Jacob Dineen

Text Mining

7/18/18

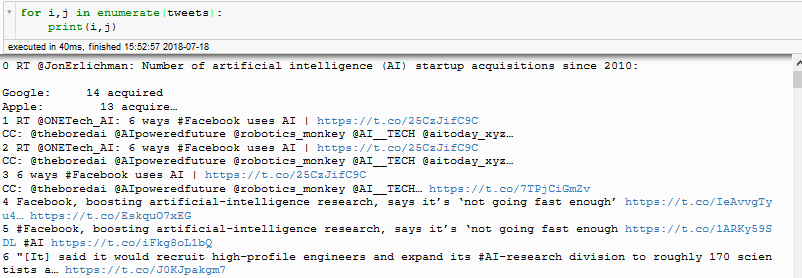
Homework 1

**Collecting a Dataset:**

To begin, under the suggestion from a fellow classmate, I decided to use an automated protocol to extract tweets from twitter regarding the state of AI with specific constraints requiring the tweet to contain some fuzzy matched combination of == [‘Artificial Intelligence’, ‘AI’, ‘Facebook’, ‘Twitter’]. This involved setting up a twitter dev account and extracting their API access tokens for use within the generators.



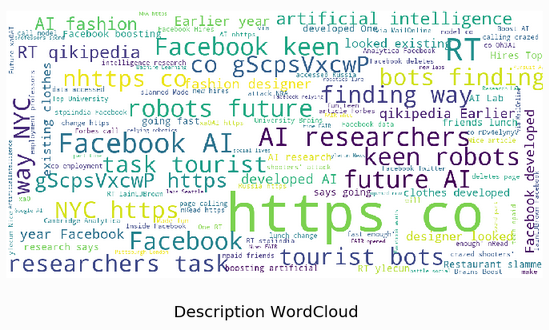
I’ve extracted the max tweets allowed before hitting limitations, mainly because I’ve read through some of them and there appears to be a lot of duplicates, mainly regarding news headlines that are prevalently retweeted.



For the purposes of this assignment, I think I can manually extract some of the sentiment derived from the text – Although I may find it more suitable to analyze comments rather than posts. I think it makes sense that posts, particularly from news sources, may have a more declarative slant than opinionated commentary from individual users. Unfortunately, twython (the python module) doesn’t appear to be able to delineate between posts and comments, so I will likely have to manually source that data into a csv file whilst searching for AI buzzwords.

I think there are two ways to look at the above expression – One is that news sources are supposed to be more factual than opinionated, so actual expressed sentiment may be harder to detect, and a lack of sentiment is not representative of the general population. Injected this into the training stage of a predictive model may unintentionally add a neutral level of bias when performing some level of approximation. Because we want our model to generalize well, we want enough variance to be adaptive. Additionally, a binary classifier with classes == positive or negative would be misguided when dealing with factual representations.

As I’m reading through my list of collected tweets, I don’t think that these postings would be a suitable representation of public sentiment toward AI and will focus instead on manual collection of user comments populated from buzzword searching on twitter and Facebook mentions. However, I do want to quickly generate a distribution visualization of the current headlines in AI:

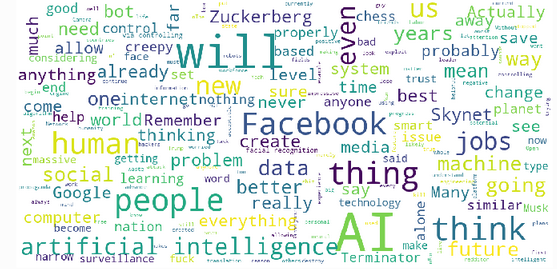


The above word cloud shows density by word for the approximately 1000 posts collected, and I think it backs up my point from above. There are no highly occurring words that would be expressive of an individual’s slant/perspective on the emergence and popularization of AI in the social media landscape. Most of the tweets were indicative of current research being conducted, or grabby headlines on product development.

Moving on to manual collection – I’ve scraped comments pertaining to Amazon/Google/Facebook specifically in the AI/Machine Learning/Deep Learning space. It is important now to realize that there is going to be an inherent bias amongst those that are not in the ML space, most of the population. People in the general population are more likely to not understand the inner working of a deep neural net as a function approximator and are instead influenced by flashy headlines denouncing progress in the space. In my opinion, however, this is the best representation of the population at large – A sometimes loud, collective voice that denigrates the use of ML in aspects of their life that they don’t want such progress to be prevalent. So, there are three main crowds of people I’ve analyzed, not just for this assignment, but in my time on reddit or reading Facebook comments:

