14/01/2020

Possible challenges

C4.xiv: Automated creativity

Given some information generate plots/story synopses

C5.C.i: Procedural interactive content

Auto generate levels and have an ai try and solve it

C8.iii.1: Novel user interfaces

Visualisation of data in vr

C6.B: Big data and data analysis

Clean and normalise tweets and categorise them

16/01/2020

Chosen challenge

C4.C.xiv

Use existing techniques such as GAN’s or constraint solving algorithms to learn how to synthesise examples that are similar to the dataset. Plots/story synopses (eg IMDB movie plot synopsis)

17/01/2020

Today I started researching around the GPT-2 topic. From this I gathered a few things to research:

* Machine learning
* Python3
* WordToVec

I started my research doc for the module and set up the basic document structure.

18/01/2020

Today I watched YouTube videos about machine learning theory and GPT-2 implementation and news to get familiar with the common phrases around the topic

To do list for week 20/01/2020 – 25/01/2020

* Learn basics of python 3
* Setup python IDE and needed library’s
* Research on how to make neural nets
* Make a simple neural net in python
* Explain the code in the research doc

20/01/2020

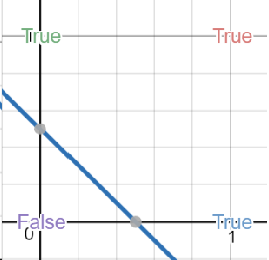
Today I started writing notes on NNs. Took notes on the makeup of a neural net and learned how to understand NN diagrams. Also took notes on how calculations occur in a NN and derived the formula of a neuron calculation. I referenced it against the actual equation from the literature and had derived it correctly. Looked into a few different activation functions and took notes.

21/01/2020

Today I took notes on how NNs learn and looked into the very basics of backpropagation, as I might need to use it at some point so understanding the basics will benefit me when and if I start to research it properly.

I finished programming my first Neural net in python. It is a perception matrix that is able to classify for the OR problem:

|  |  |  |
| --- | --- | --- |
| A | B | OR |
| 1 | 1 | 1 |
| 1 | 0 | 1 |
| 0 | 1 | 1 |
| 0 | 0 | 0 |



I followed the article from towardsdatascience.com to make the NN.

As I made it in python, I had to set up python on my computer. I downloaded python 3.8 and used notepad++ to write code in as I didn’t like using IDLE which was provided with python. I don’t like using notepad++ either so I will need to find a better text editor for writing in python.

This was my first-time using python so I’m still getting used to the syntax and am going to have to take notes on the basics and theory.

Due to this inexperience in phyton syntax there where a few errors in my code that where due to syntax errors that took a while to figure out and fix but the NN does now work.

22/01/2020

Today I mainly focused on setting up python so it would be easier to use. I decided on visual studio code as an editor and discovered a plugin called Kite that provides helpful functions within the editor (intellisence, documentation, etc)

23/01/2020

Had my first meeting today and took some notes on what else I should include in this diary:

* Talk more about my feelings for how I see the project going
* Compare what I think I need to do to what I actually do
* Talk about what im unsure about

Today I also finished the notes for the code example on the NN

24/01/2020

Today I focused mainly on getting all the basics of python down. I did a little bit of work on the NN code by adding some screenshots of it working at the compile time to the research doc.

25/01/2020

Week roundup:

This week I was able to finish everything on my to-do list. This could be an indication that I’m focusing on things that are too easy or too familiar. This week I plan to essentially throw myself into the deep end and go straight into the coded examples of poetry generation from gwern. Ill either be able to understand it or ill gather information on what I need to know. Ill think about what ill put in the to-do list over the next day and then write it up on Monday morning.

27/01/2020

To do list for week 27/01/2020 – 01/02/2020

* Follow the Gwern blog and get a working version running myself
* Work on the html webpage

I decided to only put 2 things on my to-do list this week as I expect the gwern example to generate more things to learn as I am working through it. I also want to make sure I start work on the webpage. For the rest of today I read through the Gwern site.

28/01/2020

Today I was only able to get a little bit of work done in the morning, so I used this time to create a virtual environment in python. I did this as when I was reading through some code on gwern it said about setting up a tensorflow virtual environment so after googling it, it recommended using a Venv. I created the Venv and assume that I will need to load tensorflow in as a module so I will do that tomorrow. I currently don’t really know what way I’m supposed to emulate the results from gwern at the minute I think it is basically just importing packages and running them, but I don’t really know so I’m going to need to look into it more.

