William Pembleton

10/31/2018

Monday Presentations

Hunt: Hunt's project is based on trying to solve a staffing issue in Open Source projects. He's wondering if repository structure correlates with contributor activity. He's planning on getting his data from the Github API. The main problems he's going to run into is that repositories aren't marked as pubic. The way he got around this is he's only going to pull from repositories that are public and have at least one commit from a developer who is not an owner of the repo. He's planning on using CART to make a tree that will predict how likely a repo is to be successful.

Courtney: She's doing a repeat study of a paper she found. She wasn't happy with how the paper was carried out so wants to replicate it. She's going to be reusing the data collected by the original authors because it was gone through by hand and she doesn't want to do it again. She's plans to use GAN to generate and also discriminate if a particular piece of text is positive or negative.

Kyle: Kyle wants to test if city planners have prejudice. He's planning on doing this by combining data from road length and road width from a dataset and mixing it economic data from the US census. On the plots he's going to be putting things like road width against average salary in a region. He's planning on using K-Means to do the clustering.

Andy: Andy is going to be using topic analysis on a few different epics stuff like the Odyssey. The chose this topic because it combined their interests of the Classics and DM/ML. Topic analysis is used for building up models of well topics. Topics are a mathematical way to describe groups of words with a similar context to them. For instance, words like bird and parakeet would be in the same topic while words like bird and car would be in different topics. The goal for this project is to be asking questions on what the structure of this epics are through the lens of Topic analysis and how the translation of these texts change over time.

Jake: Jake is planning on using convolutional neural networks to transform pictures taken at NCF into pictures in the style of pictures his friends will make. A convolution neural network is based on a neural network except that instead of weights there is these things called convolutions. Convolutions are a set of pixels with weights on them for how much to turn up or turn down a pixels intensity. So, when the model is training it is adjusting these convolutions.

Wednesday Presentations

Eric is planning on getting his data from MyAnimeList.net with rows like user score, genre, show length in episodes, season released, publisher and producer. He's planning on using CART, linear regression and possibly neural networks to tell what types of shows are the best for publishers to create and for viewers to decide what types of shows they should watch.

Angelo is planning on using neural networks, k nearest neighbors and a bag of words approach to classify pictures. They got their data from Kaggle which has tons of pre-classified pictures and they want to compare against the algorithms Google uses for reverse image search. They want to see where each model fails/succeeds and why.

Ty wants to use data scrapped from Ravelry (a knitting website) to guess at what type of object (hat, socks etc.) is being made based on the type of yarn that is used. Ravelry already tons of stats like color and thickness of each of the samples. They want to use K-means which will hopefully cluster into the objects, support vector machines and at the time of presenting multiclass logistic regression which they're not sure what it is but it looks like it might be useful.

Paul wants to predict movie endings based on the type of characters are in the movies. He plans on getting the scripts for 300 movies, describing what the characters' personalities are like and then building a tree with ID3 which will hopefully predict the ending of the movie. I think this project is going to be very time intensive for data collection and I wish them luck.

Rain wants to help people learn Kanji easier by creating a Convolutional Neural Network to make a connection between the components of a character and the meaning of the character. There already is a Word2Vec for Kanji characters and Rain wants to take those as inputs and build a model out of them.