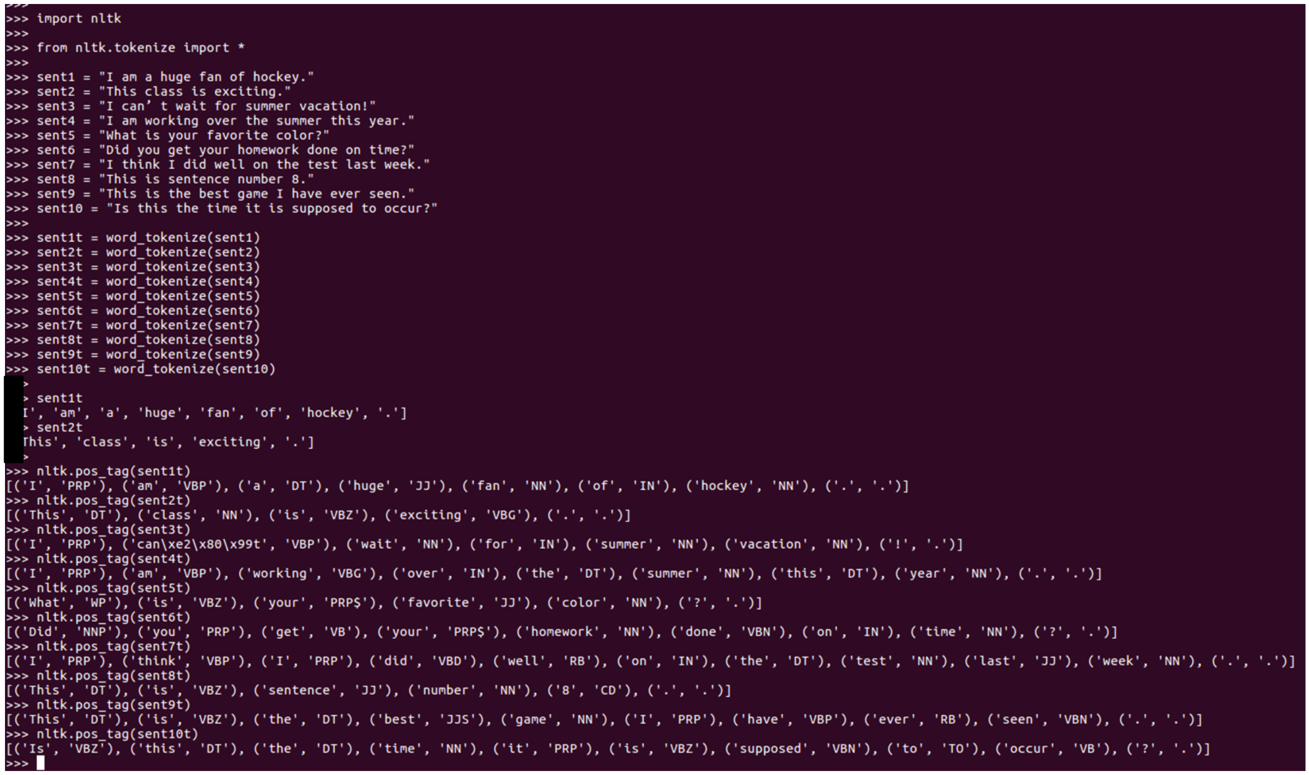
CHAPTER 5

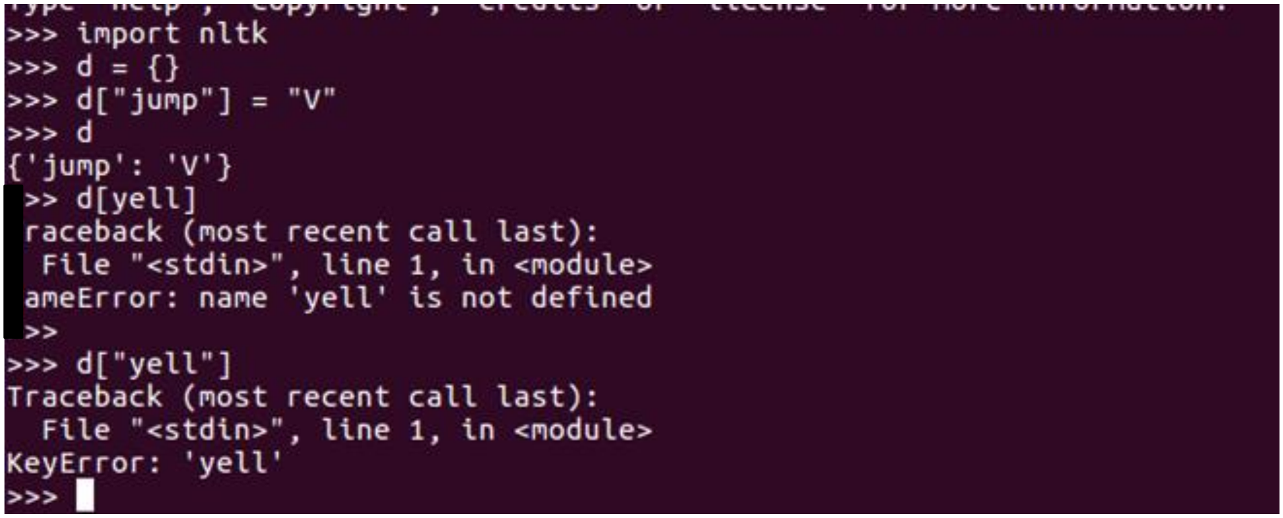
1) 10 sentences:



```
>>> import nltk
>>>
>>> from nltk.tokenize import *
>>>
>>> sent1 = "I am a huge fan of hockey."
>>> sent2 = "This class is exciting."
>>> sent3 = "I can’ t wait for summer vacation!"
>>> sent4 = "I am working over the summer this year."
>>> sent5 = "What is your favorite color?"
>>> sent6 = "Did you get your homework done on time?"
>>> sent7 = "I think I did well on the test last week."
>>> sent8 = "This is sentence number 8."
>>> sent9 = "This is the best game I have ever seen."
>>> sent10 = "Is this the time it is supposed to occur?"
>>>
>>> sent1t = word_tokenize(sent1)
>>> sent2t = word_tokenize(sent2)
>>> sent3t = word_tokenize(sent3)
>>> sent4t = word_tokenize(sent4)
>>> sent5t = word_tokenize(sent5)
>>> sent6t = word_tokenize(sent6)
>>> sent7t = word_tokenize(sent7)
>>> sent8t = word_tokenize(sent8)
>>> sent9t = word_tokenize(sent9)
>>> sent10t = word_tokenize(sent10)
[redacted]>
> sent1t
I', 'am', 'a', 'huge', 'fan', 'of', 'hockey', '.']
> sent2t
This', 'class', 'is', 'exciting', '.']
>
>>> nltk.pos_tag(sent1t)
[('I', 'PRP'), ('am', 'VBP'), ('a', 'DT'), ('huge', 'JJ'), ('fan', 'NN'), ('of', 'IN'), ('hockey', 'NN'), ('.', '.')]
>>> nltk.pos_tag(sent2t)
[('This', 'DT'), ('class', 'NN'), ('is', 'VBZ'), ('exciting', 'VBG'), ('.', '.')]
