

TEXT MINING

Lecture 05

TEXT PRE-PROCESSING I

KEUNGOU I KIM

awekim@handong.edu



Regular Expression I: Position

- Always remember the way we read and the way a computer reads the text are different
 - Computer recognizes based on the “rule (or programming grammar)”
 - Example) Issue of space...

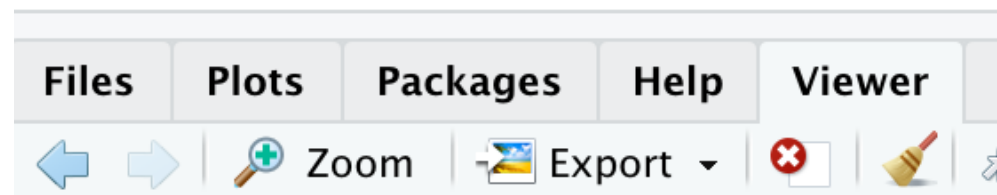
```
> women %>% head(1)
  height weight
1     58    115
> women %>%
+   head(1) %>%      nrow
[1] 1
```

- Regular expression
 - A formatted expression used for texts with certain patterns
 - Consists of characters & metacharacters
 - Metacharacter: ! + \ & ^ [] ~

Regular Expression

- Why we use regular expression
 - Verify the structure of strings
 - Extract substrings from structured strings
 - Search / replace / rearrange parts of the string
 - Split a string into tokens
- `stringr::str_view_all(text, pattern)`
 - Shows all matched patterns

```
> library(dplyr)
> library(magrittr)
> library(stringr)
> str_view_all("Text-Mining", "ing")
> |
```



Regular Expression

```
text.1 <- "Pep Guardiola expressed his admiration for Erling Haaland's attitude
after winning Man of the Match performance against Manchester United"
vector.1 <- c("I like to drink Pepsi", "I love Manchester United",
              "I have iphone", "iphone is very good")
```

- ^

- Starting point of text
- Allows us to capture the pattern that occurs at the beginning of text.
- Useful when we know the starting part of the text

```
> str_view_all(text.1, "^Pep")
```

```
Pep Guardiola expressed his admiration for Erling
Haaland's attitude after winning Man of the Match
performance against Manchester United
```

- \$

- End point of text
- Allows us to capture the pattern that occurs at the end of text.
- Useful when we know the latter part of the text

```
> str_view_all(text.1, "$")
```

```
Pep Guardiola expressed his admiration for Erling
Haaland's attitude after winning Man of the Match
performance against Manchester United
```

```
> str_view_all(vector.1, "^I")
```

```
I like to drink Pepsi
I love Manchester United
I have iphone
iphone is very good
```

```
> str_view_all(vector.1, "d$")
```

```
I like to drink Pepsi
I love Manchester United
I have iphone
iphone is very good
```

Regular Expression

```
text.1 <- "Pep Guardiola expressed his admiration for Erling Haaland's attitude  
after winning Man of the Match performance against Manchester United"  
vector.1 <- c("I like to drink Pepsi", "I love Manchester United",  
              "I have iphone", "iphone is very good")
```

- Only one letter to the left and right
- Useful when we know the middle of certain text pattern

```
> str_view_all(text.1, ".pres.")  
Pep Guardiola expressed his admiration for Erling  
Haaland's attitude after winning Man of the Match  
performance against Manchester United
```

```
> str_view_all(vector.1, ".p.")  
I like to drink Pepsi  
I love Manchester United  
I have iphone  
iphone is very good
```

Regular Expression II: Search

```
car <- c('car','cr','caar','caaar','caaaar')  
apple <- c('apple','applee','aple','appple','apleee')
```

- ? (match-zero-or-one)

- The preceding item is optional and will be matched at most once.
- Used when we remember only a certain part of the text
- Recommended to use when the text you know is not certain.

```
> str_view_all(car, "ca?r") str_view_all(apple, "ap?le")
```

car

cr

Even if that is not
matched perfectly (match-
zero), we can select them

caar

caaar

caaaar

apple

applee

aple

appple

apleee

The most
conservative way of
matching the patterns

Regular Expression

```
car <- c('car','cr','caar','caaar','caaaar')  
apple <- c('apple','applee','aple','appple','apleee')
```

- * (match-zero-or-more)

- The preceding item will be matched zero or more times.
- “anything” that matches

```
> str_view_all(car, "ca*r") > str_view_all(apple, "ap*le")
```

car

cr

caar

caaar

caaaar

← Less “strict” than +
(match-one-or-more)

apple

applee

aple

appple

apleee

```
car <- c('car','cr','caar','caaar','caaaar')  
apple <- c('apple','applee','aple','appple','apleee')
```

- + (match-one-or-more)

- The preceding item will be matched one or more times.
- “at least one occurrence”
- Recommended to use when you know the text for sure and want to edit typos

```
> str_view_all(car, "ca+r") > str_view_all(apple, "ap+le")
```

car

cr

caar

caaar

caaaar

More “strict” than *
(match-zero-or-more)

apple

applee

aple

appple

apleee

Any cases containing
‘ap’ at front and ‘le’ at
end are possible

Regular Expression III: Group

Regular Expression

- []
 - Used for expressing text in groups
 - Each text in [] is considered separately
 - -: Continuous values
 - Effective when used with other regular expressions

```
text.1 <- "Pep Guardiola expressed his admiration for Erling Haaland's attitude  
after winning Man of the Match performance against Manchester United"  
vector.1 <- c("I like to drink Pepsi", "I love Manchester United",  
              "I have iphone", "iphone is very good")
```

a, b, c, d, e

```
> str_view_all(text.1, "[a-e]")  
> str_view_all(text.1, "[ale]")  
> str_view_all(vector.1, "[a-e]")  
> str_view_all(vector.1, "[ale]")
```

a, e

Pep Guardiola expressed his admiration for Erling Haaland's attitude after winning Man of the Match performance against Manchester United	
I like to drink Pepsi	I like to drink Pepsi
I love Manchester United	I love Manchester United
I have iphone	I have iphone
iphone is very good	iphone is very good

Regular Expression

- `[] + ^`

- If `^xxx` is used with `[]`, it indicates everything except `xxx`

```
text.1 <- "Pep Guardiola expressed his admiration for Erling Haaland's attitude
after winning Man of the Match performance against Manchester United"
vector.1 <- c("I like to drink Pepsi", "I love Manchester United",
              "I have iphone", "iphone is very good")
```

```
> str_view_all(text.1, "[^a-z]")
```

