INLS 560
Programming for Information Professionals

Text Analysis

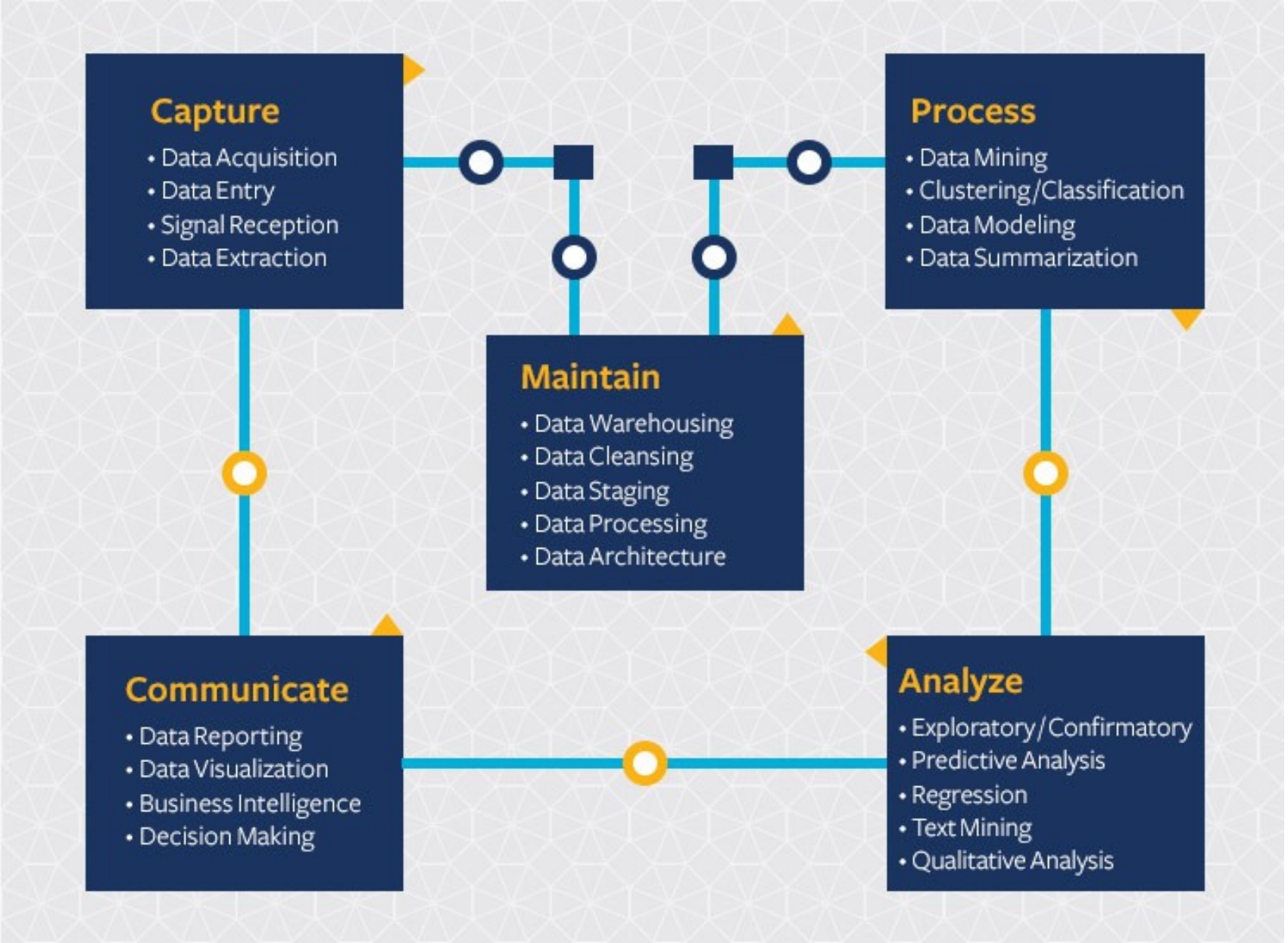
UNC
SCHOOL OF INFORMATION
AND LIBRARY SCIENCE

Joan Boone
jpboone@email.unc.edu

Part 1: Overview

Part 2: Text Analysis Techniques

What is Data Science?



