Goal: Automatic Text Generation in Bengali (Text type can be News)

Reference Implementation: Automatic Text Generation in Macedonian Using Recurrent Neural Networks

Their process:

* Generated two types of texts: News articles and Poems
* Model trained on a dataset in English, used for Macedonian
* ROUGE-N metric used for performance evaluation

Their steps:

1. Collect data using crawler from websites (based on html tags)
2. Clean the data:
   1. Remove special characters, html tags, Java-script functions, etc.
   2. Remove $, degree, mathematical ops, etc.
   3. Translate English words into ??
   4. Punctuation signs
3. A

My Possible Steps:

1. Collect a lot of articles on a particular topic, suppose, Cricket
2. Take article a1
   1. Remove htms tags, java-script function, special characters, etc.
   2. Replace numbers with <number>
   3. Add ‘space’ before and after punctuations to handle them as separate ‘units’
   4. Find names by comparing to a dictionary and replace names with <name1>, <name2> etc. accordingly
3. Do Step 2 for all the articles
   1. Encode ‘units’ with numbers using a dictionary (or, word embedding), then use One hot encoding for output modeling
   2. Build the LSTM Model

Environment Setup & Tutorial Codes:

1. <https://towardsdatascience.com/installing-keras-tensorflow-using-anaconda-for-machine-learning-44ab28ff39cb>
   1. Install and system prep
2. <https://machinelearningmastery.com/text-generation-lstm-recurrent-neural-networks-python-keras/>
   1. Character generation using lstm rrn
3. <https://www.geeksforgeeks.org/python-word-embedding-using-word2vec/>
   1. Shows the preprocessing and word2vec (bag of words and skipgram) implementation
4. <https://towardsdatascience.com/word-embeddings-exploration-explanation-and-exploitation-with-code-in-python-5dac99d5d795>
5. <https://www.tensorflow.org/tutorials/text/text_generation>
   1. Character wise text generation using GRU RNN