Possible dataset we could use?

* <https://www.kaggle.com/itachi9604/disease-symptom-description-dataset>

Implementations by other people - or related to the articles

* <https://www.kaggle.com/healthcarechatbot>
* <https://www.kaggle.com/programminghut/seq2seq-chatbot-keras-with-attention>

Skeleton code we are using for chatbot:

<https://github.com/python-engineer/pytorch-chatbot>

Pretrained models for text classification:

* <https://www.analyticsvidhya.com/blog/2020/03/6-pretrained-models-text-classification/>
* <https://arxiv.org/pdf/1906.08237v2.pdf>

Notes for meeting Sunday 19th:

1. I read through the articles that I put in the folder “Readings” (also citations are stored in a text file in case we need to include them in the proposal as part of research).
2. From what I gather there should be a way to get data from reddit, that or there should be a kaggle set that has all the healthcare info, i put one above in case thats viable.
3. I put a skeleton code for a chatbot into the “Code” folder
   1. Our NLP preprocessing Pipeline 
      "Is anyone there?" 
      tokenize 
      ["Is", "anyone", "there", "?"] 
      lower + stem 
      ["is", "anyon", "there", 
      exclude punctuation characters 
      [ is, anyon", "there"] 
      bag of words 
   2. Based on the above preprocess pipeline atm - from videos (they gave licence for free, hence why can use it as a skeleton code)
   3. This contains a few files which are linked together
   4. We need to pick a model for this: as far as i’m concerned I think we can use a pretrained one, and use transfer learning to adapt it to our purposes from what Merhtash said
      1. The papers that seemed to be investigating better than simple models - looked like RNN, deep CNN or LSTM are the way to go.
      2. Key thing here was something about an “attention model”
4. The paper on transport systems has some extra stuff about natural language processing that may be relevant to increase the encoding of the model
5. Tkinter is the easiest GUI to map to for the chat function apparently, so we can code that also whenever anyone has time (all the elements are there already)

Come up with “methodology”:

Based on our literature review, we will use Natural Language Processing and Machine Learning to implement a contextual chatbot for healthcare purposes as developed by Kandpal et al. The implementation that we will use as a skeleton for the natural language processing as described by Kandpal et al is located at <https://github.com/python-engineer/pytorch-chatbot>. Kandpal et al uses a neural network with two hidden layers, however we will be using a pretrained XLNet as it is recommended here (<https://www.analyticsvidhya.com/blog/2020/03/6-pretrained-models-text-classification/>)??? Or do we find an implementation of a model????

1. It says dataset shouldn’t need considerable preprocessing… i’m so confused as to what the “methodology” but i think its just describing the neural network and what you do with the output??????????

TL;DR

* I put skeleton code
* What does methodology mean? The model used?
  + Can we use pretrained model - I remember Merhtash mentioning.
* The proposal stuff said don’t focus too much on preprocessing, so maybe keep it basic.
* TKinter can be used to make GUI whenever someone has time