Everything except a, b, c, ..., z

```
Pep Guardiola expressed his admiration for Erling Haaland's
attitude after winning Man of the Match performance against
Manchester United
```

```
> str_view_all(text.1, "[^a-zA-Z]")
```

Everything except a, b, c, ..., A, B, ..., Z

```
Pep Guardiola expressed his admiration for Erling Haaland's
attitude after winning Man of the Match performance against
Manchester United
```

```
> str_view_all(vector.1, "[A-Z]") > str_view_all(vector.1, "[^A-Z]")
```

```
I like to drink Pepsi
I love Manchester United
I have iphone
iphone is very good
```

```
I like to drink Pepsi
I love Manchester United
I have iphone
iphone is very good
```

Exercise

- Write down codes that generates the following result

```
sample <- "abc ABC 123.!?\\"{\\}\n abcde aaa bacad .a.aa.aaa  
abbaab ababcbabcdcbabcde"
```

```
> str_view_all( )  
abc ABC 123 .!?\\({} abcde aaa bacad .a.aa.aaa abbaab ababcbabcdcbabcde  
> str_view_all( )  
abc ABC 123 .!?\\({} abcde aaa bacad .a.aa.aaa abbaab ababcbabcdcbabcde  
> str_view_all( )  
abc ABC 123 .!?\\({} abcde aaa bacad .a.aa.aaa abbaab ababcbabcdcbabcde  
> str_view_all( )  
abc ABC 123 .!?\\({} abcde aaa bacad .a.aa.aaa abbaab ababcbabcdcbabcde
```

Exercise

- Write down codes that generates the following result

```
sample <- "abc ABC 123.!?\\(\\)\\{\\}\\n abcde aaa bacad .a.aa.aaa  
abbaab ababcbababcdcbabcde"
```

```
> str_view_all([_____])
```

```
abc ABC 123 .!?\\(\\)\\{\\} abcde aaa bacad .a.aa.aaa abbaab ababcbababcdcbabcde
```

```
> str_view_all([_____])
```

```
abc ABC 123 .!?\\(\\)\\{\\} abcde aaa bacad .a.aa.aaa abbaab ababcbababcdcbabcde
```

```
> str_view_all([_____])
```

```
abc ABC 123 .!?\\(\\)\\{\\} abcde aaa bacad .a.aa.aaa abbaab ababcbababcdcbabcde
```

```
> str_view_all([_____])
```

```
abc ABC 123 .!?\\(\\)\\{\\} abcde aaa bacad .a.aa.aaa abbaab ababcbababcdcbabcde
```

Regular Expression IV: Multiple Text

- `[:lower:]`
 - Lower-case letters in the current locale.
- `[:upper:]`
 - Upper-case letters in the current locale.
- `[:alpha:]`
 - Alphabetic characters: `[:lower:]` and `[:upper:]`

```
> vector.1 %>%  
+   str_view_all("[:lower:]")
```

```
I like to drink Pepsi  
I love Manchester United  
I have iphone  
iphone is very good
```

```
> vector.1 %>%  
+   str_view_all("[:upper:]")
```

```
I like to drink Pepsi  
I love Manchester United  
I have iphone  
iphone is very good
```

```
> vector.1 %>%  
+   str_view_all("[:alpha:]")
```

```
I like to drink Pepsi  
I love Manchester United  
I have iphone  
iphone is very good
```

- `[[:digit:]]`

- Digits: '0 1 2 3 4 5 6 7 8 9'.

```
> c("one 2 3 four 5 six") %>%  
+   str_view_all("[[:digit:]]")
```

one 2 3 four 5 six

- `[[:xdigit:]]`

- Hexadecimal digits:
'0 1 2 3 4 5 6 7 8 9 A B C D E F a b c d e f'.

```
> c("one 2 3 four 5 six") %>%  
+   str_view_all("[[:xdigit:]]")
```

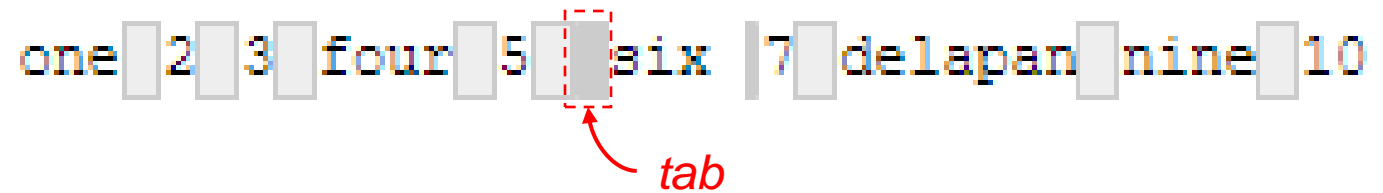
one 2 3 four 5 six

```
c("one 2 3 four 5    six  
7 delapan nine 10")
```

- `[:blank:]`

- Blank characters: space and tab, and possibly other locale-dependent characters, but on most platforms not including non-breaking space.

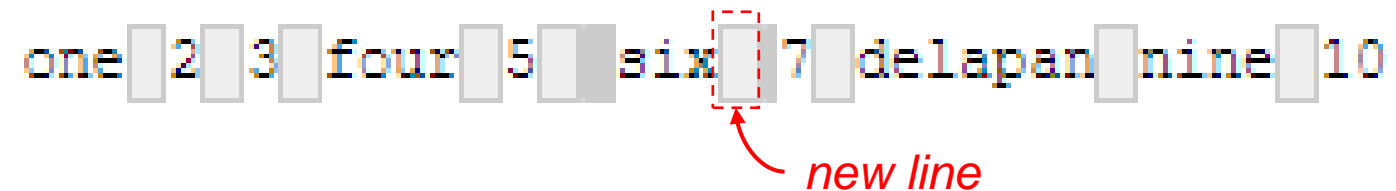
```
> c("one 2 3 four 5    six  
+   7 delapan nine 10") %>%  
+   str_view_all("[:blank:]")
```



- `[:space:]`

- Space characters: tab, newline, vertical tab, form feed, carriage return, space, and possibly other locale-dependent characters – on most platforms, this does not include non-breaking spaces

```
> c("one 2 3 four 5    six  
+   7 delapan nine 10") %>%  
+   str_view_all("[:space:]")
```



