SPECS --- CONFIDENTIAL

The POC consists of a core NLP system that can be interfaced through a simple website:

* The core system is built on Python/NLTK.
* The website interface is built on Eleventy.
* Delivered as operational on AWS.

FUNCTIONALITY:

1 - Upload/Download & Storage of Documents: Through the website interface, the system can take in and save text documents; the documents are accessible through the website for download.

2 - Creation & Storage of Corpora: Through the website, the system can be directred to concatenate multiple designated documents into a larger document (a corpus). Corpora are stored and are accessible through the w 3 - Document/Corpus Manipulation: Through the website, the system can be directred to process designated docuemnts/corpora through various NLTK tools -- such as, for example, removal/insertion of characters/text.
ebsite.

4 - Statistical Analysis and Output: The system can be directred to perform various descriptive stats on designated documents/corpora, and create and store the outputs. For example, it can be directed to identify single words and/or n-word sequences, count their respective occurrences, and output various frequency lists: how many times the word "elephant" may come up in a document or corpus; how many times the 2-word sequence "elephant in" comes up; how many times the 4-word sequence "elephant in the zoo"; output and store a list of the 10 most-frequent 5-word sequences in a designated corpus, etc.

5 - Search and Comparison and Output: The system can be directed to identify the occurence/non-occurrence or n-word sequences in designated documents/corpora, compare documents/corpora/lists to each other, and output and store the results. For example, the system can identify, for an n-word sequence in one list, whether that sequence is or is not present in one other list or designated document or corpus; if the sequence is present, the system can juxtapose the frequency of its ocurrence in the respective lists/documents/corpora; the system can store the outputs of such searches and comparisons.

6 - Action Chaining: The system can link together various functions that can then be triggered in one chain. For example, the system can be directed to create a compound function that

* first concatenates designated documents into a first corpus,

* then identifies n-word sequences in this corpus,
* then outputs a list thereof,
* then conctatenates other designated documents into a second corpus,
* then searches the second corpus for the presence/absence of each n-word sequence from the first output list,
* then outputs the final outcome.