>>> nltk.pos_tag(sent3t)
[('I', 'PRP'), ('can\xe2\x80\x99t', 'VBP'), ('wait', 'NN'), ('for', 'IN'), ('summer', 'NN'), ('vacation', 'NN'), ('!', '.')]
>>> nltk.pos_tag(sent4t)
[('I', 'PRP'), ('am', 'VBP'), ('working', 'VBG'), ('over', 'IN'), ('the', 'DT'), ('summer', 'NN'), ('this', 'DT'), ('year', 'NN'), ('.', '.')]
>>> nltk.pos_tag(sent5t)
[('What', 'WP'), ('is', 'VBZ'), ('your', 'PRP$'), ('favorite', 'JJ'), ('color', 'NN'), ('?', '.')]
>>> nltk.pos_tag(sent6t)
[('Did', 'NNP'), ('you', 'PRP'), ('get', 'VB'), ('your', 'PRP$'), ('homework', 'NN'), ('done', 'VBN'), ('on', 'IN'), ('time', 'NN'), ('?', '.')]
>>> nltk.pos_tag(sent7t)
[('I', 'PRP'), ('think', 'VBP'), ('I', 'PRP'), ('did', 'VBD'), ('well', 'RB'), ('on', 'IN'), ('the', 'DT'), ('test', 'NN'), ('last', 'JJ'), ('week', 'NN'), ('.', '.')]
>>> nltk.pos_tag(sent8t)
[('This', 'DT'), ('is', 'VBZ'), ('sentence', 'JJ'), ('number', 'NN'), ('8', 'CD'), ('.', '.')]
>>> nltk.pos_tag(sent9t)
[('This', 'DT'), ('is', 'VBZ'), ('the', 'DT'), ('best', 'JJS'), ('game', 'NN'), ('I', 'PRP'), ('have', 'VBP'), ('ever', 'RB'), ('seen', 'VBN'), ('.', '.')]
>>> nltk.pos_tag(sent10t)
[('Is', 'VBZ'), ('this', 'DT'), ('the', 'DT'), ('time', 'NN'), ('it', 'PRP'), ('is', 'VBZ'), ('supposed', 'VBN'), ('to', 'TO'), ('occur', 'VB'), ('?', '.')]
>>>
```