- `[:punct:]`
 - Punctuation characters:
'! " # \$ % & ' () * + , - . / : ; < = > ? @ [\] ^ _ ` { | } ~'
- `[:alnum:]`
 - Alphanumeric characters: `[:alpha:]` and `[:digit:]`
- `[:print:]`
 - Printable characters: `[:alnum:]`, `[:punct:]` and space
- `[:graph:]`
 - Graphical characters: `[:alnum:]` and `[:punct:]`

```
> c("'I love you.' heheh!@#") %>%  
+   str_view_all("[:punct:]")
```

```
'I love you.' heheh!@#
```

```
> c("'I love you.' heheh!@#") %>%  
+   str_view_all("[:alnum:]")
```

```
'I love you.' heheh!@#
```

```
> c("'I love you.' heheh!@#") %>%  
+   str_view_all("[:print:]")
```

```
'I love you.' heheh!@#
```

```
> c("'I love you.' heheh!@#") %>%  
+   str_view_all("[:graph:]")
```

```
'I love you.' heheh!@#
```

Not include space

Regular Expression V: Range

- {n}

- The preceding item is matched exactly n times.

```
> c("'I love you.' heheh!@#") %>%  
+   str_view_all("[:punct:]{1}")  
> c("'I love you.' heheh!@#") %>%  
+   str_view_all("[:punct:]{2}")  
> c("'I love you.' heheh!@#") %>%  
+   str_view_all("[:punct:]{3}")
```

Diagram illustrating the results of `str_view_all` for the regular expression `{n}` applied to the string `"'I love you.' heheh!@#"`:

- $n=1$: Matches the first punctuation character `'`.
- $n=2$: Matches the first two punctuation characters `.'`.
- $n=3$: Matches the first three punctuation characters `!@#`.

- {n,}

- The preceding item is matched n or more times.

```
> c("'I love you.' heheh!@#") %>%  
+   str_view_all("[:alnum:]{1,}")  
> c("'I love you.' heheh!@#") %>%  
+   str_view_all("[:alnum:]{3,}")  
> c("'I love you.' heheh!@#") %>%  
+   str_view_all("[:alnum:]{4,}")
```

Diagram illustrating the results of `str_view_all` for the regular expression `{n,}` applied to the string `"'I love you.' heheh!@#"`:

- $n \geq 1$: Matches the first alphanumeric character `I`.
- $n \geq 3$: Matches the first three alphanumeric characters `love`.
- $n \geq 4$: Matches the first four alphanumeric characters `love you.`

- {n,m}

- The preceding item is matched at least n times, but not more than m times.

```
> c("'I love you.' heheh!@#") %>%  
+   str_view_all("[:alnum:]{1,1}")  
> c("'I love you.' heheh!@#") %>%  
+   str_view_all("[:alnum:]{1,3}")  
> c("'I love you.' heheh!@#") %>%  
+   str_view_all("[:alnum:]{2,2}")
```

$1 \leq n \leq 1$ → 'I love you.' heheh!@#

$1 \leq n \leq 3$ → 'I love you.' heheh!@#

$2 \leq n \leq 2$ → 'I love you.' heheh!@#

- Regular Expression

- \s: white space

```
'I love you.' heheh!@#
```

- \S: all text except white space

```
'I love you.' heheh!@#
```

- \w: alphanumeric + '_'

```
'I love you.' heheh!@#
```

- \W: non-alphanumeric + without '_'

```
'I love you.' heheh!@#
```

- \d: numerics

```
'I love you.' heheh!@#
```

- \D: all text except numerics

```
'I love you.' heheh!@#
```

!+ !+ Regular Expression

```
> c("'I love you.' heheh!@#") %>%  
+   str_view_all("\\s")  
> c("'I love you.' heheh!@#") %>%  
+   str_view_all("\\S")  
> c("'I love you.' heheh!@#") %>%  
+   str_view_all("\\w")  
> c("'I love you.' heheh!@#") %>%  
+   str_view_all("\\W")  
> c("'I love you.' heheh!@#") %>%  
+   str_view_all("\\d")  
> c("'I love you.' heheh!@#") %>%  
+   str_view_all("\\D")
```


Regular Expression Exercise

- From Trump's speech, find texts that are not spoken by the person
 - [Crowd chants: "We love you"], [Melania Trump], [Donald Trump]

```
> trump.speech.sum$text %>%
```

[1] NA	"[Crowd chants: "We love you"]"
[3] NA	NA
[5] NA	NA
[7] NA	"[Melania Trump]"
[9] "[Donald Trump]"	NA
[11] NA	NA

- From Trump's speech, find texts that are not spoken by the person
 - Steps

Knowing the fact that texts that are not spoken by the person starts with '[' → '['

```
> trump.speech.sum$text %>%
+ str_extract('\\[[[:alpha:]]')
[1] NA "[C" NA NA NA NA NA "[M" "[D" NA
[18] NA NA NA NA NA NA NA NA NA NA NA
```

[[:alpha:]] → a-zA-Z

```
> trump.speech.sum$text %>%
+ str_extract('\\[[[:alpha:]]+')
[1] NA "[Crowd" NA NA NA
[8] "[Melania" "[Donald" NA NA NA
[15] NA NA NA NA NA
[22] NA NA NA NA NA
[29] NA
```

