Patanjali B

2021114014

REPORT:

Tokenizer:

My tokenizer basically splits the entire corpus into tokens, which it should but also has features like replacing with URLHERE, MENTIONHERE etc. For urls and mentions.

It also has features like

Replacing don’t with don’t

Won’t with wont like features.

Not only this it also converts Mr. To mr, mrs. To mrs etc.

Models:

Due to a crash in the system, I could not bring out average perplexities on the training data, but because of the shortened data in the training set, I could calculate the average perplexities.

Kneser_Ney:

a. Pride and Prejudice: Over a sequence of 1004 odd sentences,

Average perplexity was about 2014.4354896548266

b. Ulysses: Over a sequence of 2034 sentences,

Average Perplexity was 1816.3256578675645

Neural_Language_Model:

a. Ulysses: over a sequence of 456 sentences, the average perplexity is around 2329.5504752788784