Assignment 7
CSE 517: Natural Language Processing - University of Washington
Winter 2022

Solution 1.
In this problem I started by downloading the novel Alice in Wonderland from the NLTK library. I used the whitespace tokenizer method to extract tokens from sentences that did not contain whitespaces, newlines, or tabs. After that, I assigned one to the last character of each token in the training and test sets and 0 to the rest of the characters. Below is the code snippet for the training set:

```
train_data_y = []
for word in train_data:
    y = ""
    for i in range (len(word)-1):
        y = y+'0'
    y = y+'1'
    train_data_y.append(y)
```

The next step was to generate the input features by taking five characters left and five characters on the right from the center character so, for a total of eleven characters. Code snippet below for reference:

```
features = []
output = []
for i in range(len(x)):
    left = 0
    if i-5 > 0:
        left = i-5
    right = len(x)-1
    if i+5 < len(x):
        right = i+5
    xi = []
    for j in range(left, right+1):
        xi.append(x[j])
    features.append(xi)
    output.append(y[i])
```

Then, on the input feature, I utilized one hot encoding to transform it to numeric values. Following this stage, the single hot encoded data is divided into training and test sets in an 80:20 ratio.

To predict the label, I trained a Logistic regression classifier from the sklearn library. On the test set, my model achieved 89 percent per-character segmentation accuracy.

Below is a sample of segmented text from the test set:

```
To Alice's great surprise, the Duchess's armthat waslinked in t ohersbeg a
n totremble. Alice looked upand theres tood the Queeninfron tof them, wit
hhe rarms folded, frowninglike athunders torm! "Now, I g ive y ou f air wa
rning," shouted the Queen, stamping on the ground as she s poke, "eitheryo
```

```
uor y ourhead must beoff, and that ina bouthal f notime. Take y our choice
!"
```

To compute the accuracy, precision, recall, and F1 score, I used the NumPy library. Below is the code snippet to calculate the accuracy, precision, recall, and F1 score:

```
actual = test['y']
predicted = y_pred

tp = np.sum((actual=='1') & (predicted=='1'))
fp = np.sum((actual!='1') & (predicted=='1'))
tn = np.sum((actual=='0') & (predicted=='0'))
fn = np.sum((actual!='0') & (predicted=='0'))

accuracy = (tp+tn)/(tp+tn+fp+fn)
precision = tp/(tp+fp)
recall = tp/(tp+fn)
f1 = 2 * (precision * recall) / (precision + recall)
```

The performance metrics for the Logistic Regression Classifier are listed below in Table 1.

Table 1. Performance Metrics

Model	Accuracy	Precision	Recall	F1 score
Logistic Regression Classifier	0.896	0.801	0.703	0.749

Note: Solution 3 Starts from the next page using paper and pen.

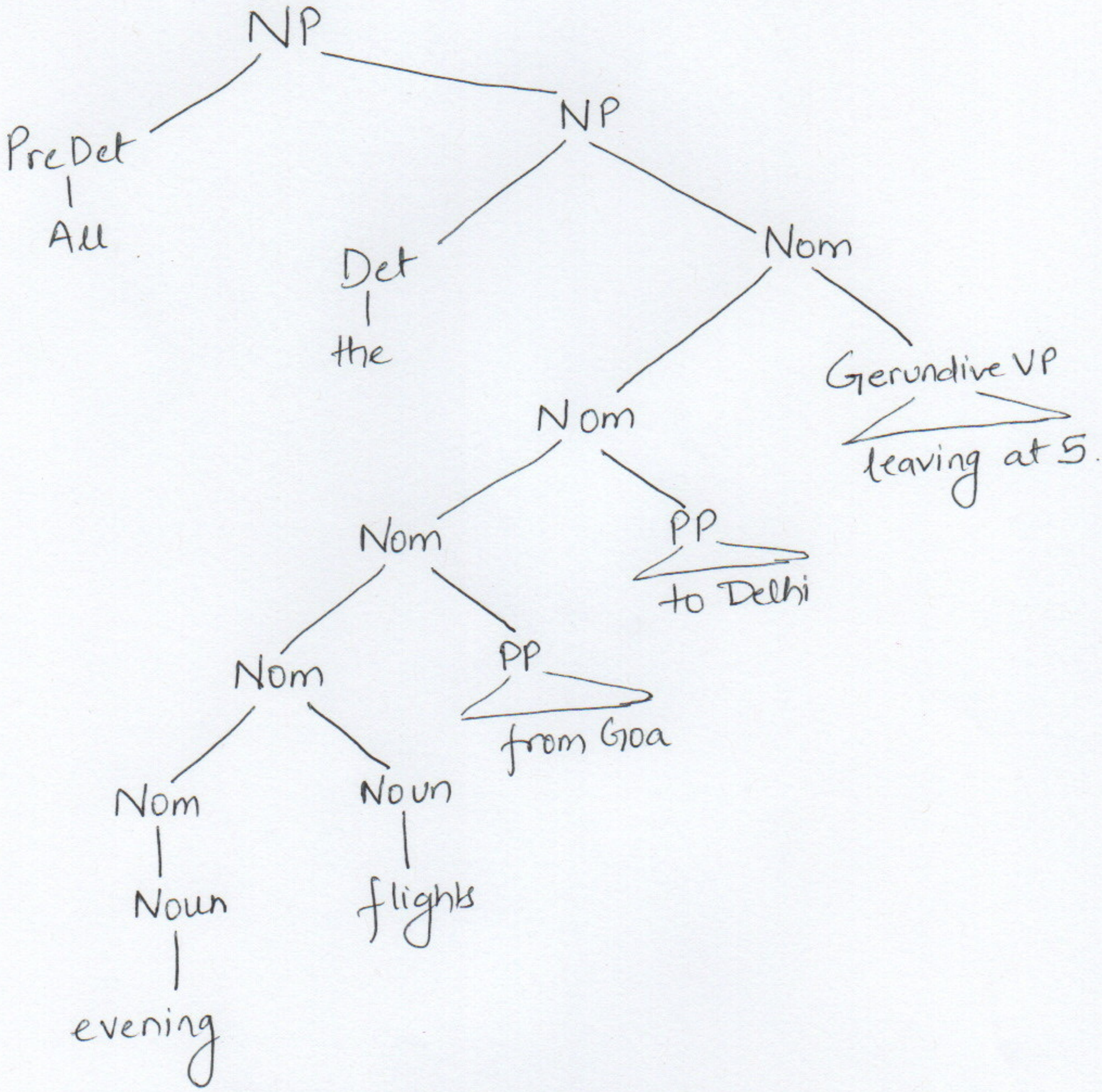
References:

[1] Speech and Language Processing. Daniel Jurafsky & James H. Martin. Copyright © 2021. All rights reserved. Draft of December 29, 2021.
[2] Murphy, Kevin P. Machine learning: a probabilistic perspective. MIT press, 2012.
[3] Link: https://assamite.github.io/cc-mas-course/parsing_NLTK.html
[4] McKinney, Wes. "Pandas, python data analysis library." URL http://pandas. pydata. org (2015).
[5] Hellmann, Doug. The Python standard library by example. Upper Saddle River, USA: Addison-Wesley, 2011.
[6]https://scikitlearn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html?highlight=logisticregression#sklearn.linear_model.LogisticRegression
[7] Ed discussion board for A1 Assignment.

Solution 3.

1) Noun phrase

example 1: All the evening flights from Goa to Delhi leaving at 5.



2) Verb Phrases

example2:: He has been hiking in the mountain.

VP ⟶ VBG NP.

3) Sentences

example 3: He likes an evening walk