29/01/2020

As I created a venv yesterday, today I was going to install the tensorflow module into it. I kept getting errors about no version and after digging around on stack overflow I discovered there’s no version of tensor flow for python 3.8 so I will have to use python 3.7. I couldn’t find a way to downgrade pyhton in the venv I made and also found out the method of creating venvs I used is exclusive to python 3.8. after looking around on YouTube for tutorials on making 3.7 environments I saw a tool called anaconda being used so I downloaded it with the hopes it would help me make the environment. Thankfully it did and it was an easy process with a gui wizard. The 3.7 environment was easily setup and I installed the TensorFlow module into it.

I don’t really know where to go from here but I did see this on the Gwern site: “I assume you have a fully-working Nvidia CUDA & GPU-enabled Tensorflow installation, and have either run other DL code successfully or run the TF installation checklist’s MNIST toy example to verify that you have working GPU training.” So I’m going to look into these 2 things to make sure I have tensor flow working.

30/01/2020

I read through the rest of the Gwern and found a link to another tutorial on medium.com. this one seemed to be a tutorial on getting the GPT-2 set up for training. The first thing it says to do is get a git clone of the pretrained model and project. I had never used Git hub to clone a project before just to download already compiled programmes. I googled how to get a clone and found I could get it through visual studio code. For some reason every time I tried to use the Git clone command in vis studio it would say command not found so I need to figure out how to fix this.

31/01/2020

I figured out the github problem, I found out that you must have the github framework installed. Once I had this installed the git clone worked fine and I got a clone of the project. The next thing I had to do was setup the venv for this clone I set it up in anaconda with python 3.7. after making the venv I ran the python --version command but kept getting “Pyhton not recognised as a command”. After a quick bit of googling I found a YouTube tutorial on how to fix it. I had to create a system global variable with a path to the python install folder. This fixed the python problem. Next I installed the 4 modules that the tutorial said I needed this was fine as I had done this before, so I knew how to do this. Next I downloaded the model. The dataset was the next thing I needed to make. This was just a plain text notepad doc with a specific separator between blocks of text. I grabbed a few poems from Gutenberg poetry. Not enough for a good dataset but should be fine for just learning how to set it up to train on this data.

Week roundup:

This week I couldn’t get the GPT-2 model trained and making samples due to the confusion at the start of the week. Currently I am working right on the edge of my understanding so each step of the tutorial I am following requires quite a bit of googling around it to get things to work. Hopefully next week I will be able to get the model working. I had a little look at the webpage and a bit of the HTML. Ill need to finish this next week. I plan to do the tutorial on setting up the GPT-2 model.

3/02/2020

To do list for week 3/02/2020 – 8/02/2020

* Finish the GPT-2 tutorial
* Finish the HTML webpage

Now I have to encode my dataset for training. This is where most of the big problems occurred. First, I found out through error messages that I was missing a few python modules, numpy and toposort. So I installed these. The next problem was with the version of tensorflow I was using. The version I had installed was the latest stable build, but it turns out functionality that was in an older version of TF but not in this one was needed. So, I uninstalled the current version of TF and installed TF v1.14. The next problem was with directory locations I spent a good part of the morning going through stack overflow and other forums to find a fix for this error. Instead of going through the source code to fix it myself I emailed my supervisor for help. He suggested going through the source code, so I did and the problem was a lot easier to fix than I thought it was going to be. So in hindsight I should have just gone through the code before I emailed him even though I was overwhelmed by it as I would have very quickly noticed that I was a rather easy fix. The problem was that the code was trying to find a directory it didn’t have access to so all I had to do was paste in the file path and then the encoding worked fine. The next step was training. Before this worked, I spent a few hours debugging error messages most of which ended up being easy directory fixes. Once these where fixed I started training and it was producing samples. So, I successfully set up the GPT-2 model.

4/02/2020

Today I just worked on wringing the tutorial for the poetry model in the Doc and an now making it in html form for the website. I am finding Html and css very easy to use. I have never used either before but have got the jist of it.

