CS 6320.002: Natural Language Processing
Fall 2019
Homework 3

Xiaojie Zhu
XXZ180012

Writeup Question 1.1: Notice that we are using the Universal Dependencies tagset instead of Penn Treebank (17 tags instead of 45). What are some advantages and disadvantages of using a smaller tagset? Give at least one advantage and one disadvantage.

Answer:

Advantages:

The universal dependencies tagset contains significantly less tags, and thus the processing would be more efficient, and the recall rate will increase.

Disadvantages:

Because now we are using a smaller tagset, most words that could be distinguished by different tags now are labeled with one tag, thus, the accuracy of the model will decrease.

Writeup Question 2.1: The two features $'\boldsymbol{prevskip}-[\boldsymbol{x}]'$ and $'[\boldsymbol{nextskip}-[\boldsymbol{x}]'$ are skip-bigrams, ie. a bigram where the two words, w_i and w_{i-1} are not consecutive. Why do we need skip-bigram features when w_2 is already captured by the word trigram features? What is the advantage of a skip-bigram over a trigram?

Answer:

1. We need skip-bigram features because the skip trigram model is able to cover more predecessor/successor words of the present word compared to the normal trigram while the same memory space is required.

2. The advantage of a skip-bigram is that its components (typically words) need not be consecutive in the text under consideration, but may leave gaps that are skipped over. It provides one way of overcoming the data sparsity problem found with conventional n-gram analysis.

Writeup Question 2.2: What is the goal of this last set of features? (Hint: why we using rare words?)

Answer:

The last set of features is to see if we can extract some useful information from the rare words. For a single word, if it is in the rare word set, because of its low occurrence, its features may all be included in the rare features set after we do *remove_rare_features* operations, however, if we can check if its prefix or suffix combinations, it may not be rare feature, and thus could provide some other useful information for the model to increase its accuracy.

Writeup Question 3.1: Why do we want to remove rare features? Give at least two reasons why we do this.

Answer:

1. The rare features are features that occurs only a small times, which cannot cause significance to our model's training and prediction. Thus, they are considered unnecessary, irrelevant, or redundant features from the dataset, and should be removed, as they will not help in improving the accuracy of the model, and may in fact lower the accuracy.

2. By removing the rare features, we reduce the size of the feature volume, and hence can increase our precision, we can also increase the efficiency of the program.

Writeup Question 3.2: Why do we want to use a sparse matrix for X train? What is the advantage of a sparse matrix over a dense (normal) matrix?

Answer:

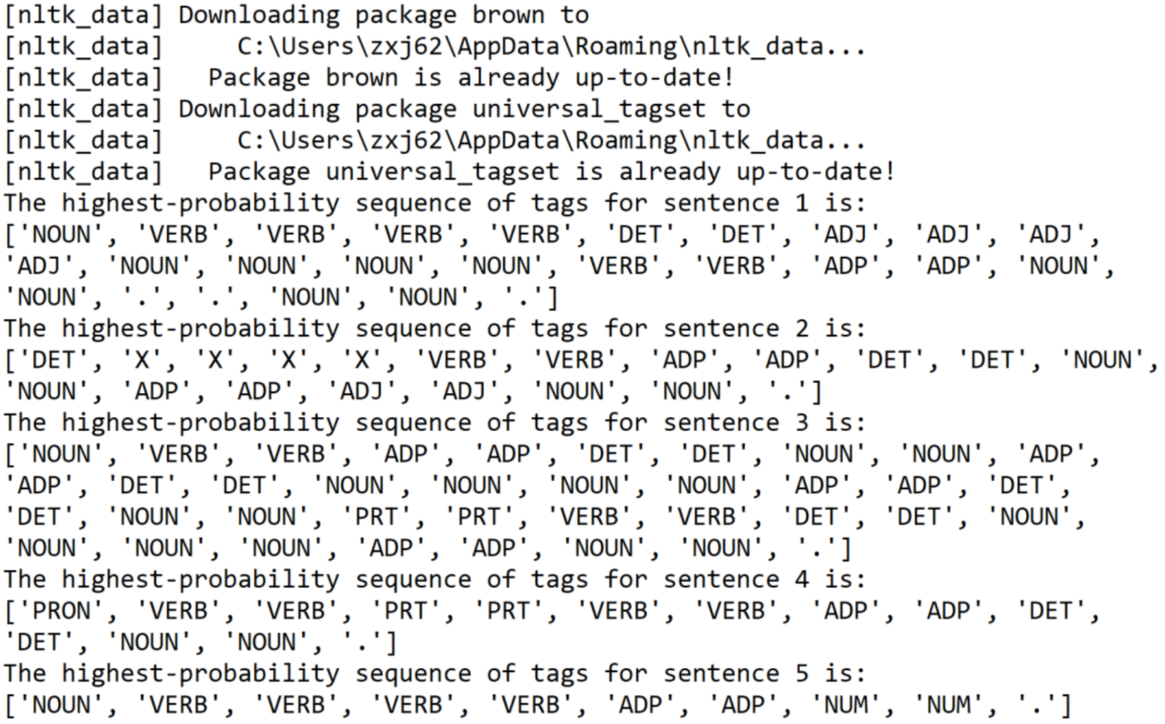
1. Since we are training the model with a very big dataset, the size of feature volume and sentence number are huge. However, for each word in the sentence, only a few of the features were hit, if we use a regular matrix, it would be very time consuming to look up values, and it would also waste a lot space to store the matrix. If we use a sparse matrix, however, it will save a lot space in memory, and the queries would be very efficient too.