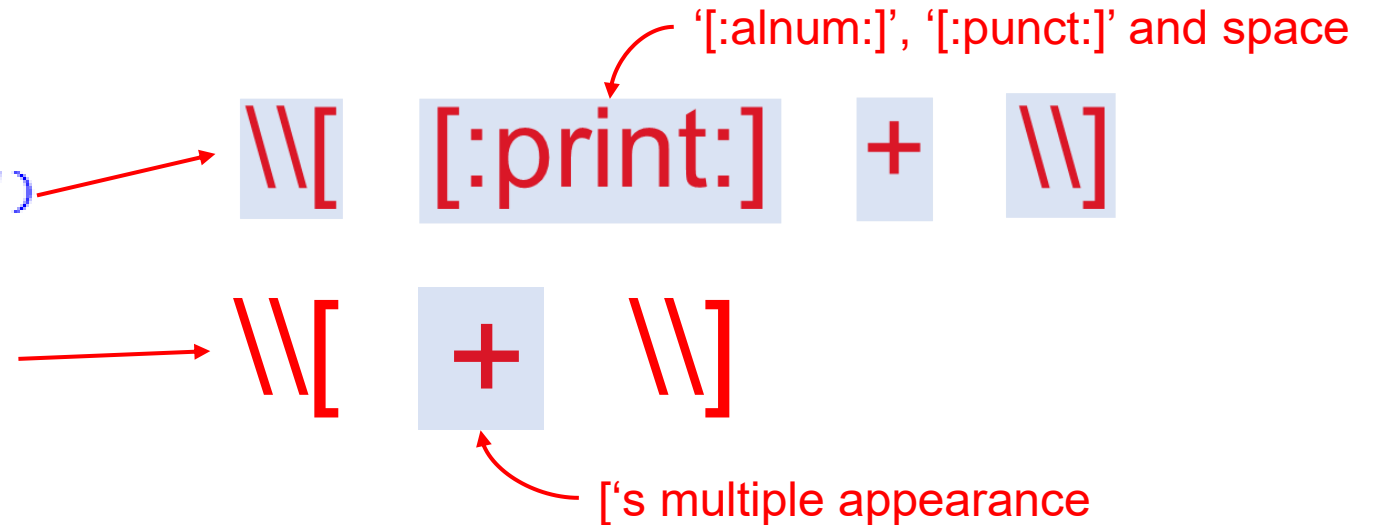
[[:alpha:]]+ → a-zA-Z & one-more-times

```
> trump.speech.sum$text %>%
+ str_extract('\\[[[:alpha:]]{1,}'))
[1] NA "[Crowd" NA NA NA
[8] "[Melania" "[Donald" NA NA NA
[15] NA NA NA NA NA
[22] NA NA NA NA NA
[29] NA
```

[[:alpha:]]{1,} → a-zA-Z & one-more-times

- From Trump's speech, find texts that are not spoken by the person
 - [Crowd chants: "We love you"], [Melania Trump], [Donald Trump]

```
> "[Donald Trump]" %>%
+ str_extract('\\[[:print:]+\\]')
[1] "[Donald Trump]"
> "[Donald Trump]" %>%
+ str_extract('\\[+\\]')
[1] NA
```



```
> "[[[[[[[Donald Trump]]]]]]]" %>%
+ str_extract('\\[+')
[1] "[[[[[[[["
> "[[[[[[[Donald Trump]]]]]]]" %>%
+ str_extract('\\[+\\]')
[1] "]]]]]]]"
```

```
> c("caaaabr", "caar", "ca!r") %>%
+ str_extract('ca*r')
[1] NA "caar" NA
> c("caaaabr", "caar", "ca!r") %>%
+ str_extract('ca*!')
[1] NA NA "ca!"
```

- Write down the desired result

(1) `c("[DonaldTrump]", "[Donald Trump]", "[!@#Donald Trump]") %>%
 str_extract('[:punct:]+')`

(2) `c("[DonaldTrump]", "[Donald Trump]", "[!@#Donald Trump]") %>%
 str_extract('\\[[:punct:]]+')`

(3) `c("[DonaldTrump]", "[Donald Trump]", "[!@#Donald Trump]") %>%
 str_extract('\\[[:alpha:]]+\\')`

(4) `c("[DonaldTrump]", "[Donald Trump]", "[!@#Donald Trump]") %>%
 str_extract('\\[[:print:]]+\\')`

(1)

(2)

(3)

(4)

Text Mining PDF

- `pdftools::pdf_text()`
 - Extracts text from PDF file
 - A character vector with extracted texts for each page is returned

```
> library(pdftools)
> mat10 <- pdf_text("R data/Matthew10.pdf")
> mat10
[1] "1\n Jesus called his twelve disciples to him and gave them authority to drive out impure\nspirits and to
heal every disease and sickness.\n\n2\n These are the names of the twelve apostles: first, Simon (who is call
ed Peter) and his\nbrother Andrew; James son of Zebedee, and his brother John; 3 Philip and\nBartholomew; Thom
as and Matthew the tax collector; James son of Alphaeus, and\nThaddaeus; 4 Simon the Zealot and Judas Iscario
t, who betrayed him.\n\n5\n These twelve Jesus sent out with the following instructions: "Do not go among the
\nGentiles or enter any town of the Samaritans. 6 Go rather to the lost sheep of\nIsrael. 7 As you go, proclai
m this message: 'The kingdom of heaven has come\nnear.' 8 Heal the sick, raise the dead, cleanse those who hav
e leprosy, drive out\ndemons. Freely you have received; freely give.\n\n9\n "Do not get any gold or silver or
copper to take with you in your belts- 10 no bag\nfor the journey or extra shirt or sandals or a staff, for t
he worker is worth his\nkeep. 11 Whatever town or village you enter, search there for some worthy person\nand
```

- Import Matthew10.pdf
 - Two pages
 - Two elements

```
> mat10
```

[1] "1\n Jesus called his twelve disciples to him and gave them authority to drive out impure\nspirits and to heal every disease and sickness.\n\n2\n These are the names of the twelve apostles: first, Simon (who is called Peter) and his\nbrother Andrew; 3 James son of Zebedee, and his brother John; 4 Philip and\nBartholomew; Thomas and Matthew the tax collector; James son of Alphaeus, and\nThaddaeus; 5 Simon the Zealot and Judas Iscariot, who betrayed him.\n\n6\n These twelve Jesus sent out with the following instructions: "Do not go among the\nGentiles or enter any town of the Samaritans. 7 Go rather to the lost sheep of\nIsrael. 8 As you go, proclaim this message: 'The kingdom of heaven has come\nnear.' 9 Heal the sick, raise the dead, cleanse those who have leprosy, drive out\ndemons. Freely you have received; freely give.\n\n10\n "Do not get any gold or silver or copper to take with you in your belts- 11 no bag\nfor the journey or extra shirt or sandals or a staff, for the worker is worth his\nkeep. 12 Whatever town or village you enter, search there for some worthy person\nand stay at their house

Page 1

¹ Jesus called his twelve disciples to him and gave them authority to drive out impure spirits and to heal every disease and sickness.

² These are the names of the twelve apostles: first, Simon (who is called Peter) and his brother Andrew; James son of Zebedee, and his brother John; ³ Philip and Bartholomew; Thomas and Matthew the tax collector; James son of Alphaeus, and Thaddaeus; ⁴ Simon the Zealot and Judas Iscariot, who betrayed him.