Berkeley School of Information: What is Data Science?

Text Analysis

Data Mining: Practical Machine Learning Tools and Techniques
by Ian H. Witten, Eibe Frank, and Mark A. Hall

- Text mining is about looking for patterns in text
- The information extracted should be potentially useful
- Output can summarize salient features from a large body of text

There are many techniques for extracting useful, high-quality information from a document, and they often rely on 3rd party Python libraries, such as

- NLTK for natural language processing (NLP)
- Sci-kit for machine learning (ML)
- NumPy, pandas for scientific computing and data analysis

Text data takes many forms

Unstructured text

- Published content, data generated from speech, tweets
- Processing often requires Natural Language Processing (NLP) tools that work with human language data to categorize words, classify text and analyze sentence structure and meaning

Tabular data (semi-structured)

- Typically organized in rows and columns
- Examples: spreadsheets, CSV files, log data

Structured data

- Organized in a specific format that describes and defines data
- Examples: JSON and XML data formats

Unstructured Text

Project Gutenberg

collection of free e-books

Project Gutenberg › 68,062 free ebooks › 119 by Robert Louis Stevenson

Treasure Island by Robert Louis Stevenson

Download This eBook

	Format	Size
	Read this book online: HTML	503 kB
	EPUB (with images)	74.8 MB
	EPUB (no images)	227 kB
	Kindle (with images)	162.7 MB
	Kindle (no images)	820 kB
	Plain Text UTF-8	391 kB
	More Files...	

The Old Sea-dog at the Admiral Benbow

SQUIRE TRELAWNEY, Dr. Livesey, and the rest of these gentlemen having asked me to write down the whole particulars about Treasure Island, from the beginning to the end, keeping nothing back but the bearings of the island, and that only because there is still treasure not yet lifted, I take up my pen in the year of grace 17__ and go back to the time when my father kept the Admiral Benbow inn and the brown old seaman with the sabre cut first took up his lodging under our roof.

I remember him as if it were yesterday, as he came plodding to the inn door, his sea-chest following behind him in a hand-barrow--a tall, strong, heavy, nut-brown man, his tarry pigtail falling over the shoulder of his soiled blue coat, his hands ragged and scarred, with black, broken nails, and the sabre cut across one cheek, a dirty, livid white. I remember him looking round the cove and whistling to himself as he did so, and then breaking out in that old sea-song that he sang so often afterwards:

"Fifteen men on the dead man's chest--
Yo-ho-ho, and a bottle of rum!"

in the high, old tottering voice that seemed to have been tuned and broken at the capstan bars. Then he rapped on the door with a bit of stick like a handspike that he carried, and when my father appeared, called roughly for a glass of rum. This, when it was brought to him, he drank slowly, like a connoisseur, lingering on the taste and still looking about him at the cliffs and up at our signboard.

"This is a handy cove," says he at length; "and a pleasant sittyated grog-shop. Much company, mate?"

My father told him no, very little company, the more was the pity.

"Well, then," said he, "this is the berth for me. Here you, matey," he cried to the man who trundled the barrow; "bring up alongside and help up my chest. I'll stay here a bit," he continued. "I'm a plain man; rum and bacon and eggs is what I want, and that head up there for to watch ships off. What you mought call me? You mought call me captain. Oh, I see what you're at--there"; and he threw down three or four gold pieces on the threshold. "You can tell me when I've worked through that," says he, looking as fierce as a commander.

And indeed bad as his clothes were and coarsely as he spoke, he had none of the appearance of a man who sailed before the mast, but seemed like a mate or skipper accustomed to be obeyed or to strike. The man who came with the barrow told us the mail had set him down the morning before at

Part 1: Overview

Part 2: Text Analysis Techniques

Text Analysis

Before applying advanced techniques such as NLP and ML, text is often pre-processed – it is transformed to make it more computation-friendly, and to enable some preliminary text analysis.

