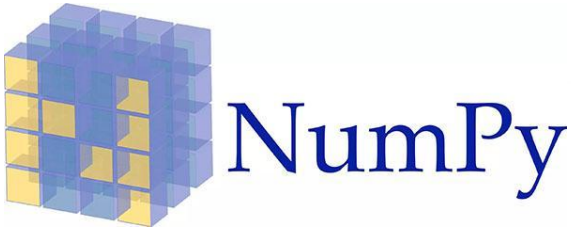
NLTK tutorial

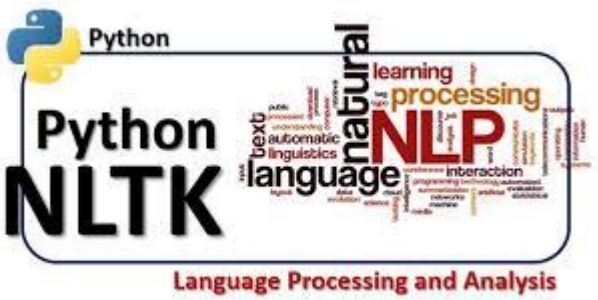
- the Natural Language Toolkit

#2 class

- Numpy tutorial
- NLTK tutorial

Numy & NLTK


NumPy


Python
Python
NLTK
natural
learning
processing
NLP
text
automatic
linguistics
language
interaction
Language Processing and Analysis

NLTK Tutorial

- Natural language processing (NLP) is the automatic or semi-automatic processing of human language. NLP is closely related to linguistics and has links to research in cognitive science, psychology, physiology, and mathematics.

What Is NLTK?

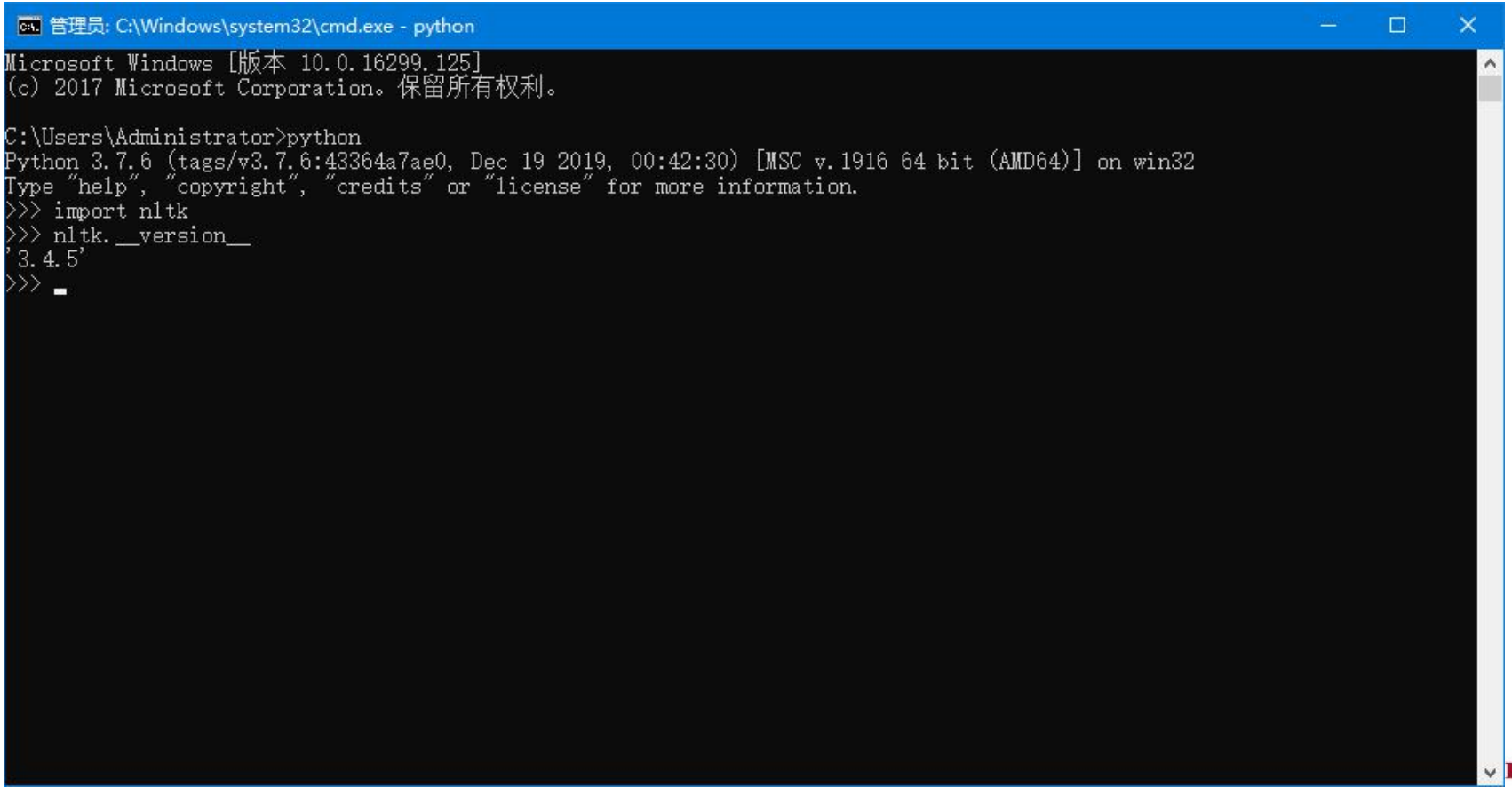
- The Natural Language Toolkit (NLTK) is a platform used for building programs for text analysis. The platform was originally released by Steven Bird and Edward Loper.
- There is an accompanying book for the platform called Natural Language Processing with Python.