2. Advantage of a sparse matrix: a sparse matrix is much less expensive both in time and in space to build and store since it only needs to store the entries with the value 1. Also, because of its reduced storage size, the algorithms working on it would be much more efficient than on a regular matrix.

Writeup Question 4.2: Put your predicted tag sequence for each of the test sentences.

Answer:

The tag sequence for each sentence is below:



```
[nltk_data] Downloading package brown to
[nltk_data]     C:\Users\zxj62\AppData\Roaming\nltk_data...
[nltk_data]   Package brown is already up-to-date!
[nltk_data] Downloading package universal_tagset to
[nltk_data]     C:\Users\zxj62\AppData\Roaming\nltk_data...
[nltk_data]   Package universal_tagset is already up-to-date!
The highest-probability sequence of tags for sentence 1 is:
['NOUN', 'VERB', 'VERB', 'VERB', 'VERB', 'DET', 'DET', 'ADJ', 'ADJ', 'ADJ',
'ADJ', 'NOUN', 'NOUN', 'NOUN', 'NOUN', 'VERB', 'VERB', 'ADP', 'ADP', 'NOUN',
'NOUN', '.', '.', 'NOUN', 'NOUN', '.']
The highest-probability sequence of tags for sentence 2 is:
['DET', 'X', 'X', 'X', 'X', 'VERB', 'VERB', 'ADP', 'ADP', 'DET', 'DET', 'NOUN',
'NOUN', 'ADP', 'ADP', 'ADJ', 'ADJ', 'NOUN', 'NOUN', '.']
The highest-probability sequence of tags for sentence 3 is:
['NOUN', 'VERB', 'VERB', 'ADP', 'ADP', 'DET', 'DET', 'NOUN', 'NOUN', 'ADP',
'ADP', 'DET', 'DET', 'NOUN', 'NOUN', 'NOUN', 'NOUN', 'ADP', 'ADP', 'DET',
'DET', 'NOUN', 'NOUN', 'PRT', 'PRT', 'VERB', 'VERB', 'DET', 'DET', 'NOUN',
'NOUN', 'NOUN', 'NOUN', 'ADP', 'ADP', 'NOUN', 'NOUN', '.']
The highest-probability sequence of tags for sentence 4 is:
['PRON', 'VERB', 'VERB', 'PRT', 'PRT', 'VERB', 'VERB', 'ADP', 'ADP', 'DET',
'DET', 'NOUN', 'NOUN', '.']
The highest-probability sequence of tags for sentence 5 is:
['NOUN', 'VERB', 'VERB', 'VERB', 'VERB', 'ADP', 'ADP', 'NUM', 'NUM', '.']
```

Writeup Question 5.1: How long did this homework take you to complete (not counting extra credit)?

Answer:

The homework took me 3 days to complete, and another 2 hours to run the training set. The total programming time is estimated to be 5 hours.

Writeup Question 5.2: Did you discuss this homework with anyone?

