THE UNIVERSITY OF RHODE ISLAND

Introduction to Text Mining

Mohammad A Al-Mamun, PhD
Department of Pharmacy Practice
November 22, 2019

THINK BIG WE DO

THE UNIVERSITY OF RHODE ISLAND

Text Analysis

- Text analysis refers to the representation, processing, and modeling of textual data to derive useful insights.

- It is a technique that computers use to extract worthwhile information from the human natural language in a smart and efficient manner.

THINK BIG WE DO

THE UNIVERSITY OF RHODE ISLAND

What is Text Mining?

THINK BIG WE DO

What is Text Mining?

Text mining is the process of discovering relationships and interesting patterns in large text collections

Text Analysis Process

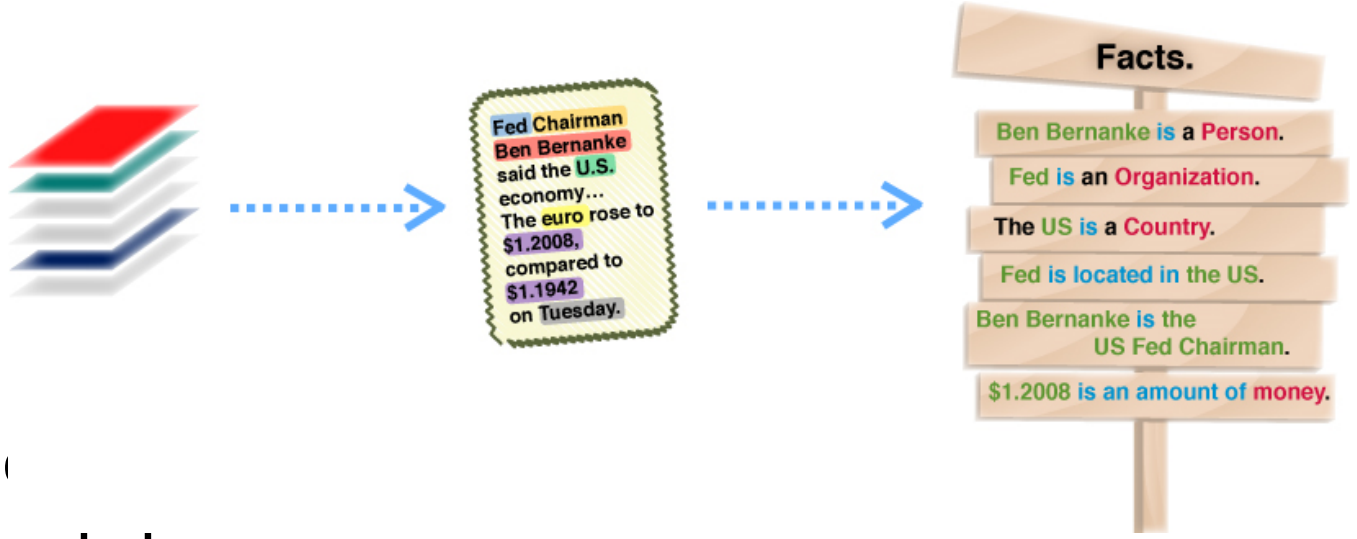
Usually, there are three steps involved in text analysis:

1. Parsing
2. Search and Retrieval
3. Text Mining

Parsing

- Parsing is the process that takes unstructured text and imposes a structure for further analysis

- Example: The unstructured text could be a plain text file, weblog, an HTML file, or a Word document

