**Recurrent Neural Network**

Zheng Lu (ECE692)

## Dataset: The Count of Monte Cristo

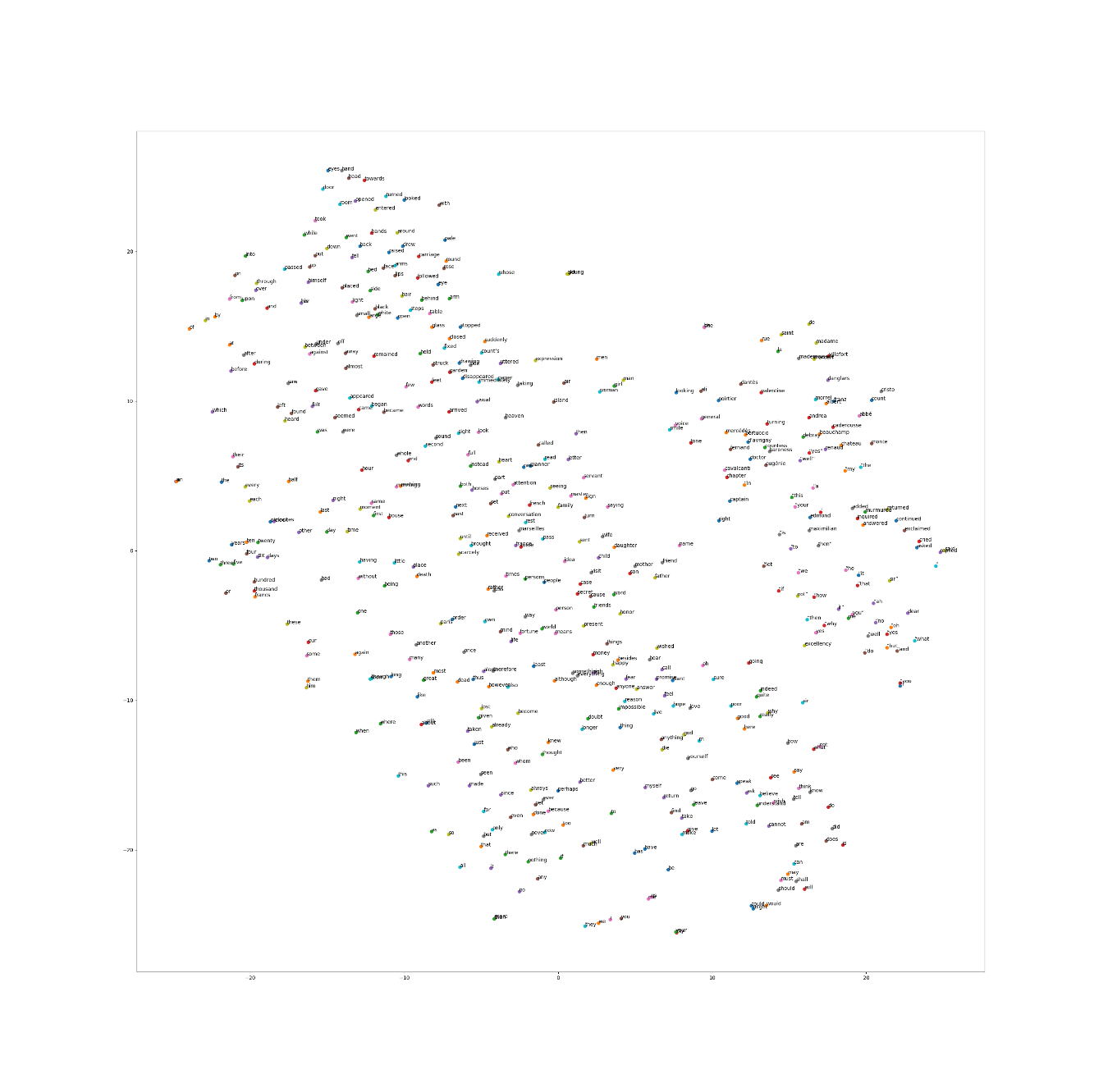
We choose the novel “The Count of Monte Cristo” written by Alexandre Dumas, père.

The novel contains 2,601,150 characters, 21,557 sentences and 468343 words.

There are in total 22,173 unique words.

## Word2Vec visualization

We choose the 500 most frequently used word in the novel and show their 2D word2vec visualization below:



## Samples of the generated text

We run the training for 170,000 iterations, and show the sample of generated text below:

* iteration 40,000 generated sample: colonial pre cious jewels jewels hat miles jewels jewels shirts shirts shirts
* iteration 80,000 generated sample: communicative understudy mcclenahan shrewd shrewd shrewd shrewd expects
* iteration 120,000 generated sample: palmetto bungalow etc etc gestures gestures gestures 30 30 m 30 30 m m 30 m m etc etc
* iteration 170,000 generated sample: educated nurse cious secretary glorious poise poise stare hams jewels chiefly miles miles

## Paper review (ECE692 only)

Both paper studies the text classification problem.

[HAN:2016] try to achieve better performance by using a two-level attention mechanism at both word level and sentence level to learn which individual words and sentences should it pay more attention when constructing the representations of the text. The hierarchy in the text is well defined in this model, the model mainly learns at each layer of the hierarchy which part contributes most towards the next level.

[TextCNN:2016] explores the “benefit of depth” of convolutional neural networks in NLP by using a deep stack of local operations to learn a hierarchical representation of a sentence from character level. The hierarchy in the text is not directly defined in this model. The model learns the hierarchy directly by itself.