Day 10

Took a look at the various corpora that I can use to test my algorithm on

* Gutenberg
* Brown
* Web and Text Chat
* Reuters
* Inaugural address corpus.

I also detailed an algorithm sketchup for what I think will be the fluff cutter function

Here’s the pseudocode (as of June 1st)

function fluff\_cutter(text):

highlighted 🡸 []

#eliminate stop words.

recurrent\_words 🡸 re.findall(r’ ‘) //all repeated

for word in text:

repeat\_stemmed 🡸 stemmer.stem(word)

#if stem equal for more than two words, and the distance between them is small, add them to the highlighted list

Return highlighted