⁵ These twelve Jesus sent out with the following instructions: "Do not go among the Gentiles or enter any town of the Samaritans. ⁶ Go rather to the lost sheep of Israel. ⁷ As you go, proclaim this message: The kingdom of heaven has come near.' ⁸ Heal the sick, raise the dead, cleanse those who have leprosy, drive out demons. Freely you have received; freely give.

9 "Do not get any gold or silver or copper to take with you in your belts— 10 no bag for the journey or extra shirt or sandals or a staff, for the worker is worth his keep. 11 Whatever town or village you enter, search there for some worthy person and stay at their house until you leave. 12 As you enter the home, give it your greeting. 13 If the home is deserving, let your peace rest on it; if it is not, let your peace return to you. 14 If anyone will not welcome you or listen to your words, leave that home or town and shake the dust off your feet. 15 Truly I tell you, it will be more bearable for Sodom and Gomorrah on the day of judgment than for that town.

16 "I am sending you out like sheep among wolves. Therefore be as shrewd as snakes and as innocent as doves. 17 Be on your guard; you will be handed over to the local councils and be flogged in the synagogues. 18 On my account you will be brought before governors and kings as witnesses to them and to the Gentiles. 19 But when they arrest you, do not worry about what to say or how to say it. At that time you will be given what to say, 20 for it will not be you speaking, but the Spirit of your Father speaking through you.

²¹ "Brother will betray brother to death, and a father his child; children will rebel against their parents and have them put to death. ²² You will be hated by everyone because of me, but the one who stands firm to the end will be saved. ²³ When you are persecuted in one place, flee to another. Truly I tell you, you will not finish going through the towns of Israel before the Son of Man comes.

²⁴ "The student is not above the teacher, nor a servant above his master. ²⁵ It is enough for students to be like their teachers, and servants like their masters. If the head of the house has been called Beelzebul, how much more the members of his household!

²⁶ "So do not be afraid of them, for there is nothing concealed that will not be disclosed, or hidden that will not be made known. ²⁷ What I tell you in the dark,

Page 2

speak in the daylight; what is whispered in your ear, proclaim from the roofs. ²⁸ Do not be afraid of those who kill the body but cannot kill the soul. Rather, be afraid of the One who can destroy both soul and body in hell. ²⁹ Are not two sparrows sold for a penny? Yet not one of them will fall to the ground outside your Father's care. ³⁰ And even the very hairs of your head are all numbered. ³¹ So don't be afraid; you are worth more than many sparrows.

³² "Whoever acknowledges me before others, I will also acknowledge before my Father in heaven. ³³ But whoever disowns me before others, I will disown before my Father in heaven.

³⁴ "Do not suppose that I have come to bring peace to the earth. I did not come to bring peace, but a sword. ³⁵ For I have come to turn

"a man against his father,
 a daughter against her mother,
 a daughter-in-law against her mother-in-law—
 36 a man's enemies will be the members of his own household."

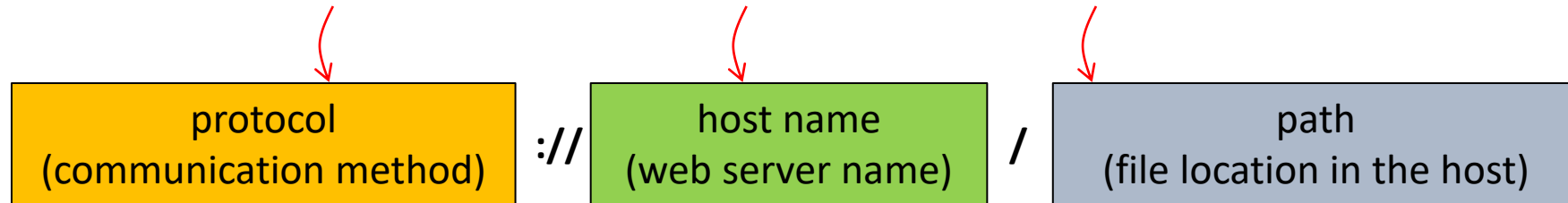
37 "Anyone who loves their father or mother more than me is not worthy of me; anyone who loves their son or daughter more than me is not worthy of me. 38 Whoever does not take up their cross and follow me is not worthy of me. 39 Whoever finds their life will lose it, and whoever loses their life for my sake will find it.

40 "Anyone who welcomes you welcomes me, and anyone who welcomes me welcomes the one who sent me. 41 Whoever welcomes a prophet as a prophet will receive a prophet's reward, and whoever welcomes a righteous person as a righteous person will receive a righteous person's reward. 42 And if anyone gives even a cup of cold water to one of these little ones who is my disciple, truly I tell you, that person will certainly not lose their reward."

Text Mining Web Data

- What is the Web?
 - Abbreviation of the World Wide Web (a.k.a. WWW)
 - Information space where web resources (documents, images, videos, etc.) can be accessed via the Internet
 - Uniform Resource Locators (URLs) are used to uniquely identify the resources

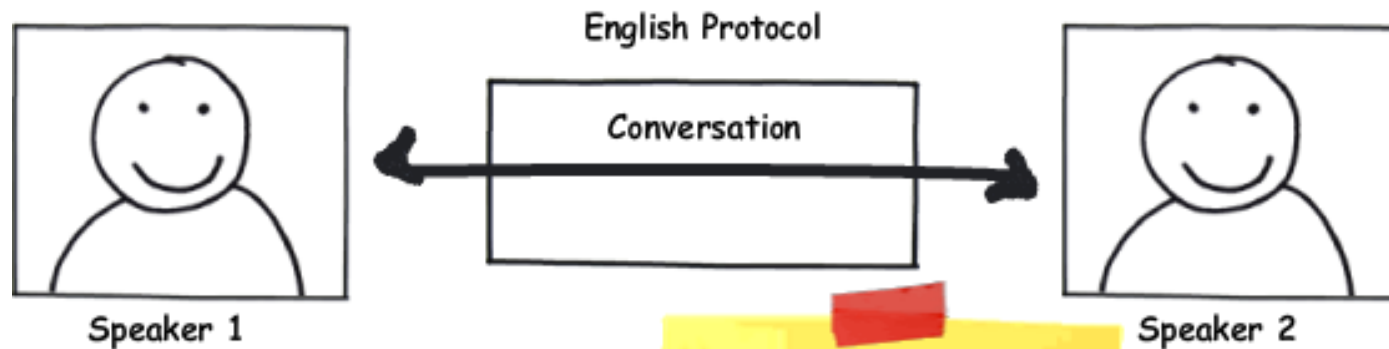
<http://www.example.com/index.html>



- Web browser
 - Applications that support the use of web documents
 - Supports graphical interface support
 - Translates html document

