CSE4022 Natural Language Processing

Digital Assignment -1

1. Utilize Python NLTK (Natural Language Tool Kit) Platform and do the following. Install relevant Packages and Libraries

   (03 Marks)

   - Explore Brown Corpus and find the size, tokens, categories,
   - Find the size of word tokens?
   - Find the size of word types?
   - Find the size of the category "government"
   - List the most frequent tokens
   - Count the number of sentences

2. Explore the corpora available in NLTK (any two)   **(02 Marks)**
   - Raw corpus
   - POS tagged
   - Parsed
   - Multilingual aligned
   - Spoken language
   - Semantic tagged

3. Create a text corpus with a minimum of 200 words (unique content). Implement the following text processing

   (05 Marks)

   - Word segmentation
   - Sentence segmentation
   - Convert to Lowercase
   - Stop words removal
   - Stemming
   - Lemmatization
   - Part of speech tagger

   Create GitHub account and upload this.