Project Proposal

by Sophie Bulman (XXXXXX), Sevag Hanssian (XXXXXX) & François Milot (260997547)

# Hypothesis

*“Is a machine learning algorithm able to generate poems from prose literatures?”*

In other words, are we able to train a model to generate simple quality couplets when everything the model has seen is no-poetic texts? How can we force rhyming scheme in the model so that it produces proper couplets?

We want to explore mainly two aspects in this project:

1. **Clustering of words by their phoneme:** we want to use an unsupervised learning algorithm that would group similar rhyming words near each other. From that algorithm, we would be able to have a pair of word that rhymes with each other. After this task, we would have our ending word for each sentence of our couplet.
2. **Generating sequence of words backward**: Now that we have our sentence-ending words that rhymes with each other, we need to come up with two sentences that are related and that has a correct English structure. First, we would generate by starting with the last word (one of the generated words from the unsupervised technique) and move backward until the sentence is completed. Next, we do a similar process for the first sentence. We start with the ending word; we take into consideration the sentence that was generated and then generate backwardly the first sentence.

From the above combination, we would have a couplet that was generated with an explicit structure that we gave to the model.

# Method

We basically have to train two model:

1. **Unsupervised learning – Clustering of words by their phoneme:** The generated words could be selected either from the corresponding *k-mean cluster* or from the *k-nearest neighbor* of the seeded word. We would need to add some stochasticity in the model to randomize words that are selected (choose randomly from a cluster for instance). A possible improvement would be to filter the possible rhyming word to be coherent in context.
2. **Supervised learning – Generating sequence of words backward:** We generated sentence would be done by *LSTM*, *RNN* and *Transformers*. The main concept to keep in mind is that we are going to learn to create sentence in a backward fashion and over two sentences.

# Datasets

We will be using books from the <https://www.gutenberg.org/> website. All books are free of copyrights. We decided to go with creating our own datasets so that the type of text used during training is what we would anticipate for our couplets. We are aiming to create datasets of around XXXX words from XXXX different books.

# Related Work

* Phoneme unsupervised algorithm: <https://www.researchgate.net/publication/312194885_Phonemes_based_Speech_Word_Segmentation_using_K-Means>
* Word2vec for phoneme: <https://arxiv.org/abs/1912.08011>
* Generating rhyming poetry using LSTM: <https://dspace.library.uvic.ca/bitstream/handle/1828/10801/Peterson_Cole_MSc_2019.pdf?sequence=3&isAllowed=y>
* Automatic Poetry classification: <https://ruor.uottawa.ca/bitstream/10393/37309/1/Kesarwani_Vaibhav_2018_thesis.pdf>
* Generating rhyming sentences: <https://nlp.stanford.edu/courses/cs224n/2013/reports/shotan.pdf>
* Guided learning: <https://arxiv.org/pdf/2006.03626.pdf>
* Reinforcement learning for sequence: <https://arxiv.org/pdf/1510.09202.pdf>

# Possible Proposal Adjustment

* We know that a greater model would be to use reinforcement learning and add rhyming score in the reward function of the model. However, we have limited knowledge of the concept and prefer to focus on the concept that we saw in class.
* We discussed the possibility of having the two rhyming words being related in sense. However, we feel that it would bring complexity to the model while still not reflecting fully what a couplet might be.
* Change the dataset to Gigaword. However, we felt that news article won’t create good poem. If we see that generating our own datasets is too time consuming, we would use those texts.
* Create longer poem if we see great results when generating couplets.