- Protocol

- A set of rules that governs data communications between the network devices
- Without a protocol, two devices may be connected but not communicating with each other.
- Hypertext Transfer Protocol (HTTP): A Protocol to exchange or transfer hypertext



Both of the speaker use english language to communicate each other. Conceptually a language can be considered a protocol where the rules are the grammar rules

Examples

HTTP: Hypertext Transfer Protocol
HTTPS: HTTP Secure
FTP: File Transfer Protocol
SMTP: Simple Mail Transfer Protocol

- Hypertext Markup Language (HTML)
 - Standard markup language for writing documents designed to be displayed in a web browser.
 - Hypertext (ex. Hyperlink) + Markup language (Google colab, Jupyter Notebook)

- When we read, how do we understand?
 - Without any explanation, we can quickly notice few things.
 - Chapter title
 - Paragraphs
 - Contents, etc.
- We can understand because we all know these “rules”

Paragraph

Chapter 1 Global value chains: the face of 21st-century international commerce

Technology, business innovations and falling trade costs have profoundly transformed the organization of global production. The production process has been unbundled, and different production stages spread across different locations. Complex international supply chains – also referred to as global value chains – have emerged, whereby firms ship intermediate goods across the world for further processing and, eventually, final assembly. Among the most far-reaching changes unleashed by the growth of global value chains has been the integration of selected developing economies into the global economy, coinciding with rapid economic growth in those economies. One prominent scholar has characterized this development as “perhaps the most momentous global economic change in the last 100 years.”¹

The rise of global value chains has gone hand in hand with the growing importance of intangible assets in economic activity. Previous editions of the *World Intellectual Property Report* have documented the rapid growth of investments in technology, design and branding – outpacing the growth of traditional bricks-and-mortar investments.² In fact, the two trends are directly connected. Intangible assets shape global value chains in two important ways. First, the organization of international supply chains – and especially the offshoring of labor-intensive manufacturing tasks to lower-wage economies – entails the transfer of technological and business knowledge from one location to another. Such knowledge is often subject to various forms of intellectual property (IP), including registered

Despite a large number of studies on global value chain trade, relatively little is known about how companies manage their intangible assets when offshoring production abroad, and how much production value derives from those assets. This report seeks to help fill that knowledge gap. It does so in two parts. First, it distills the insights from existing global value chain studies and reveals original research on the macroeconomic contribution of intangible assets to value added. Second, it explores the role of intangible assets at the microeconomic level in the case of three industries – coffee, photovoltaics and smartphones. These case studies will be presented in chapters 2, 3 and 4, respectively.

This opening chapter seeks to set the scene by reviewing how global value chains have come about, exploring economic research on their organization and providing new evidence on the contribution of intangible assets. In particular, section 1.1 provides a brief summary of the growth of global value chains over recent decades and section 1.2 introduces key concepts surrounding the organization and governance of global value chains. Against this background, section 1.3 presents original estimates of the returns accruing to intangible assets in global value chain production. Section 1.4 then takes a closer look at how firms participating in global value chains manage their intangible assets, and how firms in economies at early stages of industrial development may acquire them. This discussion provides the context for the case studies in chapters 2, 3 and 4. Finally, section 1.5 offers some policy-oriented reflections on the evolution of global value chains.

Name and
number of
Chapter

Paragraph

- When we visit website, how do we read and understand?
 - Website itself does not have much a big difference in understanding
 - Computer recognizes the websites differently, with its own “rule”.
 - If we know this rule, we can read and create the website with computer’s perspective.



지금 일어나고 있는 일

오늘 트위터에 가입하세요.

 Google에서 가입하기

 Apple에서 가입하기

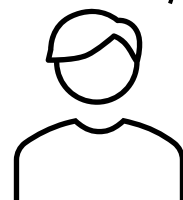
휴대폰 번호나 이메일 주소로 가입하기

By signing up, you agree to the [Terms of Service](#) and [Privacy Policy](#).

이미 계정이 있으세요? [로그인하기](#)

소개 고객센터 이윤관리 개인정보처리방침 쿠팡정책 광고정보 불로그 상태 채용 브랜드리소스 쿠팡 마케팅 비즈니스를드워드 개발자 디렉터리

How we see
and
understand



```
<!DOCTYPE html>
<html dir="ltr" lang="ko" style="overflow-y: scroll; overscroll-behavior-y: none; font-size: 15px;">
  <head>...</head>
  <body style="background-color: #FFFFFF;">
    <noscript>...</noscript>
    <div id="react-root" style="height:100%;display:flex;">
      <div class="css-1dbjc4n r-13awgt0 r-12vffkv">
        <div class="css-1dbjc4n r-13awgt0 r-12vffkv">
          <div class="css-1dbjc4n r-13qz1uu r-417010" aria-hidden="false" style="min-height: 849px;">
            <main role="main" class="css-1dbjc4n r-16y2uox r-1wbh5a2">
              <div class="css-1dbjc4n r-150rngu r-16y2uox r-1wbh5a2"> <flex>
                <div class="css-1dbjc4n r-13awgt0"> <flex>
                  ...
                  <div class="css-1dbjc4n r-tv6buo"> <flex> == $0
                    <div class="css-1dbjc4n r-1777fci r-ywje51 r-1ye8kvj r-1qmwkhh r-nsbfu8 r-13qz1uu"> <flex>
                      <div class="css-1dbjc4n r-1pcd215 r-13qz1uu"> <flex>
                        <svg viewBox="0 0 24 24" aria-hidden="true" class="r-k200y r-1cv12hr r-4qtq9 r-yyyyoo r-55kf15 r-dnmrzs r-kzkbwu r-bnwqim r-1plcui r-1rvibr">...</svg>
                        <div dir="auto" class="css-901oao r-18jsvk2 r-fm7h5w r-19oahor r-b88u0q r-nm9kes r-1ncnk10 r-4afqvc r-bcqeeo r-qvut0">...</div>
                        <div dir="auto" class="css-901oao r-18jsvk2 r-fm7h5w r-1jyipy1 r-b88u0q r-ueyrd6 r-zd98yo r-bcqeeo r-qvut0">...</div>
                        <div class="css-1dbjc4n">...</div> <flex>
                      </div>
                    </div>
                    <div class="css-1dbjc4n r-1dhjt1p r-13awgt0 r-1777fci r-12zvaga r-1udh08x r-t60dpp">...</div> <flex>
                    </div>
                    <nav aria-label="바닥글" role="navigation" class="css-1dbjc4n r-18u37iz r-1w6e6rj r-1777fci r-ymttW5 r-1fisjgu">...</nav> <flex>
                    </div>
                  </div>
                </main>
              </div>
            </body>
          </html>
```


- Tag
 - Like a block of building

- Two types of tag

- Tag with content: `<tag> content </tag>`

`<h1> Hello world </h1>`

`<p> I like this! </p>`

Start tag

End tag

- Tag without contents: `<tag>`

`
`

``

Bigger text?