Installing NLTK

- Installing NLTK is very simple. In Windows 10, with Command Prompt (MS-DOS), type the following command:

```
pip install nltk
```

Installing NLTK



Stop Words

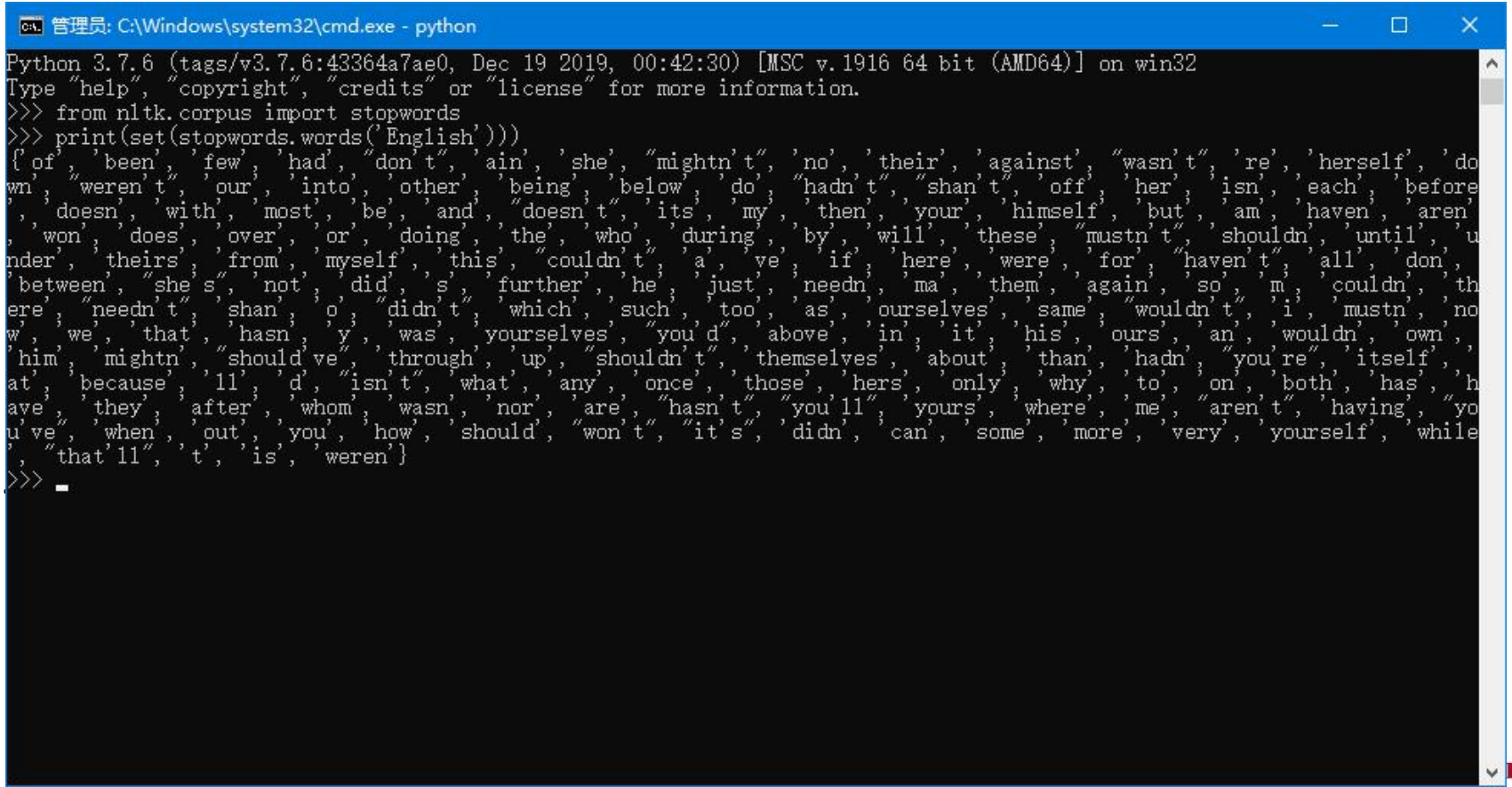
- Sometimes we need to filter out useless data to make the data more understandable by the computer. In natural language processing (NLP), such useless data (words) are called stop words. So, these words to us have no meaning, and we would like to remove them.

