

Auburn University
Assignment 3

COMP 5630/ COMP 6630/ COMP 6630 - D01 (Fall 2024)
Machine Learning

Deadline: Nov 3, 2024, 11:59 PM CST

1 Word Embeddings and N-gram (25 Points)

1. You will examine two-word embeddings. You are given the following words.
Dog
Bark
Tree
Bank
River
Money
 - (a) Use word2vec and compute nxn matrices using cosine similarities for the given words. Use the configuration
sentences=common_texts, vector_size=50, window=5, min_count=1
 - (b) Now use Fasttext Embedding from Genism and compute nxn matrices as question a. Use the following configuration
FastText(vector_size=50, window=5, min_count=1, sentences=common_texts, epochs=10)
 - (c) Which embedding captures better semantics? Justify your answer.

[Link to word2vec](#)

[Link to FastText](#)

2. N-grams and Classification

Download Twitter Sample Data from nltk

Kaggle link to download

- (a) Split the data 70% training and 30% testing.
- (b) Extract n-grams for n in [1, 4]. unigram, bigram, trigram, 4-grams.
- (c) Build a logistic regression model using n-gram features and evaluate your model's performance.
- (d) How does the value of n in n-gram affect the model's performance? Explain your answer. You can draw a plot with n-gram and the model's performance.

2 RNN and Machine Translation (25 Points)

You will be training a Seq2seq model using RNN. Your input will be a text and the output will be a summary of the text.

1. Load the California State bill subset of the BillSum dataset from HuggingFace. Load the test split as your entire dataset for this task. Split the dataset into a train and test set with the `train_test_split` method as done in the Hugging Face .
`billsum = load_dataset("billsum", split="ca_test")`
2. Use the number of neurons, dropout, and your selection of RNN architecture. Report BLEU as the model's performance.
3. Vary the input seq length by truncating the main text at 1024, 2048 and the summary text as 128, 256. How does the sequence length impact the model's performance?
4. Try different hyperparameters to obtain the best accuracy on the test set. What is your best performance and what were the hyperparameters?

Link of the dataset from Hugging Face .
An example of seq2seq in Keras