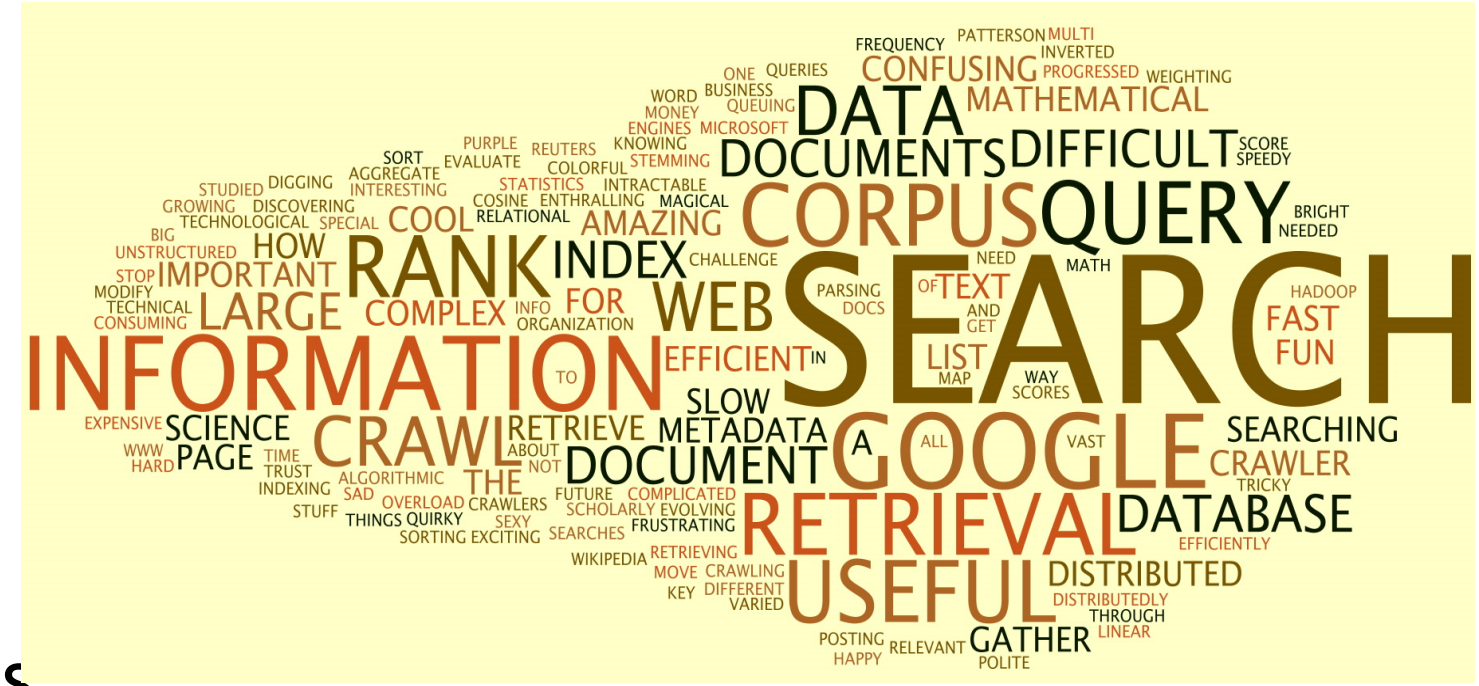


- Parsing deconstructs the provid structured way for the subsequent steps

Search and Retrieval

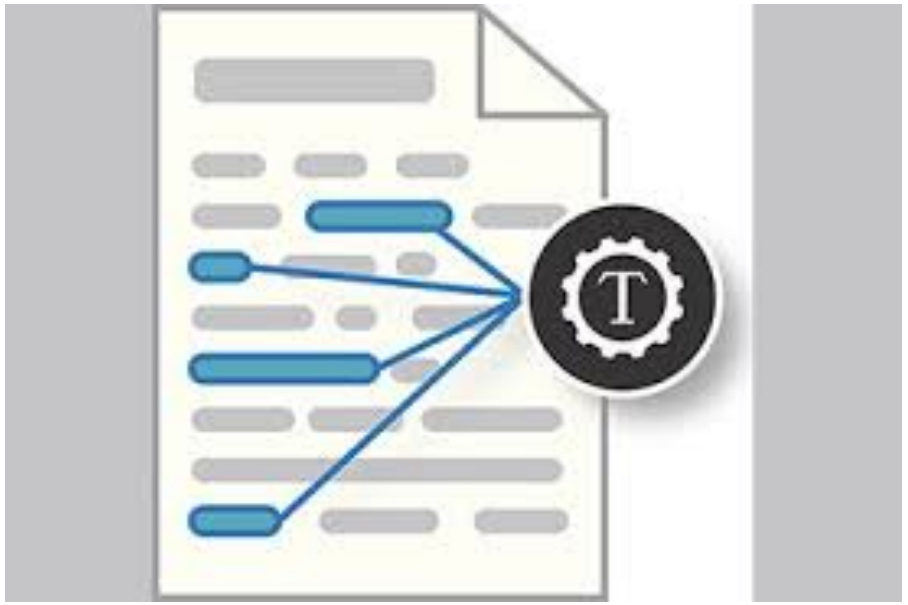
- This process involves identification of the documents in a corpus that contain search items such as specific words, phrases, topics, or entities like people or organizations

