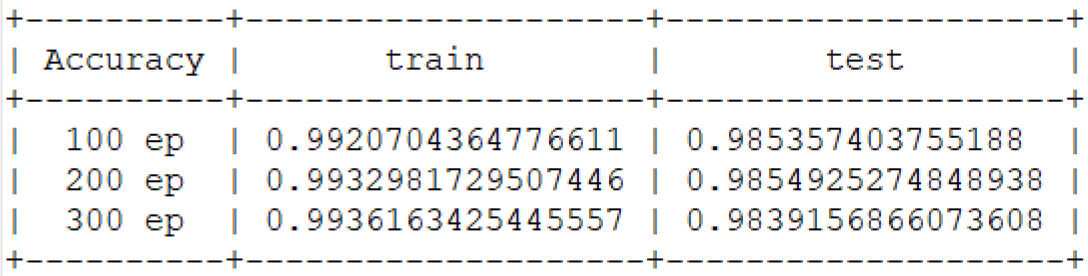
Week 10

Work done

The week has been spent attempting protein entity recognition and pubmed data fetching.

The researcher has created an ad-hoc list of some 200,000 biological entities such as proteins and genes based on the BioCreative II data. This set of organic molecule names intersected with an NLTK set of common 200,000 english words consists the training and dataset int an 80-20 % ratio. We encoded the words as 96 dimensional vectors with scispacy, an independent embedding trained on biomedical corpora.
For our purposes, even a simple mlp produced what could be considered good results: a two layer neural network with dense layers and sigmoid activations produced a 97% accuracy on the test set. We therefore the same architecture in with a 0.1 learning rate for 100, 200 and 300 epochs respectively. The results are as follows.



```
+----------+--------------------+--------------------+
| Accuracy |       train        |        test        |
+----------+--------------------+--------------------+
|  100 ep  | 0.9920704364776611 | 0.985357403755188  |
|  200 ep  | 0.9932981729507446 | 0.9854925274848938 |
|  300 ep  | 0.9936163425445557 | 0.9839156866073608 |
+----------+--------------------+--------------------+
```

Results seem encouraging, until one considers that in a set of 400 000 words 2% means that 8000 words are wrongly classified.

We attempted a rudimentary application by fetching 500 articles from pubmed which feature dystrophin, since there is a list of known interactions.When applied at a practical level, this network cannot differentiate most proteins in a text. Indeed, we obtain better practical results by simply directly comparing with our 200 000 protein list.

Overall, the protein entity recognition is a failure.

Unfortunately, neither approach is fast enough to apply directly on the text, our classification function taking some 10-20 minutes for our 500 query set.

This reaffirms the researcher's original intuition that an independent web crawler is necessary.

Problems

The main problem with entity recognition is the lack of a properly curated protein list. The ad-hoc one improvised from the BioCreative II data is jumpled with genes and other organic molecules, and often contains even simple numbers. A better dataset is needed for more precision.

Once a better protein dataset is found, a limited database can be created for a local database, which can then be used for various website representations.

To do for next week:

Finding a better protein list, and implementation of a rudimentary crawler.