Assignment 3: Part of Speech Tagging
Report
1. Which is the best machine learning algorithm (classifier) for this task (for both the baseline and the improved classifier)? You need to discuss this per metric used to compute the performance.
- For the baseline classifier, the Naive Bayes algorithm had the higher performance at 0.312 for all 3 metrics. For the improved classifier, the Decision Tree algorithm had the higher performance at 0.366 for all 3 metrics. The Naive Bayes algorithm actually decreased in performance to 0.241 for all 3 metrics for the improved classifier (Table 1).

2. Which attributes contributed the most to each of the performance metrics for your improved model? Which contributed the least? (Write about this for each algorithm considered.)
- Naive Bayes algorithm: Attribute 4 seemed to contribute the most for this algorithm, and attributes 5, 6, and 7 seemed to contribute the least to the algorithm.
- Decision Tree algorithm: Attribute 4 seemed to contribute the most for this algorithm, and attributes 2 and 6 seemed to contribute least to the algorithm.
- This makes sense as attribute 4 is the target token in the context window and thus contributes the most to the model. It is interesting to see which other attributes contribute more/less to the model for each algorithm, and generally it seems that the further attributes contribute less to the algorithm. It's also interesting to note that for the Naive Bayes algorithm, attribute 5 (the token following the target token) contributes slightly less to the model, whereas for the Decision Tree algorithm, the following attribute contributes more to the model (Table 2).

3. How good is your feature set for this task (for each algorithm)? (Base your response to this question to your answer from Question 2)
- For the improved classifier, the feature set seems to be better suited for the Decision Tree algorithm. For the overall model, performance actually decreased for the Naive Bayes algorithm when compared to the baseline system. When assessing attribute performance, less features seemed to negatively contribute to the model for the Decision Tree algorithm compared to the Naive Bayes algorithm.

4. If you had more time to work on this problem and do it more efficiently (in terms of performance), which features/text representation would you choose? Write 1-2 short paragraphs about the feature sets you might want to try for this problem and why.
- My feature set is not particularly good; the highest performance I reached was an F1 score, precision, and recall score of 0.366 (improved classifier, Decision Tree algorithm) which is not particularly impressive.
- If I had more time, I would play around with the size of my context window. For the Decision Tree algorithm, the closely-surrounding tokens seems most important to consider.
- I would also consider breaking the word forms into their respective prefixes/suffixes and lemma, because this morphological information can better inform a word's part of speech.

Extra Credit 1

Which pairing yielded the best metrics? Why do you think that is?

- The orth/part of speech pair with the Decision Tree algorithm yielded the best metrics, followed by the lemma/part of speech pair. I believe this is because the full ctag requires the classifier to classify a significantly larger set of labels; there are 896 unique full ctags and 35 unique part of speech tags.

	Baseline System		Improved System	
	Naive Bayes	Decision Tree	Naive Bayes	Decision Tree
F1 Score	0.31207469340875127	0.30504542778690313	0.24172852008743873	0.36643571978444955
Precision	0.31207469340875127	0.30504542778690313	0.24172852008743873	0.36643571978444955
Recall	0.31207469340875127	0.30504542778690313	0.24172852008743873	0.36643571978444955

Table 1. Performance metrics for baseline system vs improved system

	Attribute Removed						
	a1	a2	a3	a4 (target word)	a5	a6	a7
Naive Bayes	0.0037544305986093873	-0.00281993964917 4791	-0.0095466 278050165 83	0.0271002375 33602578	-0.01356041 0517345689	-0.0153470 584118594 04	-0.0133216419 04714854
Decision Tree	0.0017619476932070333	-0.0040343 6621341 7961	-0.0062450 342303623 23	0.1639270026 717384	0.05857241 0698480515	-0.0041661 006203867 48	-0.0005063541 267861793

Table 2. Difference in performance for each model with one removed feature using the improved classifier. Note: F1, Precision, and Recall metrics are the same, so only one value is reported.

	Model			
Pairing	*Orth / POS* tag	*Orth / ctag*	*Lemma / POS* tag	*Lemma / ctag*
Naive Bayes	0.24172852008743873	0.16628998859674 04	0.2727272727272727	0.16782963447818766
Decision Tree	0.36643571978444955	0.17127119586024 625	0.3104609469233841	0.09685360602355578

Table 3. Extra Credit 1; Model performance for each predictor/label pair. Note: F1, Precision, and Recall metrics are the same value, so only one value is reported.