2)

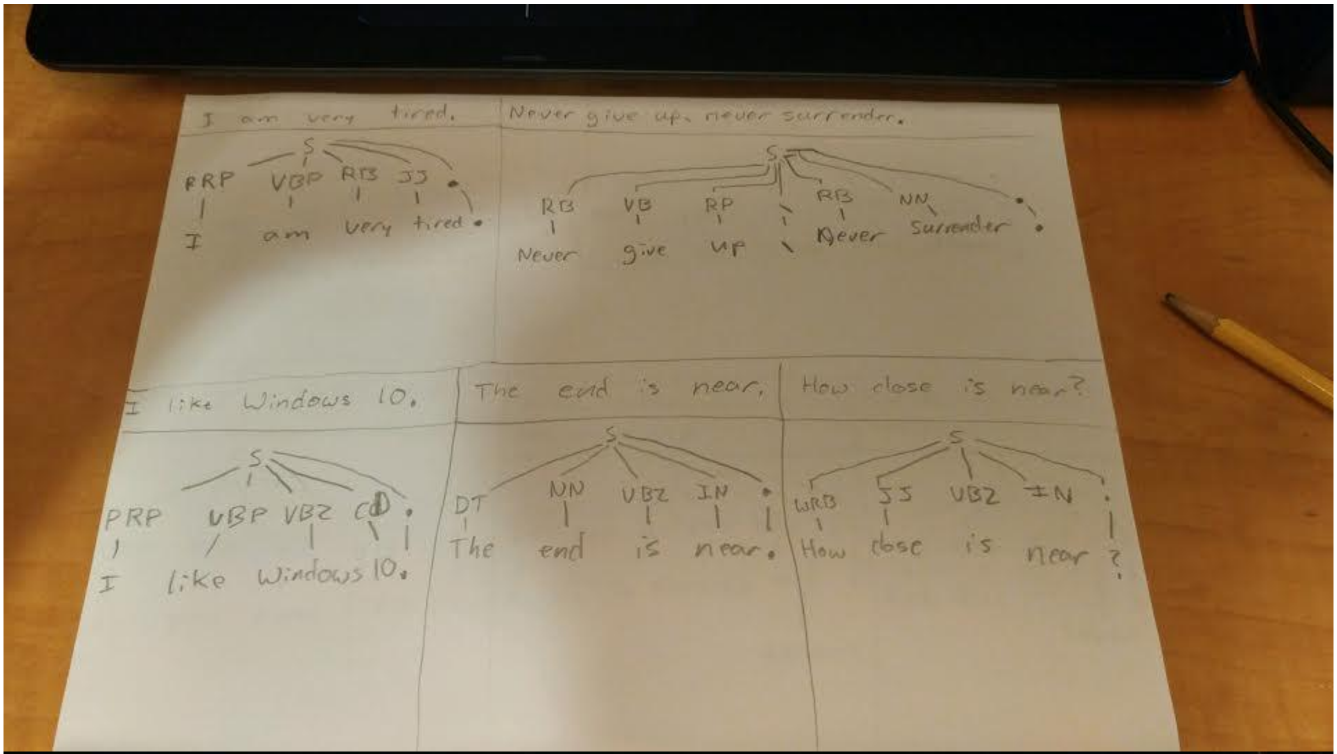
- When you type in something not in the dictionary it gives a key error.

