COMP 550 NLP, FALL 2020 FINAL PROJECT PROPOSAL

Sophie Bulman (260738494), Sevag Hanssian (260398537), François Milot (260997547)

Research Question: "Can a machine learning model generate poems from prose?"

We hypothesize that we will be able to train a model to generate simple rhyming couplets without training on any poetry. We want to train the model on prose/novels and introduce features such as phonemes and syllables to enforce rhyming schemes, with the goal of generating rhyming 2-sentence couplets.

Specifically, we plan to explore the following two domains as part of our project:

1) **Unsupervised learning – Clustering of words by their phonemes:** we want to use an unsupervised learning algorithm that would group similar rhyming words near each other, to create rhyming word pairs. 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).
   We could introduce greater complexity by choosing words that are also *related in meaning.*
   A more advanced (but difficult) variant could be to use *reinforcement learning* to add a rhyming score in the reward function – but we may avoid this if we don't study it in class.

2) **Supervised learning – Generating sequence of words backward**: We can set these two rhyming words as the fixed last word of two sentences. We can then create sentences (using *LSTMs*, *RNNs*, or *Transformers*) of the couplet by generating a sequence backwards starting from the final rhyming words.
   Depending on the workload associated with dataset collection and problem 1, we may modify this section to simply use pre-packaged tools to generate sentences.
   We could add more complexity by considering that one sentence is created independently, while the other should depend on the first to create semantic cohesion, in order to create a couplet that makes sense.
   We could also try to generate rhymes longer than couplets (3+ sentences), and introduce different types of rhyme schemes, to see how far our model can go.

For our dataset, we will be using free, public domain classic novels written in the English language from the https://www.gutenberg.org/ website, most likely from a distinct category (e.g. "popular American novels from the 19th century"). We may also resort to using existing datasets (e.g. Gigaword) if generating our own datasets causes us any problems. We plan to introduce pronunciation data by using pre-made tools such as the CMU Pronouncing Dictionary and corresponding functionality available as part of NLTK.

Related works, existing literature:

- 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