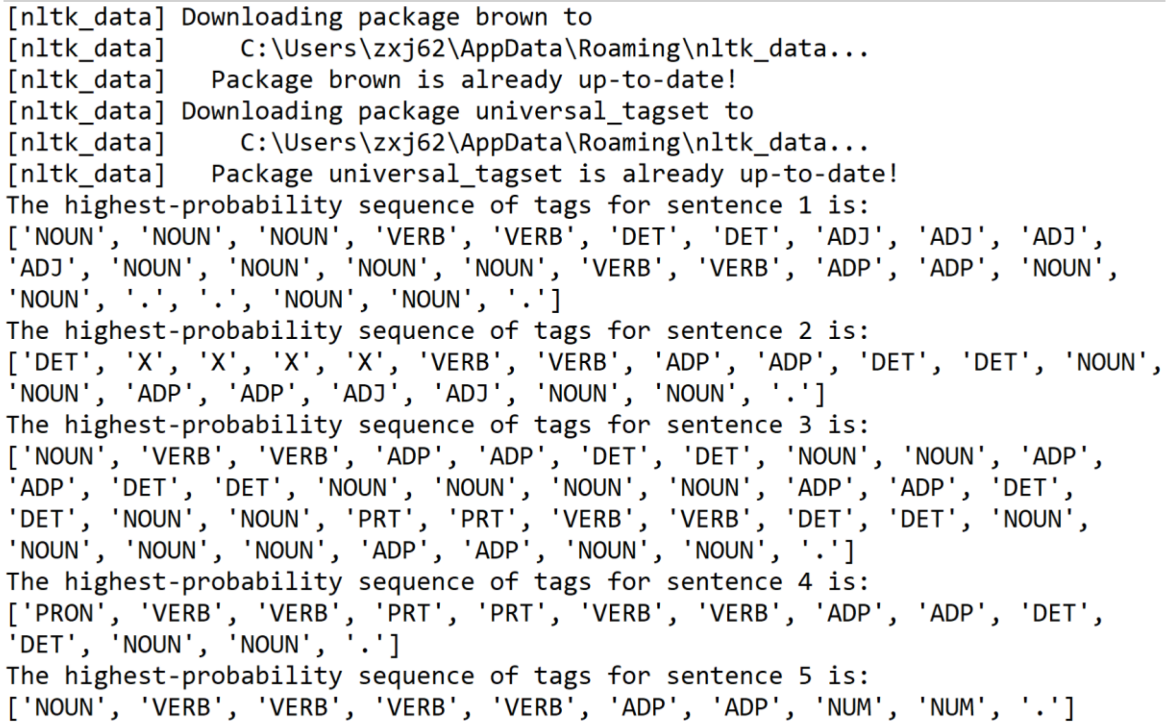
Answer:

I did not discuss this homework with anyone.

Writeup Question 6.1: Put your new predicted tag sequence for each of the test sentences. Are they better or worse than the predictions from Part 4, and if so, in what way?

Answer:

The new predicted tag sequence for the test sentence is below:



```
[nltk_data] Downloading package brown to
[nltk_data]     C:\Users\zxj62\AppData\Roaming\nltk_data...
[nltk_data]   Package brown is already up-to-date!
[nltk_data] Downloading package universal_tagset to
[nltk_data]     C:\Users\zxj62\AppData\Roaming\nltk_data...
[nltk_data]   Package universal_tagset is already up-to-date!
The highest-probability sequence of tags for sentence 1 is:
['NOUN', 'NOUN', 'NOUN', 'VERB', 'VERB', 'DET', 'DET', 'ADJ', 'ADJ', 'ADJ',
 'ADJ', 'NOUN', 'NOUN', 'NOUN', 'NOUN', 'VERB', 'VERB', 'ADP', 'ADP', 'NOUN',
 'NOUN', '.', '.', 'NOUN', 'NOUN', '.']
The highest-probability sequence of tags for sentence 2 is:
['DET', 'X', 'X', 'X', 'X', 'VERB', 'VERB', 'ADP', 'ADP', 'DET', 'DET', 'NOUN',
 'NOUN', 'ADP', 'ADP', 'ADJ', 'ADJ', 'NOUN', 'NOUN', '.']
The highest-probability sequence of tags for sentence 3 is:
['NOUN', 'VERB', 'VERB', 'ADP', 'ADP', 'DET', 'DET', 'NOUN', 'NOUN', 'ADP',
 'ADP', 'DET', 'DET', 'NOUN', 'NOUN', 'NOUN', 'NOUN', 'ADP', 'ADP', 'DET',
 'DET', 'NOUN', 'NOUN', 'PRT', 'PRT', 'VERB', 'VERB', 'DET', 'DET', 'NOUN',
 'NOUN', 'NOUN', 'NOUN', 'ADP', 'ADP', 'NOUN', 'NOUN', '.']
The highest-probability sequence of tags for sentence 4 is:
['PRON', 'VERB', 'VERB', 'PRT', 'PRT', 'VERB', 'VERB', 'ADP', 'ADP', 'DET',
 'DET', 'NOUN', 'NOUN', '.']
The highest-probability sequence of tags for sentence 5 is:
['NOUN', 'VERB', 'VERB', 'VERB', 'VERB', 'ADP', 'ADP', 'NUM', 'NUM', '.']
```

From the results we can see that the new model predicts the sentence with relative similar sequence of tag as the previous model, except for the first sentence, where the 2nd and 3rd verbs were replaced with nouns. This could be partly due to the model is already well-trained with large feed of data. And the reason why the first sentence has changes is that now the modified model can recognize **“Apple Inc.”** as a company name, thus, the newly modified model will be able to increase its accuracy on this new pattern.