For example, generating word frequencies can assist in classifying content, especially for very large text documents.

- Words that occur most frequently are likely to convey what a document is about
- Word frequencies are used to build indexes to improve the speed of search, and to define metadata that can improve content discovery

Several common techniques for analyzing text include

- Tokenization
- Stopword removal
- Normalization (stemming and lemmatization) where words are reduced to their root form

Tokenization: extracting words

Tokenization is the process of converting unstructured text into a list of single words. The built-in Python String methods greatly simplify this task.

For example, if you have read a line of text from a file, this text can be converted to a list of words using the String method **`split()`** to split a string using the specified separator – if not specified, the separator defaults to any whitespace.

```
line_of_text = "In the meantime, we had found nothing of any value but
the silver and the trinkets, and neither of these were in our way."

list_of_words = line_of_text.split()

print(list_of_words)

['In', 'the', 'meantime,', 'we', 'had', 'found', 'nothing', 'of',
'any', 'value', 'but', 'the', 'silver', 'and', 'the', 'trinkets,',
'and', 'neither', 'of', 'these', 'were', 'in', 'our', 'way.']
```

w3schools: Python String split() Method

Tokenization: removing punctuation

Once we have a list of words, there is often additional data cleaning to be done. In the previous example, the list contained the following words:

```
'long,' 'hall,' and 'roof.'
```

The String method `strip()` can remove punctuation, by specifying the set of characters to be removed. The Python String constant, `string.punctuation` defines this set of characters:

```
!"#$%&'()*+,-./:;<=>?@[\]^_`{|}~.
```

Lastly, use the String method `lower()` to convert characters to lower case.

```
import string
...
for word in list_of_words:
    print(word.strip(string.punctuation).lower(), end=' ')


in the meantime we had found nothing of any value but the silver
and the trinkets and neither of these were in our way
```

w3schools: Python String strip() Method

Stopword Removal

Stopwords are words in a text that do not carry useful information, such as articles (“the”), conjunctions (“and”), and prepositions (“from”). These words occur frequently, and generally convey little about the meaning or content of a document, so they are often removed from text before further analysis.

There are many stopword lists available, and they can be customized for domain-specific needs of an application. The one used with this example was developed by Buckley and Salton.

Below is a paragraph from the *Treasure Island* text that highlights the stopwords found, and subsequently removed from the list of words.

```
In the meantime, we had found nothing of any value but the
silver and the trinkets, and neither of these were in our way.
Underneath there was an old boat-cloak, whitened with sea-salt
on many a harbour-bar. My mother pulled it up with impatience,
and there lay before us, the last things in the chest, a
bundle tied up in oilcloth, and looking like papers, and a
canvas bag that gave forth, at a touch, the jingle of gold.
```

Tokenization + Stopword Removal

Below is the version of the original paragraph text where the stopwords and punctuation are highlighted:

```
In the meantime, we had found nothing of any value but the
silver and the trinkets, and neither of these were in our way.
Underneath there was an old boat-cloak, whitened with sea-salt
on many a harbour-bar. My mother pulled it up with impatience,
and there lay before us, the last things in the chest, a
bundle tied up in oilcloth, and looking like papers, and a
canvas bag that gave forth, at a touch, the jingle of gold.
```

This is the revised version where the stopwords and the leading and trailing punctuation have been removed, and all characters converted to lower case.

```
meantime found silver trinkets underneath boat-cloak
whitened sea-salt harbour-bar mother pulled impatience lay
things chest bundle tied oilcloth papers canvas bag gave
touch jingle gold
```

Generating Word Frequencies

Tokenization and stopword removal are the pre-processing steps taken before generating word frequencies.

We now have a list of (somewhat) meaningful words that occur in the text, but how many times does each word occur?

The general algorithm is to loop through the list and count how many times each word occurs. A `Python dictionary` is the best data structure to represent the results:

- ***key*** is a word in the text
- ***value*** is its frequency in the text

Word	Frequency
man	232
captain	208
silver	201
doctor	159
time	130
good	123
hand	119
long	113
back	106
. . .	. . .