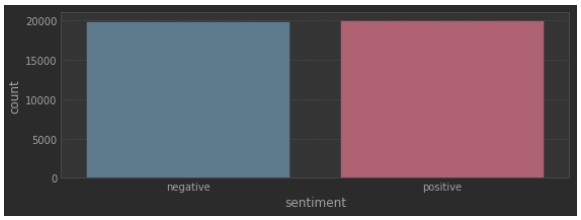
NLP assignment 2

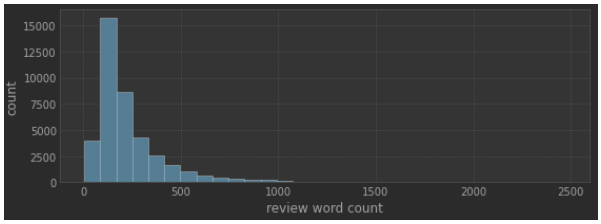
Ram Cohen & Jonathan fuchs

1 EDA

The datasets are IMDB reviews labeled positive/negative which are dispersed equally:



Each review is composed of words, with varying number of word count:



This large variation in word count will affect the feature choosing ahead

2 Word embedding

As required, we used a pre-trained tokenizer for the word embeddings. Due to the large variation in sentence lengths we choose to use a single vector in the size of the embedding for each review. Therefore to embed a lot of information in a single vector we chose a large embedding: the - ”glove-twitter-200” with a 200 size.

Tokenizing has been performed so for each review there is a single vector containing the mean of the embeddings of the entire review. Each word is changed to lower case, stripped of punctuation characters and passed to the tokenizer. The method is a @staticmathod so it can be used by the class for post training inference on new reviews.

```
@staticmethod
def make_review_token(tokenizer, review):
    """
    A static method which tokenizes a single review, it may be accessed by either the class instance or
    as a service for a new external review
    :param tokenizer: the tokenizer which was used to train the reviews
    :param review: a string which represents the review
    :return:  tokenized review as a float vector
    """
    review = [word.strip('.,;!@#$%^&*()/"\'<>~') for word in review.lower().split() if len(word) > 1]
    tokens = []
    for word in review:
        if word not in tokenizer.key_to_index:
            continue
        tokens.append(tokenizer[word])
    feat_array = np.concatenate(tokens, axis=0).reshape(-1, tokenizer.vector_size)
    X_i = feat_array.mean(0)
    return X_i
```

3 Model and hyper-parameters

We used a 4-layer fully connected NN (200-512-1024-512-2) to allow room for a lot of learning. Relu as activation and Cross-entropy as loss. **We chose a 2 class output and not a binary cross-entropy so in the future we can enlarge the model for more calsses.**

Optimizer – Adam, lr=0.001 – we had to choose a small value for the network to actually start learning. Batch size of 128 seems as a good value and 100 epochs are enough.

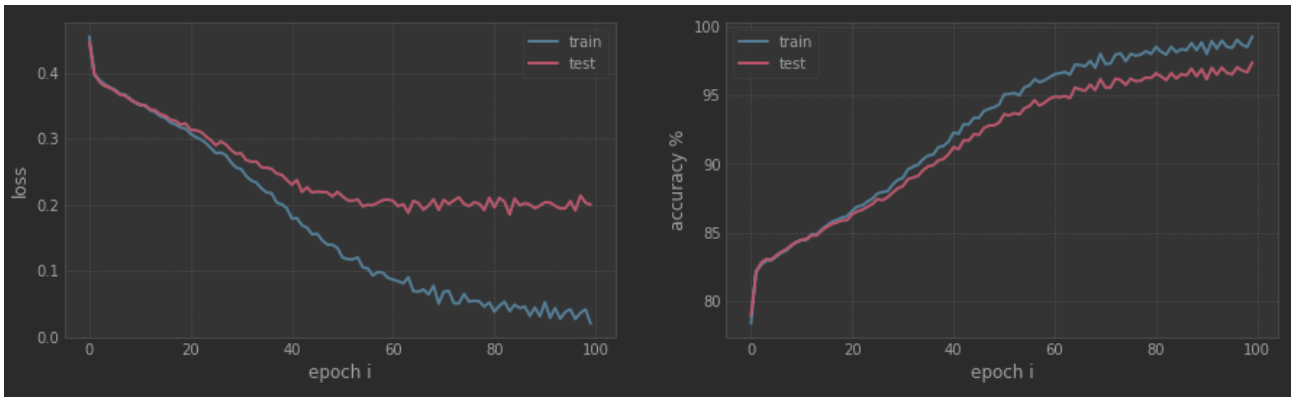
We also built a training function with a checkpoint indicator so the best model will be saved.