5/02/2020

Today I wanted to analyse the output I was getting from my model. In last weeks meting John said something towards the end of the meeting about how GPT-2 has some inherent problems so I wanted to see if I could notice some major flaws in the samples it created. I had quite a large number of samples from my poetry model as I had left it training when I want to class. So I started to go through them and noticed they where all a bit weird. I noticed it would find a phrase or word it knew and would repeat it over and over in the same sample or in some cases it would even just generate a poem I have as an input word for word. I thought this could be due to the fact I used a rather small data set with only 3 or 4 poems, as my main focus was to get it working first. I decided the best way to analyse the output would be to setup a new model and train it on a dataset that I would know all the contents off so I could easily pick up things like duplication and replication. For my new dataset I used all the lyrics from the beetle’s album Sgt pepper’s lonely hearts club band.

I Picked this album specifically as it had a good few songs in it that don’t entirely follow a generic song in how its written. I believed the absurdity in the lyrics would give the model a bit of leeway in creating new songs so that if it didn’t make a sample that really made sense you could still kind of see its similarities from songs in the albums.

This essentially turned out to be a complete failure in the hope I had in GPT-2 making something of value. The samples actually managed to get worse as generations went on. It managed to write an entire song that was over 100 lines long where every line started with eh word “Where”, and this was an example where it at least produced something that resembled a song.

This experiment really changed my outlook on this project. Originally, I assumed I would be able to do this with GPT-2 with just tweaking of the model and a specific type of training. Now I know I will need to find at least some other method to help if not take the place of GPT-2 in this project.

6/02/2020

Today I just tidied up the website a bit and thought about the future of this project due to my concerns with GPT-2. I didn’t speak much about how I made the website earlier so ill cover it here. To make my website I used visual studio code I learned the basics from w3Schools.com. I didn’t use many fancy implementations on the website except a small accordion that required a little bit of java script. I am happy enough to work with html and found it easy to pick up.

7/02/2020

Week roundup

This week I got the model up and working and was nothing less than disappointed with the results. The main thing I noticed is that I will need to apply some other sort of method to GPT-2. I’m going to start looking into NLP processes and some research papers of GPT-2 NLP combinations like john suggested. I think I’m going to need to apply some sort of structure with NLP to a GPT-2 generation like process. And I think ill need some sort of after pass to go over the result as I think structuring GPT-2 will just result in structured nonsense. So some sort of cleaning up will need to occur to make the end result make any sense. The though of starting from scratch to now learn a new method is very daunting but seeing as I was able to do it for GPT-2 I don’t see why I shouldn’t be able to do it for NLP.

12/02/2020

I only got around to starting work today as ive been trying to get coursework up to date for csc 1031 as ive had a bug in the code that ive been working on so this has been on the back of my mind. Today I read a few articles on nlp and am looking for research papers on nlp and gpt-2 as john suggested.

13/02/2020

Today I started transferring my diary to the format provided by john itl probably take me a few hours to get everything organised and then im going to start looking at the uber research model.

15/02/2020

Week Roundup

This week I didn’t really get much done. I tried to find some papers on language processing but didnt get to far. I tried using the top 100 journal rankings like john said but got frustrated when I couldn’t figure it out so just left it. Coursework for csc 1031 is still priority at the minute and looks like it will be until after next week so im going to try and chip away at finding research papers next week but I don’t think ill make any significant progress.

18/02/2020

Getting this coursework done for csc 1031 is still main priority but I felt that I needed to at least say something in the diary so that I wouldn’t have just a big gap of nothing done. I haven’t been extensively researching into the topic but have watched a few videos on YouTube here and there related to neural networks and text generation with the hope that even if I cant sit and devote time to learning something new I can at least stay familiar with the material.

19/02/2020

FINALLY!! I have a bit of breathing room now so got time to start properly searching for these research papers. I figured out how to use google scholar and am going through the Engineering and computer science topic as well as the artificial intelligence topic and gathering all the papers that have anything to do with text generation and then to an extent things that have keywords that I extracted about the topic at the start such as GAN’s.

20/02/2020

I plan to spend this morning just gathering papers. I got the github repo setup and downloaded the desktop gui as john suggested. Today in our meeting we talked about other ways of thinking to solve the problem. This included trying to convince people who are higher in the field to help or guide us. I will be visiting Cambridge in a few weeks to visit a friend so taking this thinking into account I am going to try and see if I am able to talk to any of the researchers about my project and see if they could point me in the right direction or give me any help. I also talked to john about the future of my project as currently my project is more researched based than practical. It is more likely I will not create any end product of use and more likely I will produce small amounts of work that could help people who would be able to make more significant results. For me this is a little disappointing as the main drive and joy that I get for projects is mainly based on the end product that I can see myself making, But I also see the value in this as it is a more realistic outcome if I where to take on a researcher role in the future.