Stop Words

- NLTK provides us with some stop words to start with. To see those words, use the following script:

```
from nltk.corpus import stopwords
print(set(stopwords.words('English')))
```

Stop Words


管理员: C:\Windows\system32\cmd.exe - python
Python 3.7.6 (tags/v3.7.6:43364a7ae0, Dec 19 2019, 00:42:30) [MSC v.1916 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license" for more information.
>>> from nltk.corpus import stopwords
>>> print(set(stopwords.words('English')))
{'of', 'been', 'few', 'had', "don't", 'ain', 'she', "mightn't", 'no', 'their', 'against', "wasn't", 're', 'herself', 'down', "weren't", 'our', 'into', 'other', 'being', 'below', 'do', "hadn't", "shan't", 'off', 'her', 'isn', 'each', 'before', 'doesn', 'with', 'most', 'be', 'and', "doesn't", 'its', 'my', 'then', 'your', 'himself', 'but', 'am', 'haven', 'aren', 'won', 'does', 'over', 'or', 'doing', 'the', 'who', 'during', 'by', 'will', 'these', "mustn't", 'shouldn', 'until', 'under', 'theirs', 'from', 'myself', 'this', "couldn't", 'a', 've', 'if', 'here', 'were', 'for', "haven't", 'all', 'don', 'between', "she's", 'not', 'did', 's', 'further', 'he', 'just', 'needn', 'ma', 'them', 'again', 'so', 'm', 'couldn', 'there', "needn't", 'shan', 'o', "didn't", 'which', 'such', 'too', 'as', 'ourselves', 'same', "wouldn't", 'i', 'mustn', 'now', 'we', 'that', 'hasn', 'y', 'was', 'yourselves', "you'd", 'above', 'in', 'it', 'his', 'ours', 'an', 'wouldn', 'own', 'him', 'mightn', "should've", 'through', 'up', "shouldn't", 'themselves', 'about', 'than', 'hadn', "you're", 'itself', 'at', 'because', 'll', 'd', "isn't", 'what', 'any', 'once', 'those', 'hers', 'only', 'why', 'to', 'on', 'both', 'has', 'have', 'they', 'after', 'whom', 'wasn', 'nor', 'are', "hasn't", "you'll", 'yours', 'where', 'me', "aren't", 'having', "you've", 'when', 'out', 'you', 'how', 'should', "won't", "it's", 'didn', 'can', 'some', 'more', 'very', 'yourself', 'while', "that'll", 't', 'is', 'weren'}
>>>