- These search items are generally called key terms



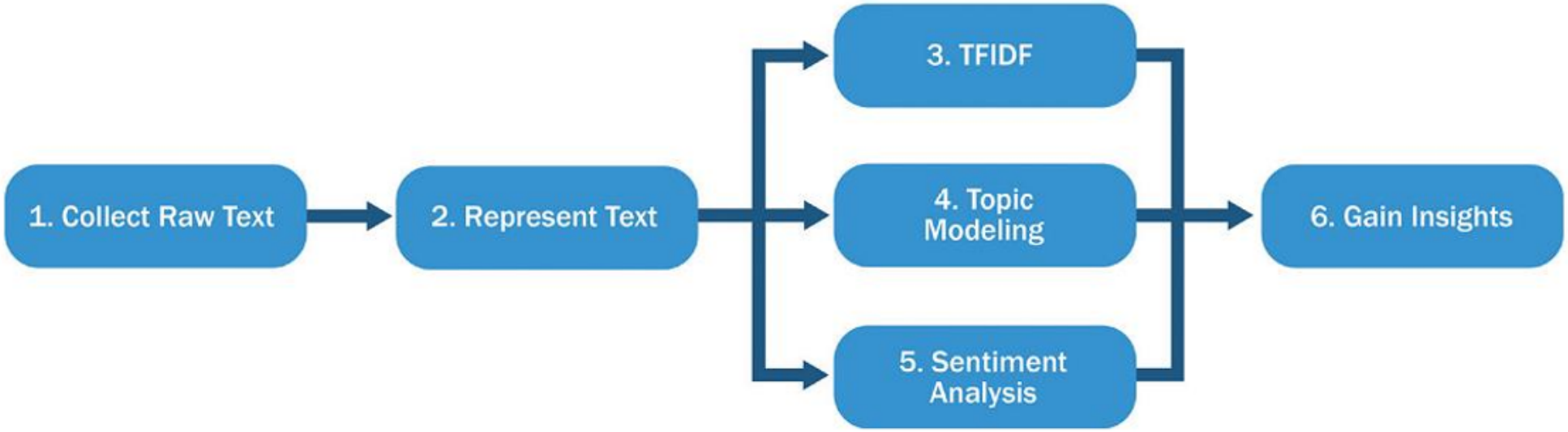
Mining Text

- Text mining uses the terms and indexes produced by the prior two steps to discover meaningful insights pertaining to domains or problems of interest



Text Analysis – An example:

- An analysis of some literature works of Jane Austen



Text Analysis Pipeline

1. Collecting raw text: Loading the data from six of Jane Austen's books as text using R packages (janeaustenr, gutenbergr)

2. Representing text: To work with this data we need to restructure it into a one-token-per-row format

3. TFIDF (Term Frequency – Inverse Document Frequency): Compute the usefulness of each word in the books

Text Analysis Pipeline

4. Topic Modeling: Categorize all the books by topics

5. Sentiment Analysis: Determine sentiments in our data. Use the tools of text mining to approach the emotional content of text programmatically

6. Gain Insights: Review the results and gain deeper insights from our data

Step 1: Collecting Raw Text

- For our example, we are using the **janeaustenr** package to load the data.
- Optionally, we can also use the **gutenbergr** package to obtain data from vast collections of literary works
- After loading data which is in **a one-row-per-line format**, transform them into a **tidy** format with a line number and chapter numbers

Step 1: Collecting Raw Text

Output looks like this:

	text	book	linenumber	chapter
	<chr>	<fct>	<int>	<int>
1	SENSE AND SENSIBILITY	Sense & Sensibility	1	0
2	""	Sense & Sensibility	2	0
3	by Jane Austen	Sense & Sensibility	3	0
4	""	Sense & Sensibility	4	0
5	(1811)	Sense & Sensibility	5	0
6	""	Sense & Sensibility	6	0
7	""	Sense & Sensibility	7	0
8	""	Sense & Sensibility	8	0
9	""	Sense & Sensibility	9	0
10	CHAPTER 1	Sense & Sensibility	10	1