24/02/2020

My coursework for csc 1031 has finally been submitted and I have finished my exam for my databases module so I can Finlay put a lot more of my time into this module. Like john mentioned last Thursday as the subject is rather new there is most likely not a survey paper or a goto place to find the most relevant and best papers. So I am going to work on creating a physical map to show where I got papers from and this should visualize the paths to follow to find papers on the subject.

25/02/2020

Today I went through all the top conferenced and got any relevant papers. I had to email someone to request one as it was behind a paywall so hopefully, they can get back to me otherwise I will use the university’s library o try and request it. I also started looking at the authors of the most relevant papers the first paper being : towards controlled generation of text. One of the authors is the PhD student Zhiting Hu. I found his website and he has worked on a number of related appears and even produced software for controlled text generation. I am now going to look into his colleuges and see if they have research in the area and I might email him to see if he can point me in any good directions.

26/02/2020

Today I searched through the other names that have appeared along side Zhiting Hu. None of their work seems to be related in the field of text generation so must have been on the paper for neural network reasons. I am going to email him and ask him if he has any recommendations as of where to research

27/02/2020

In our meeting today we talked about what we need to hand in next Friday. For me I am going to convert my research doc which is a walkthrough of what I have done to this point to learn in this area. I also need to add some code to my GIT repository so I will add the GPT-2 stuff that I did and whatever code I produce from trying out texar.

29/02/2020

Today I woke up to an email back from Zhiting Hu. He recommended trying out the following library’s for implementation: Huggingface transformers, Texar or fairseq. I’ve heard of the first two, one from john and one which was made by Zhiting Hu. The last one I will need to research. Got an email from john today linking to a reddit post, ill need to check that out. Just uploaded all my research papers and my first neural network to my github and invited john as a collaborator. Started using a tool called Mobirise4 to make the blog today.

02/03/2020

Today I have just been working on getting this blog up to date with my current research. The tool I am using morbirise4 is very handy and I have found the best way to use it is to make all of the website functionality and basics in the program and add the rest of the content I need to manually as HTML. The only problem is that if you add something with HTML and then keep working on the application it wipes all the HTML: you did so unfortunately I need to add back all the code in formatted boxes that I have been working on for the past half hour but at least I know now before I did more HTML

04/03/2020

Today I focused on tying all the little pieces I have left to do together; I now have the diary in the 3 style format john wanted but the planning needs more work for the hand in on Friday. I need to ad a few more things to the website for Friday but the main thing I want to do is get some more code into my git repo. I have some code but I don’t think it will be enough so I will have to play around with theses new library’s I’ve been told about and upload any work I can do to the repo

06/03/2020

Today I finished up the planning log and learning log. I also tidied up a few things on the website. I tried to get the models onto the Repo but cant get them separated from their own git code so it wont let me add it. I worked on it for about half an hour and couldn’t get it done. I’m a bit time constrained today as I have to catch a flight soon so am prioritizing important things. Its my own fault for not getting enough done during the week.

09/03/2020

Today I found a video series on NLP on YouTube, I sent the link to Ronan as well as I thought it might also be useful to him. From now until the end of my project I just need to refine what I have found and created to the best of my ability. I plan to work a lot more on code now as apart from the code in the first 4 weeks I have only done a small amount of code. I would like to be able to get some other text generation methds working like setting up texar, have a proper dive into huggingfaces librarys.

11/03/2020

Today I started messing about with texar I got a venv set up fopr running it in. getting it loaded in and ready was easy enough as I understand the basics of seeting up things like that in a venv now. I was looking at the models they have on it and one of them is the same gpt-2 model that I used for making poetry. Maybe a possible approach to furthering my project could be to do a comparison of all the models for text generation there are and see what works in each or what isnt as good etc and propose or even try and implement a combination of the models. Currently I think maybe possibly passing the output samples into another model for training data. This is only a wild guess but in theory if the model works as it should this could be possible however it is more likely this went be able to happen due to the uncertainty of text generation I have experienced this far.