Stop Words

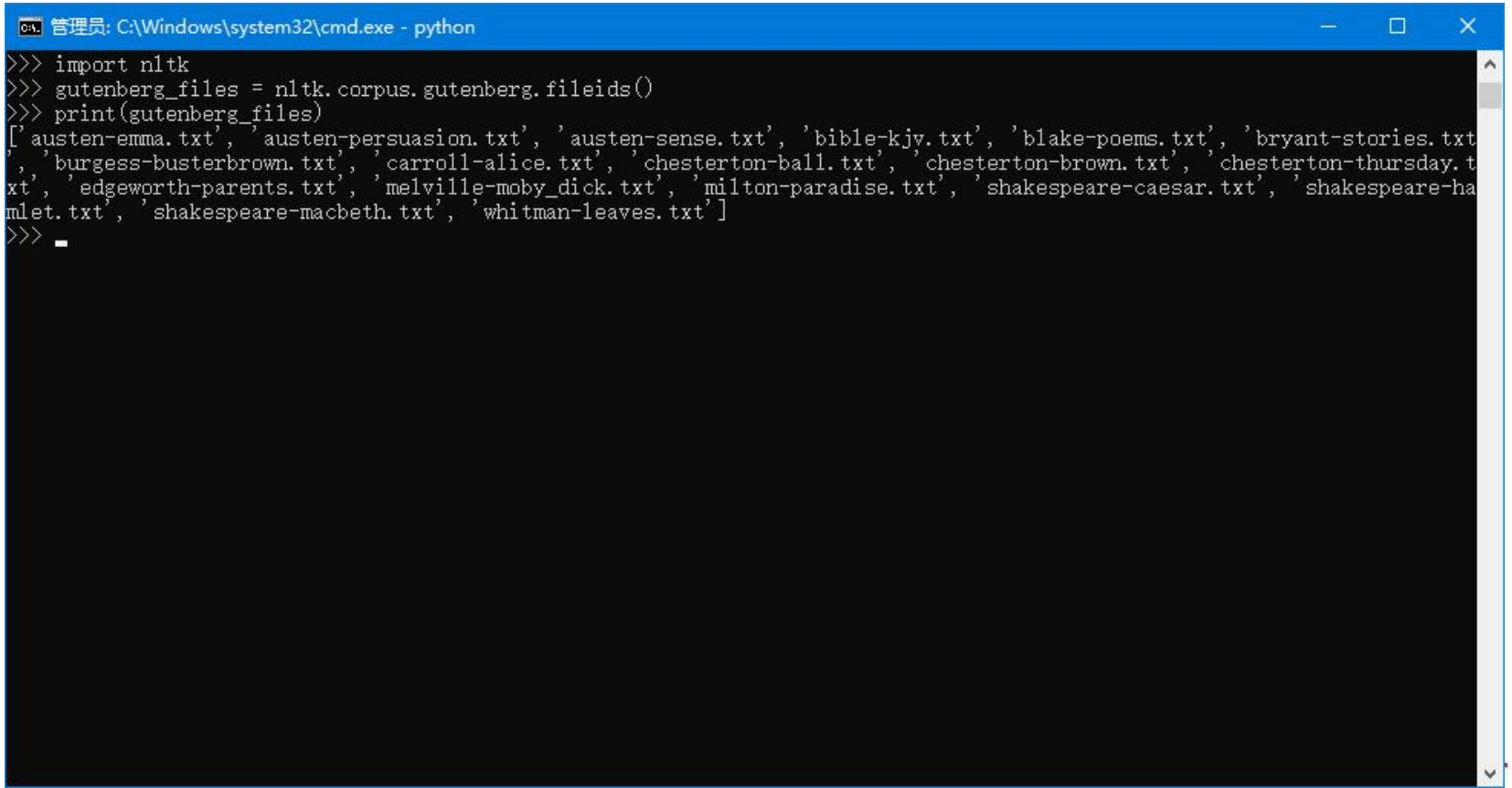
```
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize


text = 'In this tutorial, I\'m learning NLTK. It is an interesting platform.'
stop_words = set(stopwords.words('english'))
words = word_tokenize(text)
 new_sentence = []
 for word in words:
    if word not in stop_words:
        new_sentence.append(word)
 print(new_sentence)
```

The Gutenberg Corpus

- NLTK contains a small selection of texts from Project Gutenberg. To see the included files from Project Gutenberg, we do the following:

```
import nltk
gutenberg_files = nltk.corpus.gutenberg.fileids()
print(gutenberg_files)
```

The Gutenberg Corpus



The Gutenberg Corpus

- If we want to find the number of words for the text file bryant-stories.txt for instance, we can do the following:

```
import nltk

bryant_words = nltk.corpus.gutenberg.words('bryant-stories.txt')
print(len(bryant_words))
```

Conclusion

- As we have seen in this tutorial, the NLTK platform provides us with a powerful tool for working with natural language processing (NLP).
- If you would like to go deeper into using NLTK for different NLP tasks, you can refer to NLTK's accompanying book: Natural Language Processing with Python.