if similar things have been done. I will devote more time to this, probably around 12 hours of work, as I will be attempting to clone and demo some of the most relevant models locally to investigate their abilities beyond what I can read online. If I find particularly relevant models, I may be able to incorporate them as is into my project, or simply fine tune them and include them. My selection and demo of a couple key models will be evidence of my research on this.

I will likely not need to spend much time researching technologies for supporting infrastructure, like api frameworks, as I already have a couple in mind that I have worked with in the past and am comfortable with. I will set aside around 3 hours, towards the end of my research, to determine if the frameworks I am comfortable with have the capabilities I need to preform the necessary tasks for my project. I will also set aside an additional 4-6 hours with this investigating the capabilities of data visualization programs, especially focusing on their ease of use and their integration with my chosen api framework. I’ll make sample APIs/data visualizations from the documentation of the various api and data visualization frameworks to test this, which will be evidence of my completion of this research task. I will also set aside an additional 3 hours to briefly investigate Google colab and other potential services for training models, if necessary. This will be evidenced by my implementation of the previously mentioned example models on my chosen hosting platform, which may ultimately be my local machine.