... with 73,412 more rows

Step 2: Representation

- To proceed further, we need to restructure the data into the one-token-per-row format
- The function **unnest_tokens()** from the tokenizers package is used for this
- Each line of text in the original data is separated into tokens.
- Other examples are **case folding** and **bag-of-words** representation

Step 2: Representation

- After tokenization, the data now looks something like this

	book	linenumber	chapter	word
1	Sense & Sensibility	1	0	sense
2	Sense & Sensibility	1	0	and
3	Sense & Sensibility	1	0	sensibility
4	Sense & Sensibility	3	0	by
5	Sense & Sensibility	3	0	jane
6	Sense & Sensibility	3	0	austen
7	Sense & Sensibility	5	0	1811
8	Sense & Sensibility	10	1	chapter
9	Sense & Sensibility	10	1	1
10	Sense & Sensibility	13	1	the
11	Sense & Sensibility	13	1	family
12	Sense & Sensibility	13	1	of

Step 3: TF-IDF

- Refers to analysis of word and document frequency
- **Term Frequency (TF)** - how frequently a word occurs in a document
  It is one measure of how important a word may be in our data
- **Inverse Document Frequency (IDF)** - decreases the weight for commonly used words and increases the weight for rarely used words

Step 3: TF-IDF

TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document).

IDF(t) = log_e(Total number of documents / Number of documents with term t in it)

Step 3: TF-IDF- an example

A document containing 100 words wherein the word *brain* appears 3 times.

The term frequency (i.e., TF) for *brain* is then (3 / 100) = 0.03.

Now, assume we have 10 million documents and the word *brain* appears in one thousand of these.

The inverse document frequency (i.e., idf) is calculated as log(10,000,000 / 1,000) = 4

Thus, the TF-IDF weight is the product of these quantities: 0.03 * 4 = 0.12.

Step 3: TF-IDF

- Term frequency in our data (TF)
- There is one row for each word-book combination
- n is the number of times that word is used in that book
- total is the total words in that book
- The usual words are here with the highest n - “the”, “and”, “to”, and so on

	book	word	n	total
1	Mansfield Park	the	6206	160460
2	Mansfield Park	to	5475	160460
3	Mansfield Park	and	5438	160460
4	Emma	to	5239	160996
5	Emma	the	5201	160996
6	Emma	and	4896	160996
7	Mansfield Park	of	4778	160460
8	Pride & Prejudice	the	4331	122204
9	Emma	of	4291	160996
10	Pride & Prejudice	to	4162	122204
11	Sense & Sensibility	to	4116	119957
12	Sense & Sensibility	the	4105	119957

Step 4: TF-IDF

- The bind_tf_idf function is used for evaluating this value
- idf and thusTF-IDF are zero for extremely common words
- The inverse document frequency (and thus TF-IDF) is very low (near zero) for words that occur in many of the documents in a collection

Step 4: TF-IDF

- This approach decreases the weight for common words
- We see that all proper nouns names that are important in these novels, have highest tf-idf values.

	book	word	n	tf	idf	tf_idf
	<fct>	<chr>	<int>	<dbl>	<dbl>	<dbl>
1	Sense & Sensibility	elinor	623	0.00519	1.79	0.00931
2	Sense & Sensibility	marianne	492	0.00410	1.79	0.00735
3	Mansfield Park	crawford	493	0.00307	1.79	0.00551
4	Pride & Prejudice	darcy	373	0.00305	1.79	0.00547
5	Persuasion	elliot	254	0.00304	1.79	0.00544
6	Emma	emma	786	0.00488	1.10	0.00536
7	Northanger Abbey	tilney	196	0.00252	1.79	0.00452
8	Emma	weston	389	0.00242	1.79	0.00433
9	Pride & Prejudice	bennet	294	0.00241	1.79	0.00431
10	Persuasion	wentworth	191	0.00228	1.79	0.00409

... with 40,369 more rows

THE UNIVERSITY OF RHODE ISLAND

Step 5: Topic Modelling

- Topic modeling provides tools to automatically organize, search, understand, and summarize from vast amounts of information.
- More details on this next week

Step 6: Sentiment Analysis

- Sentiment analysis refers to a group of tasks that use statistics and natural language processing to mine opinions, in order to identify and extract subjective information from texts.