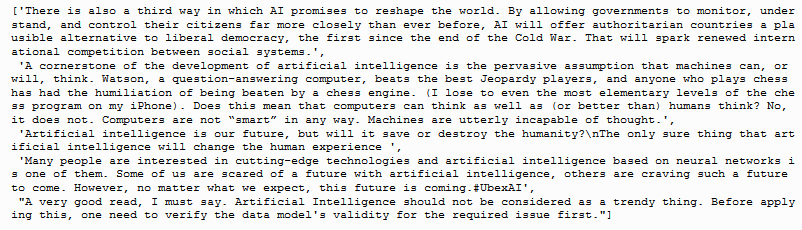
1. The doomsday-er. They claim that any small advance in this sector is an impending sign of Skynet and the end of mankind.
2. The informed – These people aren’t as concerned with the weaponization of AI in its current state as they are with the use of ML algorithms on data that they deem private and personal. These people generally see the benefits brought about by things like facial recognition for security and reco engines for movie selections, but fear overreach from Silicon Valley giants.
3. The realist – People in the ML space don’t see AI or deep learning as being close to general intelligence. They do relish in some of the things that the current state brings along, such as Reinforcement learning for gameplay(Deepmind/OpenAI), convnets and R-CNNs for autonomous driving (Tesla), and RNNs w/ LSTMs for translation and sequence to sequence modeling (Alexa/Siri).

A word cloud of my manually scraped comments/posts isn’t as insightful as I imagined. The curated list was mainly focused on negative reactions, as mentioned above, so I had anticipated some of the more fear-mongering terms to be more densely represented here, but that wasn’t the case:



To validate that point, beyond the articles that could be removed to reduce dimensionality, most of the highly occurring words are not words that would be indicative of any kind of opinion. I think what this mainly shows is that 80-100 samples are simply not enough to build up a vectorized representation of what the population truly feels about AI in social media.

*Sample of containing list (n=4):*



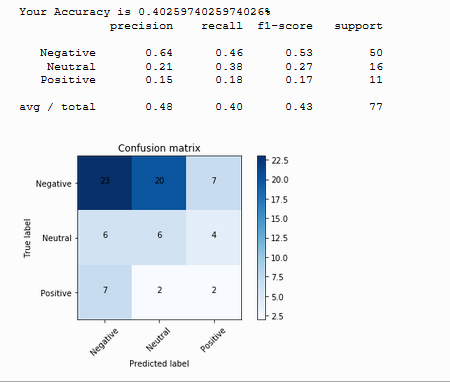
*Using Counter for word freq:*



**Testing NLTK and SentiStrength**

Admittedly, sentiment analysis could likely be conducted solely in python with some of the pretrained models defined under the NLTK library (& maybe sklearn), but that is slightly beyond the scope of this assignment. Instead, I have manually labelled my own opinion about the polarity of a string (in this case, a post/comment/reply,etc..) and will define objective(ish) success in that manner. I haven’t done a whole lot of work using parametrics for text mining, as I’ve mainly focused on nearest neighbors of CART trees, but I imagine as with most classification problems, the cost is generally defined using cross entropy – Binary sentiment analysis likely uses binary cross entropy, and multiclass probably uses cross entropy with some form of softmax activation (eg. softmax regression).

To begin, I will say that SentiStrength is not a viable option for detecting sentiment in the data that I have collected – There is a character limit that many of my posts exceed, and a partial analysis isn’t reflective of the entire corpus – I’ll instead focus on NLTK which has a char limit far exceeding any nchars of my collected strings. Plugging each result of shown by NLTK as one of three classes (Pos/Neg/Neutral), against my previously labelled dataset I can use some standard features of Sklearn’s API to show how well NLTK did:



First off, as mentioned above, most of the labels are negative, so there’s severe class imbalance, but because this is more descriptive than predictive, we’ll ignore it. The above results show that NTLK predicted the same sentiment as myself 40% of the time. As the confusion matrix and the classification report show, most of the mistakes (over half) were construing negative sentiment as neutral. It also had trouble with some of the positive comments, labelling them as having negative connotations or keywords that led to their classification as being negative in general.

This isn’t terribly surprising in context... I believe that a human has the benefit of understanding contextual clues better than machines, and that isn’t a knock on machines – I’m pretty sure that sarcasm is being accurately detected in some of these examples. However, when I was purging reddit and Facebook for some implication of the prompt, I had the benefit of labelling the data based on what it was in response to and prior world knowledge – This seems like it would be a better task for some kind of neural net with LSTM layers. A string/reply in the context of a bigger picture often doesn’t tell the whole story (I’m sure the architecture of the NLTK pretrained model is complex, however).

To wrap this up, I’ll conclude by saying that NLTK may be suitable for most sentiment analysis tasks, but it could also be advantageous to train a model from scratch rather than using pretrained weights on whatever data source this was induced on. To do so, however, would require a much more balanced, curated representation of sentiment. I also think it may have been an easier task if the prompt was referencing AI sentiment in general, rather than Facebook/Twitter AI sentiment. The narrower scope made retrieval a bit more difficult and sparse.