`<h1> ~~ </h1>`

Paragraph?

`<p> ~~~ </p>`

Chapter 1 Global value chains: the face of 21st-century international commerce

Technology, business innovations and falling trade costs have profoundly transformed the organization of global production. The production process has been unbundled, and different production stages spread across different locations. Complex international supply chains – also referred to as global value chains – have emerged, whereby firms ship intermediate goods across the world for further processing and, eventually, final assembly. Among the most far-reaching changes unleashed by the growth of global value chains has been the integration of selected developing economies into the global economy, coinciding with rapid economic growth in those economies. One prominent scholar has characterized this development as “perhaps the most momentous global economic change in the last 100 years.”¹

The rise of global value chains has gone hand in hand with the growing importance of intangible assets in economic activity. Previous editions of the *World Intellectual Property Report* have documented the rapid growth of investments in technology, design and branding – outpacing the growth of traditional bricks-and-mortar investments.² In fact, the two trends are directly connected. Intangible assets shape global value chains in two important ways. First, the organization of international supply chains – and especially the offshoring of labor-intensive manufacturing tasks to lower-wage economies – entails the transfer of technological and business knowledge from one location to another. Such knowledge is often subject to various forms of intellectual property (IP), including registered

Despite a large number of studies on global value chain trade, relatively little is known about how companies manage their intangible assets when offshoring production abroad, and how much production value derives from those assets. This *report* seeks to help fill that knowledge gap. It does so in two parts. First, it distills the insights from existing global value chain studies and reveals original research on the macroeconomic contribution of intangible assets to value added. Second, it explores the role of intangible assets at the microeconomic level in the case of three industries – coffee, photovoltaics and smartphones. These case studies will be presented in chapters 2, 3 and 4, respectively.

This opening chapter seeks to set the scene by reviewing how global value chains have come about, exploring economic research on their organization and providing new evidence on the contribution of intangible assets. In particular, section 1.1 provides a brief summary of the growth of global value chains over recent decades and section 1.2 introduces key concepts surrounding the organization and governance of global value chains. Against this background, section 1.3 presents original estimates of the returns accruing to intangible assets in global value chain production. Section 1.4 then takes a closer look at how firms participating in global value chains manage their intangible assets, and how firms in economies at early stages of industrial development may acquire them. This discussion provides the context for the case studies in chapters 2, 3 and 4. Finally, section 1.5 offers some policy-oriented reflections on the evolution of global value chains.

- Attribute
 - Additional information assigned to a tag
- Example
 - ``
 - Name of attribute* (points to `src`)
 - Attribute value* (points to `"cat.png"`)
 - `<h1 title="HGU"> Handong Global University </h1>`

Want to add link?
Change color or
font type? Etc.

Chapter 1 Global value chains: the face of 21st-century international commerce

Technology, business innovations and falling trade costs have profoundly transformed the organization of global production. The production process has been unbundled, and different production stages spread across different locations. Complex international supply chains – also referred to as global value chains – have emerged, whereby firms ship intermediate goods across the world for further processing and, eventually, final assembly. Among the most far-reaching changes unleashed by the growth of global value chains has been the integration of selected developing economies into the global economy, coinciding with rapid economic growth in those economies. One prominent scholar has characterized this development as “perhaps the most momentous global economic change in the last 100 years.”¹

The rise of global value chains has gone hand in hand with the growing importance of intangible assets in economic activity. Previous editions of the *World Intellectual Property Report* have documented the rapid growth of investments in technology, design and branding – outpacing the growth of traditional bricks-and-mortar investments.² In fact, the two trends are directly connected. Intangible assets shape global value chains in two important ways. First, the organization of international supply chains – and especially the offshoring of labor-intensive manufacturing tasks to lower-wage economies – entails the transfer of technological and business knowledge from one location to another. Such knowledge is often subject to various forms of intellectual property (IP), including registered

Despite a large number of studies on global value chain trade, relatively little is known about how companies manage their intangible assets when offshoring production abroad, and how much production value derives from those assets. This *report* seeks to help fill that knowledge gap. It does so in two parts. First, it distills the insights from existing global value chain studies and reveals original research on the macroeconomic contribution of intangible assets to value added. Second, it explores the role of intangible assets at the microeconomic level in the case of three industries – coffee, photovoltaics and smartphones. These case studies will be presented in chapters 2, 3 and 4, respectively.

