**NLP Pipeline:**

**Clean Corpus:**

Remove problematic elements that will disrupt/confound the NLP process

Make Lower case

Remove accented characters

Expand contradictions

Remove special characters

**Sentence segmentation**:

Break each Article/Section (and the paragraphs that make it up) into sentences

**Tokenization:**

Sentence tokenization into words

Remove punction & numbers

**Text Lemmatization:**

Lemmatize words so that we can see base words

Is this better than stemming? Or should I do both?

**ID & remove stop words:**

Remove filler words using a stop word list

**PoS Prediction:**

Predict part of speech for each token

**Constituency parsing:**

Break each sentence into sub-phrases

Will this already be somewhat accomplished by the removal of stop word?

Should I also do dependency parsing?

Text summarization

Topic segmentation

Relationship extraction

Bag of Word model?