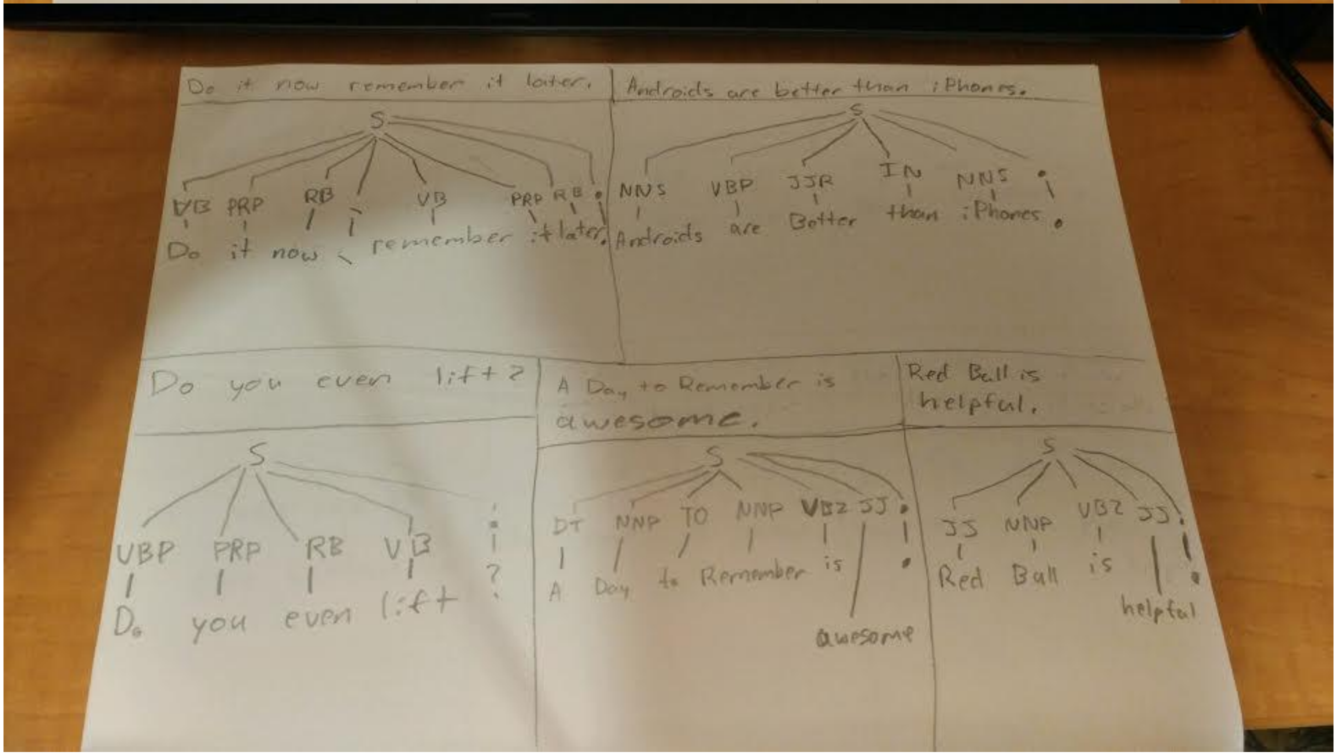


```
Type "help", "copyright", "credits" or "license" for more information.
>>> import nltk
>>> d = {}
>>> d["jump"] = "V"
>>> d
{'jump': 'V'}
[redacted]>> d[yell]
raceback (most recent call last):
  File "<stdin>", line 1, in <module>
ameError: name 'yell' is not defined
>>
>>> d["yell"]
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
KeyError: 'yell'
>>>
```

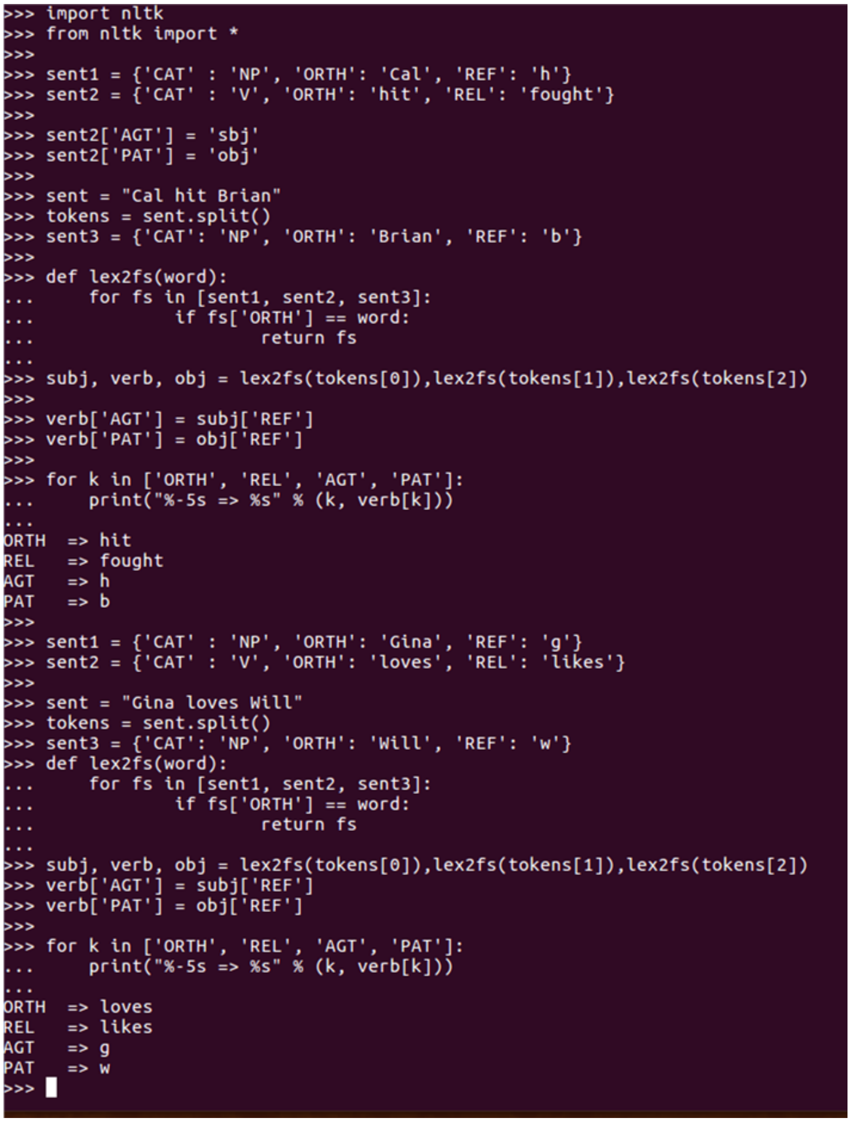
3) A large amount of data is required because when you take a large amount of data, and because it is common era on the side of safety by using 10% of the overall data, it allows for random samples of the data for testing.