- There are a variety of methods and dictionaries that exist for evaluating the opinion or emotion in text. For example: AFINN, bing, nrc

THE UNIVERSITY OF RHODE ISLAND

Step 6: Sentiment Analysis

Most common positive and negative words in our example:

Filter

	word	sentiment	n
1	miss	negative	1855
2	well	positive	1523
3	good	positive	1380
4	great	positive	981
5	like	positive	725
6	better	positive	639
7	enough	positive	613
8	happy	positive	534
9	love	positive	495
10	pleasure	positive	462
11	poor	negative	424
12	happiness	positive	369

miss

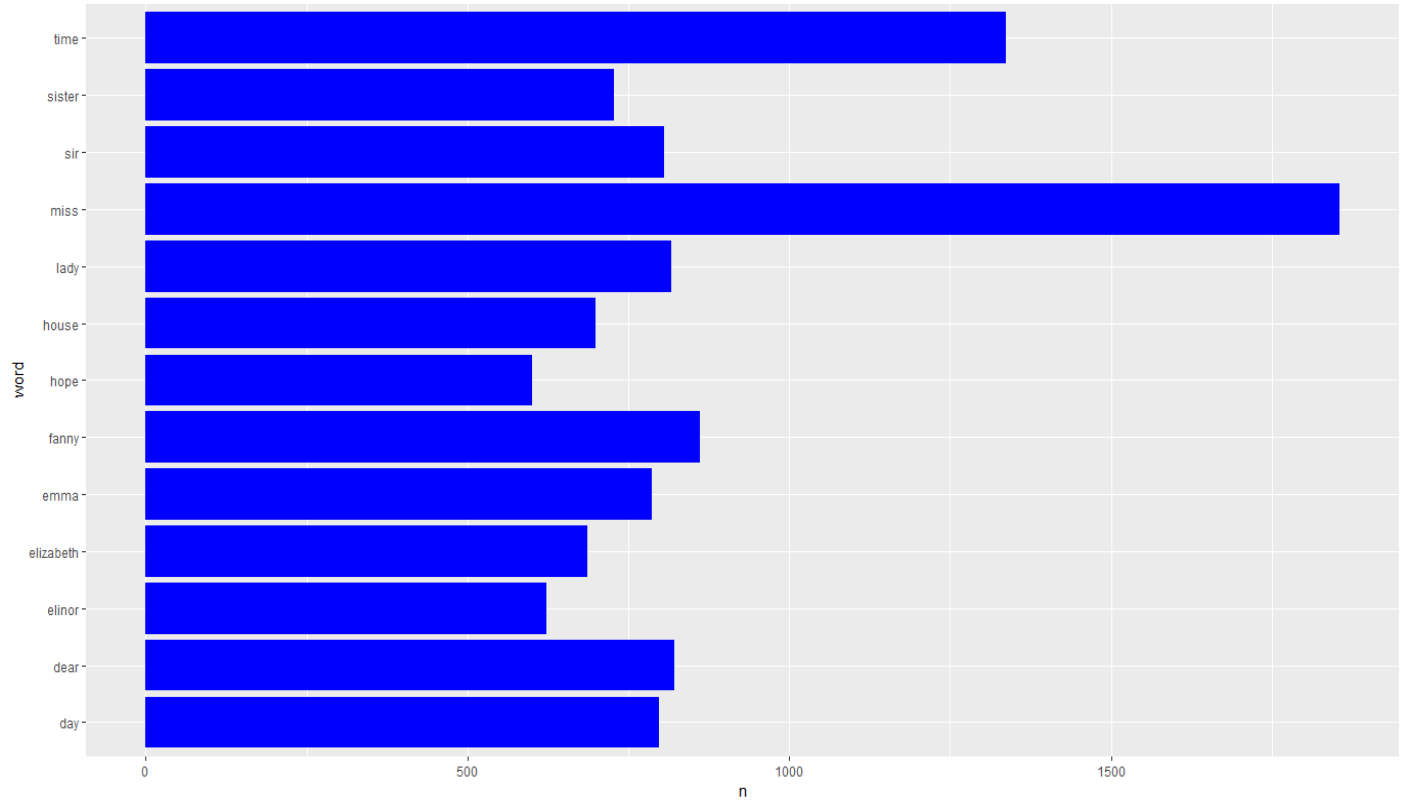
THINK BIG WE DO

Step 6: Sentiment Analysis

Word clouds with the most frequent positive and negative words

Step 7: Gain Insights

- The plot shows a visualization of the most common words in these books



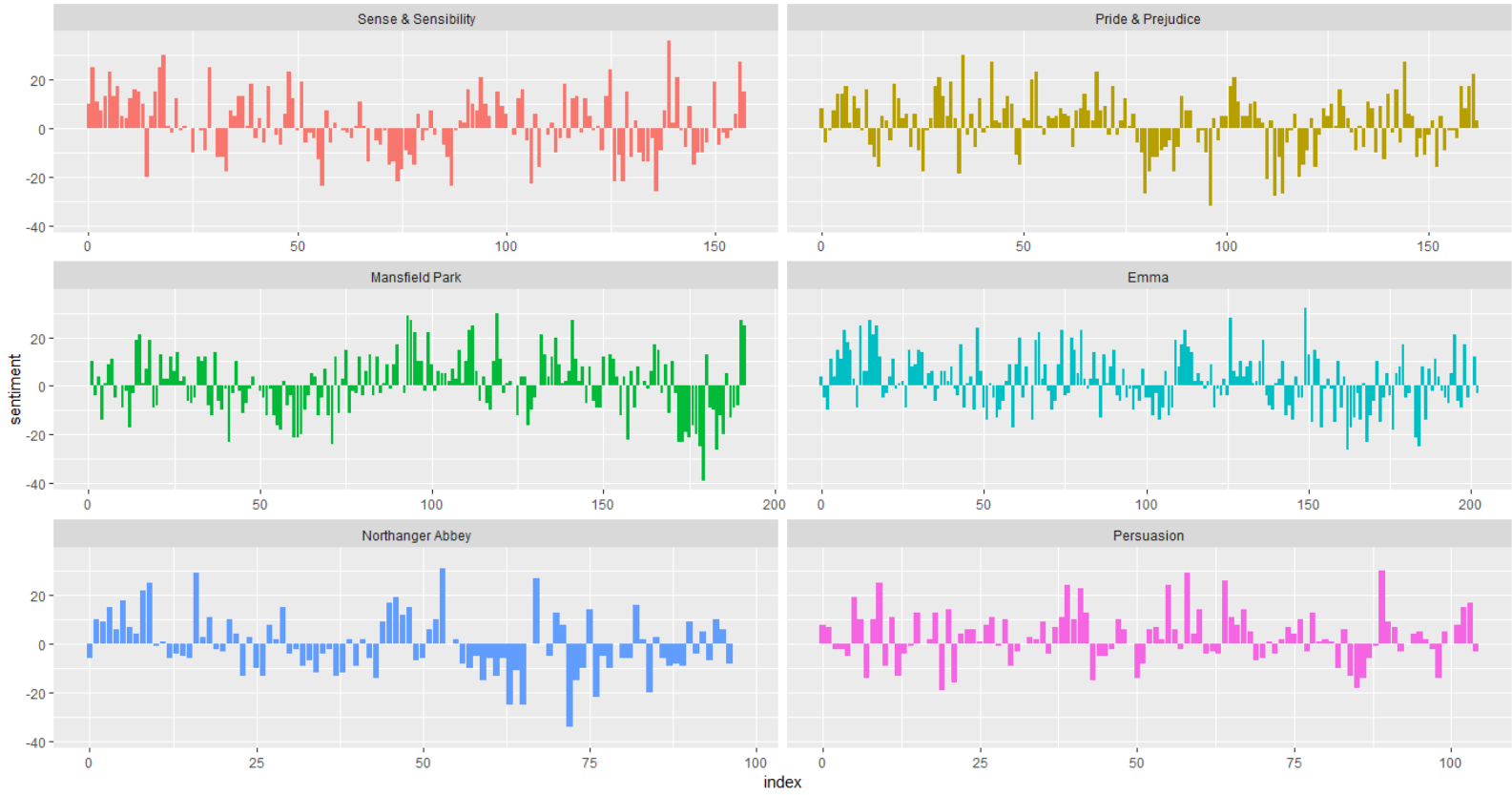
Step 7: Gain Insights

A word cloud (or tag cloud) is a visual representation of textual data

THE UNIVERSITY OF RHODE ISLAND

Step 7: Gain Insights

We can see how the plot of each novel changes toward more positive or negative sentiment over the trajectory of the story.

