W2 Assignment – NLTK

Computational Linguistics

John Gamboa

April 22, 2022

1 NLTK basics

1.1 Text objects

Consider the following piece of code:

```
 1  import nltk
 2  import nltk.book
 3
 4  t = nltk.book.text7
 5  print(t)
 6  type(t)
 7  len(t)
 8
 9  set(t)
10  len(set(t))
11
12  len(set(t)) / len(t)
13
14  t.concordance('walk')
15  t.index('walk')
```

Answer *TRUE* or *FALSE*

a) Line 5 will output the value

```
<Text: Wall Street Journal>
```

indicating that `text7` was data collected from the Wall Street Journal.

answer: ____________________________

b) Line 6 will output the value

```
nltk.text.Text
```

indicating that `t` is of the type `Text`.

answer: ____________________________

c) Line 7 will output the value

```
100676
```

indicating that `t` contains 100676 characters.

answer: ____________________________

d) Line 5 will output the value

```
<Text: Moby Dick>
```

indicating that `text7` is the Moby Dick book.

answer: ____________________________

e) Line 6 will output the value

```
list
```

indicating that `t` is of the type `list`

answer: ____________________________

f) Line 9 will output a data structure of the type `set` containing all words in `t`.

answer: ____________________________

g) Line 7 will output the value

```
100676
```

indicating that `t` contains 100676 words.

answer: ____________________________

h) Line 9 will set `t` as the default variable to be used in all function calls to the NLTK.

answer: ____________________________

i) Line 6 will output the value

```
set
```

indicating that `t` is of the type `set`.

`answer:` ____________________________

j) Line 7 will output the value

```
3
```

because that `t` contains the words “Wall”, “Street” and “Journal” words.

`answer:` ____________________________

k) Line 10 calculates the number of unique words in `t`.

`answer:` ____________________________

l) Line 12 calculates the number of sections in the Wall Street Journal data.

`answer:` ____________________________

m) Line 10 calculates the number of sections in the Wall Street Journal data.

`answer:` ____________________________

n) Line 12 provides a measure of the lexical diversity of `t`.

`answer:` ____________________________

o) Line 14 will output the list

```
[walk, walks, walking, walked]
```

indicating all the possible endings of the word “walk”.

p) Line 14 will output all the contexts in which the word “walk” was used in `t`.

`answer:` ____________________________

q) Line 14 will output all the contexts in which the word “walk” was used in `t`, along with whether it was used with the correct ending.

`answer:` ____________________________

r) The output of the line 14 follows the Key Word outside of Context format.

`answer:` ____________________________

s) The output of the line 14 follows the Key Word in Context format.

`answer:` ____________________________

t) Line 15 outputs a new data structure that allows for an efficient search of all occurrences of the word “walk”.

answer: ____________________________

u) Line 15 outputs a list containing the index of all occurrences of the word “walk”.

answer: ____________________________

v) Line 15 outputs the index of the first occurrences of the word“walk”.

answer: ____________________________

1.2 Plots 1

Consider the following piece of code:

```
import nltk
import nltk.book

t = nltk.book.text7
```

Write the line of code that would create a plot that shows the positions of occurrences in t of the word “knowledge”.

command: ____

1.3 Plots 2

Consider the following piece of code:

```
import nltk
import nltk.book

t = nltk.book.text7
freqs = nltk.FreqDist(t)
```

Write the line of code that would create a plot showing the frequencies of the 20 most common words t.

command: ____