SGN-24007: Advanced Audio Signal Processing

Exercise 4

05.04.2019

In this exercises we will attempt to design a RNN based model which can learn language model from a text corpus and then can generate some text using that model. This exercise is directly related to the topics covered on Thursday lecture of this week.

1 Framing Language Model

A statistical language model learns from the raw text, prior probability of a word sequence. Language models allow reducing the search space and ambiguity and enables use of contextual information. Language models are a key component in larger models for challenging natural language processing problems, like machine translation and speech recognition. They can also be developed as standalone models and used for generating new sequences that have the same statistical properties as the source text.

2 Text Corpus

Text corpus is the raw text from which language model learns its prior probabilities about word sequences. In the exercise I have used couple of paragraphs from Bram Stoker's famous novel Dracula. You are free to use another large text or a text corpus.

3 RNN Model

Vocabulary of a corpus means the distinct words that are present in the corpus. We train our RNN model to predict a probability distribution across all words in the vocabulary.

Please see further instructions on the supplied python script.