4 Training

Using a GPU the training was relatively fast (a couple of minuets), training was steady with loss decreasing and accuracy increasing steadily. A small acceptable overfit was recorded.

```
tokenizing: 40000/40000|==================================================================================================|
tokenizing: 8356/10000|==================================================================================================----------|
```

```
epoch:26/100|loss Train/Test: 0.278/0.290|accuracy Train/Test: 87.84%/87.43% *checkpoint*
epoch:27/100|loss Train/Test: 0.279/0.296|accuracy Train/Test: 87.94%/87.34%
epoch:28/100|loss Train/Test: 0.275/0.291|accuracy Train/Test: 88.01%/87.57% *checkpoint*
epoch:29/100|loss Train/Test: 0.264/0.283|accuracy Train/Test: 88.49%/87.90% *checkpoint*
epoch:30/100|loss Train/Test: 0.257/0.278|accuracy Train/Test: 88.83%/88.21% *checkpoint*
epoch:31/100|loss Train/Test: 0.254/0.278|accuracy Train/Test: 88.98%/88.38% *checkpoint*
epoch:32/100|loss Train/Test: 0.244/0.268|accuracy Train/Test: 89.60%/88.88% *checkpoint*
epoch:33/100|loss Train/Test: 0.237/0.266|accuracy Train/Test: 89.79%/89.00% *checkpoint*
epoch:34/100|loss Train/Test: 0.234/0.266|accuracy Train/Test: 89.93%/89.12% *checkpoint*
epoch:35/100|loss Train/Test: 0.225/0.257|accuracy Train/Test: 90.35%/89.55% *checkpoint*
epoch:36/100|loss Train/Test: 0.219/0.256|accuracy Train/Test: 90.61%/89.82% *checkpoint*
batch:203/313|==============================================================----------------------------------|loss:0.247|accuracy:92.19%
```



5 Final results

Final results (epoch 100/100):

loss Train/Test: 0.020/0.200|accuracy Train/Test: 99.30%/**97.41%**

To decided to test two new reviews (a 2 star and a 10 review), the trained model did well:

```
bad_review = "Great visuals, actors and actresses do their best to make it believable. Benedict Wong's character proves to be a bad-ass. Olsen also
 does her best. Cumberbatch is as usual. No bad things there. Even the cameos can, at some level, sell the plot points. OK! There is one important plot
  hole: at the end of WandaVision, Wanda learns hot to neutralize other people's magic. She does that at the magnitude of a town. In this movie,
 instead of using that same magic and by-pass the Sorcerer Supremes and their wizards, she tortures and kills them. Why? One important issue in
 character development: Wanda learns that her happiness means nothing if it is based on other people's misery at the end of WnadaVision. In this film,
 she has completely forgotten that lesson and at the end she learns it, again! Come on! Writers! Learn to respect these characters! Disney identity
 politics: Wanda kills the 3 male cameos (Picard, Jim from the Office, and some guy) easily. Literally, with her mind. OK. But she fights Carter and
 Captain Marvel, and Captain Marvel dies because of a statue falling on her! The same Captain Marvel, who pierced a hole in Thanos's spaceship, dies
 because of rubble?! And of course, she does not look for a universe where Vision is still alive. If you have noticed, I did not mention Dr. Strange.
 It is because Dr. Strange has NO function or importance in this movie. Cut out his scenes completely, and the plot will not change. In a Dr. Strange
 movie, Dr. Strange does not matter, at all. Edit: other than the VFX scenes, the worst photography and the worst editing in MCU."
```

```
X_bad = ReviewDataset.make_review_token(tokenizer, bad_review)
model(torch.tensor(X_bad)).argmax()
```

```
tensor(0)
```

```
good_review = "Absolutely blew my mind! I went in expecting to be disappointed because of the early reviews and concerns of it being only 2 hours! But
 wow!!! It was beyond what I expected! Coming from someone who is a huge marvel and comics fan along with horror this movie was a piece of art!!! As
 someone who thought spider man no way home was peak, this blows spider man no way home out the park! They didn't hold back at all Sam really did a
 great job putting his expert horror style into this! The cast was incredibly! Gomez as America Chavez is undoubtedly going to be one of the huge faces
  of marvel in the future she did an amazing job and made the character impossible not to like and root for! Olsen of course played the scarlet witch
 as expected making her horrifying in every second of the film! And of course Dr strange showcasing his raw power and his true kindness that not
 everyone sees, he might be one of the most hubris heroes but his genuine concern and protection of America Chavez showed why he is a true hero!!! The
 Illuminati scene was insane and easily showed the raw massive power that the scarlet witch holds. The only only thing I would change would be to make
 it 30 min to 1hour longer! It was so packed with non stop action and story telling I feel like more time would've allowed to give a little more and a
 little less of a bit rushed ending! Regardless it goes down on my list on top 3 marvel movies/comic book movies of all time!!!"
```

```
X_good = ReviewDataset.make_review_token(tokenizer, good_review)
model(torch.tensor(X_good)).argmax()
```

```
tensor(1)
```

```
dataset_train.label_2_idx
{'negative': 0, 'positive': 1}
```