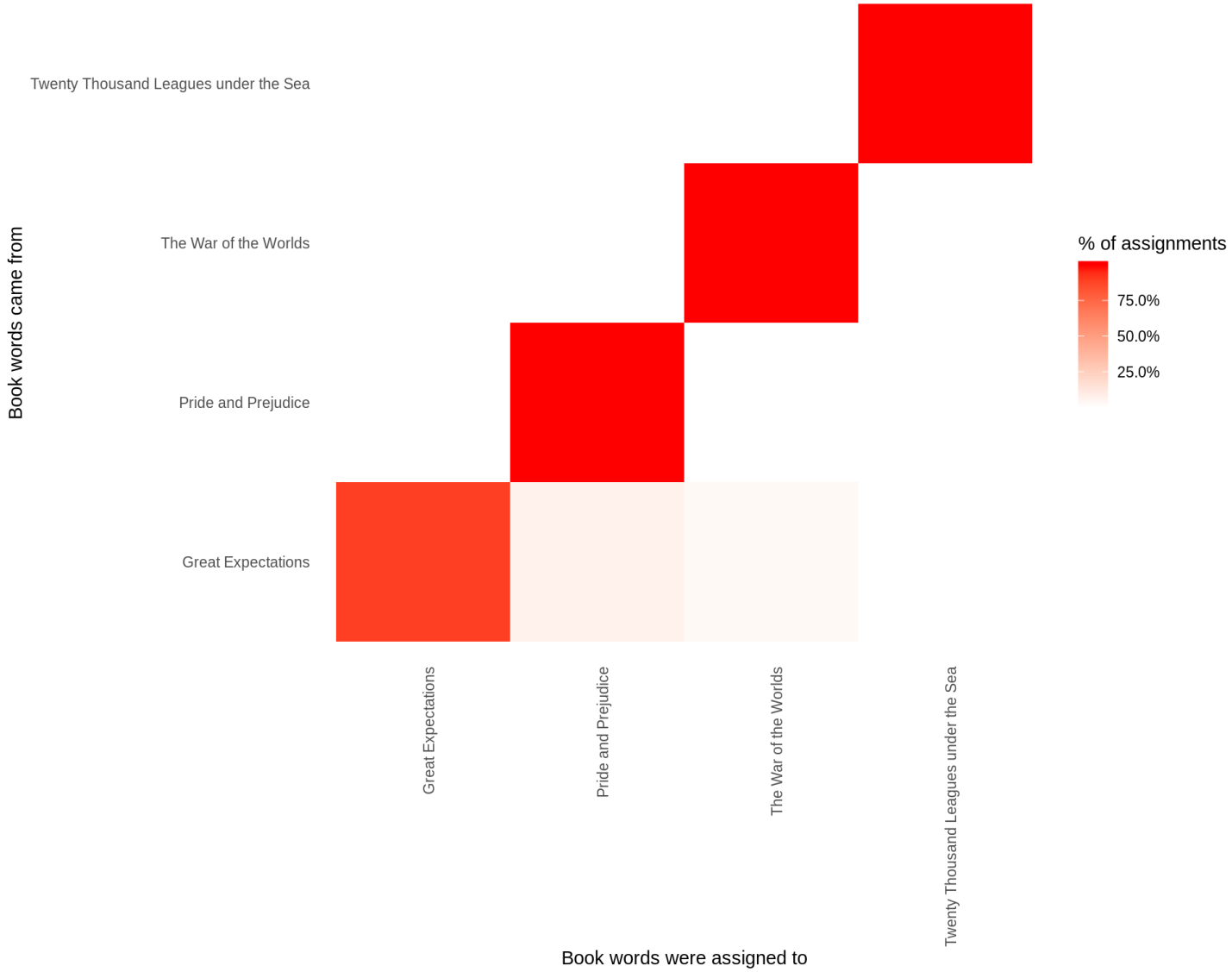


THINK BIG WE DO

THE UNIVERSITY OF RHODE ISLAND

Next Week

Topic Modelling



THINK BIG WE DO

Session 2 - R Hands-On

References:

- **Data Science and Big Data Analytics -** *by EMC*
- **Text Mining with R** – *by Julia Silge and David Robinson*