4)


I am very tired.
S
PRP VBP RB JJ .
I am very tired .
Never give up, never surrender.
S
RB VB RP , RB NN .
Never give up , Never surrender .
I like Windows 10.
S
PRP VBP VBZ CD .
I like Windows 10.
The end is near.
S
DT NN VBZ IN .
The end is near.
How close is near?
S
WRB JJ VBZ IN .
How close is near ?


Do it now remember it later.
S
VB PRP RB , VB PRP RB .
Do it now , remember it later.
Androids are better than iPhones.
S
NNS VBP JJR IN NNS .
Androids are Better than iPhones .
Do you even lift?
S
VBP PRP RB VB .
Do you even lift ?
A Day to Remember is awesome.
S
DT NNP TO NNP VBZ JJ .
A Day to Remember is awesome .
Red Ball is helpful.
S
JJ NNP VBZ JJ .
Red Ball is helpful .

5)



```
>>> import nltk
>>> from nltk import *
>>>
>>> sent1 = {'CAT' : 'NP', 'ORTH': 'Cal', 'REF': 'h'}
>>> sent2 = {'CAT' : 'V', 'ORTH': 'hit', 'REL': 'fought'}
>>>
>>> sent2['AGT'] = 'sbj'
>>> sent2['PAT'] = 'obj'
>>>
>>> sent = "Cal hit Brian"
>>> tokens = sent.split()
>>> sent3 = {'CAT': 'NP', 'ORTH': 'Brian', 'REF': 'b'}
>>>
>>> def lex2fs(word):
...     for fs in [sent1, sent2, sent3]:
...             if fs['ORTH'] == word:
...                     return fs
...
>>> subj, verb, obj = lex2fs(tokens[0]),lex2fs(tokens[1]),lex2fs(tokens[2])
>>>
>>> verb['AGT'] = subj['REF']
>>> verb['PAT'] = obj['REF']
>>>
>>> for k in ['ORTH', 'REL', 'AGT', 'PAT']:
...     print("%-5s => %s" % (k, verb[k]))
...
ORTH  => hit
REL   => fought
AGT   => h
PAT   => b
>>>
>>> sent1 = {'CAT' : 'NP', 'ORTH': 'Gina', 'REF': 'g'}
>>> sent2 = {'CAT' : 'V', 'ORTH': 'loves', 'REL': 'likes'}
>>>
>>> sent = "Gina loves Will"
>>> tokens = sent.split()
>>> sent3 = {'CAT': 'NP', 'ORTH': 'Will', 'REF': 'w'}
>>> def lex2fs(word):
...     for fs in [sent1, sent2, sent3]:
...             if fs['ORTH'] == word:
...                     return fs
...
>>> subj, verb, obj = lex2fs(tokens[0]),lex2fs(tokens[1]),lex2fs(tokens[2])
>>> verb['AGT'] = subj['REF']
>>> verb['PAT'] = obj['REF']
>>>
>>> for k in ['ORTH', 'REL', 'AGT', 'PAT']:
...     print("%-5s => %s" % (k, verb[k]))
...
ORTH  => loves
REL   => likes
AGT   => g
PAT   => w
>>>
```