Assignment 1

Task 1:

Let's consider detection of plagiarism.

Strategies

a. Tokenization - Can be used to find boundaries in sentences which can detect parts of sentences or phrases. This can be used to find common phrases or parts of sentences to determine if the document is plagiarized.
b. Part-of-Speech tagging - Can be used for detecting cases in which words have been replaced but have similar style in terms of grammar.
c. Stemming and Lemmatization - Used to transform words to their stem and dictionary base forms for efficient comparison
d. Finding similarity between N-grams or bag-of-words using similarity algorithms like Jaccard Similarity or cosine similarity
e. Constituency Parse trees - can be compared to detect similar sentence/paragraph constructs

References -

1. https://en.wikipedia.org/wiki/Plagiarism_detection
2. Using Natural Language Processing for Automatic Detection of Plagiarism.

Task 2:

Part-of-Speech:

Token	CoreNLP	Spacy
John, Susan, Europe	NNP	PROPN
met, told	VBD	VERB
In, that	IN	ADP
the	DET	DET
mall, week	NN	NOUN
She, him	PRP	PRON
is	VBZ	VERB
travelling	VBG	VERB
to	TO	ADP
next	IN	ADJ
.	.	PUNCT

The main difference here is that CoreNLP finds the tense of the verb and Point of View (1st, 2nd or 3rd person). Eg VBD for Verb, past tense, VBZ for 3rd person singular present and VBG for gerund or present participle. Other difference 'to' has its own different tag in CoreNLP.

Lemmas:

Token	CoreNLP	Spacy
John	John	john
met	meet	meet
Susan	Susan	susan
in	in	in
the	the	the
mall	mall	mall
She	she	-PRON-
told	tell	tell
him	he	-PRON-
that	that	that
she	she	-PRON-
is	be	be
travelling	travel	travel
to	to	to
Europe	Europe	europe
next	next	next
week	week	week
.	.	.

Differences:

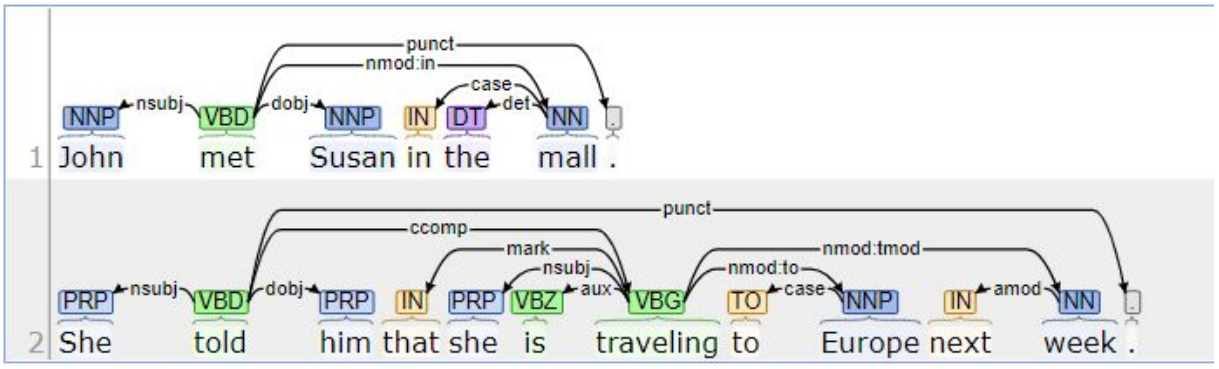
1. Proper Nouns are capitalized in CoreNLP
2. She and he are surprisingly not lemmatized in Spacy as output is given as -PRON-.

Named Entity Recognition:

Token	CoreNLP	Spacy
John, Susan	PERSON	PERSON
Europe	LOCATION	LOC
next week	DATE-2019-W09	DATE

The difference is that CoreNLP points to us the date like Week 9 instead of just marking it as a date tag.

Parse Tree:
Output of CoreNLP:



Output of Spacy:
#For Navigating Parse tree
John nsubj met VERB []
met ROOT met VERB [John, Susan, in, .]
Susan dobj met VERB []
in prep met VERB [mall]
the det mall NOUN []
mall pobj in ADP [the]
. punct met VERB []
She nsubj told VERB []
told ROOT told VERB [She, him, traveling, .]
him dobj told VERB []
that mark traveling VERB []
she nsubj traveling VERB []
is aux traveling VERB []
traveling ccomp told VERB [that, she, is, to, week]
to prep traveling VERB [Europe]
Europe pobj to ADP []
next amod week NOUN []
week npadvmod traveling VERB [next]
. punct told VERB []

Spacy is better in certain situations than CoreNLP as it can detect that traveling is connected to Europe and week. Whereas, CoreNLP just connects it to week.

Task 3:
Vectorized Parts-of-Speech which can be used for detecting cases in which words have been replaced but style in terms of grammar has not been altered.

Task 4:

FastText:

Used the supervised model of FastText. Parameters - Epoch - 50, ngrams -3. Precision and Recall is around 56% as training data is small.

Spacy:

Tokenized the sentences using spacy. Took the lemma of each token and converted it to lowercase. Then vectorized those tokens using Scikit-Learn's CountVectorizer. Used Scikit-Learn's SVM as the classifier for prediction.

Accuracy = 87.5%