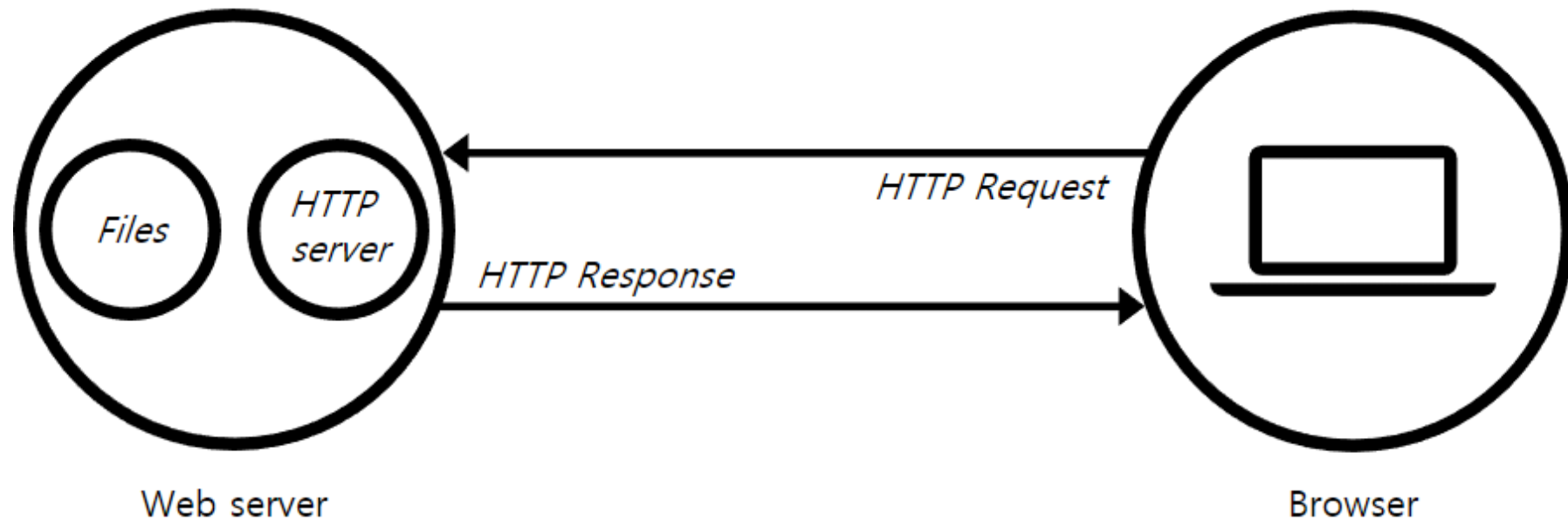
This opening chapter seeks to set the scene by reviewing how global value chains have come about, exploring economic research on their organization and providing new evidence on the contribution of intangible assets. In particular, section 1.1 provides a brief summary of the growth of global value chains over recent decades and section 1.2 introduces key concepts surrounding the organization and governance of global value chains. Against this background, section 1.3 presents original estimates of the returns accruing to intangible assets in global value chain production. Section 1.4 then takes a closer look at how firms participating in global value chains manage their intangible assets, and how firms in economies at early stages of industrial development may acquire them. This discussion provides the context for the case studies in chapters 2, 3 and 4. Finally, section 1.5 offers some policy-oriented reflections on the evolution of global value chains.

- Developing a web is like writing an e-book with the computer-friendly language
- Web is operated on the computer system. In other words, we need to learn the language that computer can understand

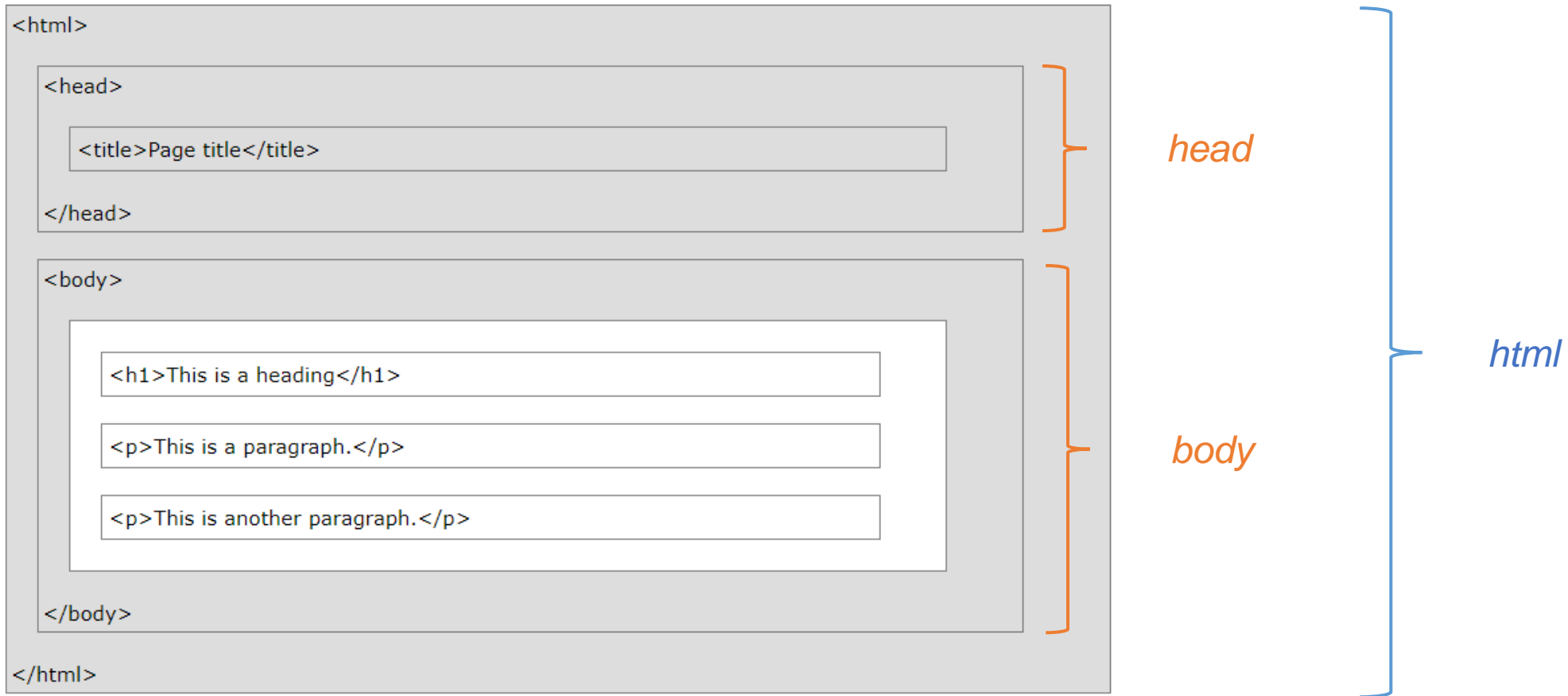
```
2 <body>
3   <h1>My Books</h1>
4   <section>
5     <h2>Green Eggs and Ham</h2>
6     <p>Great eggs and even greater ham.</p>
7   </section>
8   <section>
9     <h2>Don Quixote</h2>
10    <p>Gallant imitator of protagonists worldwide.</p>
11  </section>
12  <section>
13    <h2>The Stand</h2>
14    <p>Everyone loves a good apocalypse story.</p>
15  </section>
16 </body>
17
```

<https://www.codingdojo.com/blog/html5-examples>

- Mechanism
 - Writing HTML documents (coding)
 - Web browser
 - Web server
 - Website hosting server providing services



- HTML Structure



- `<html>` tag
 - Defines the root of a HTML document.
- `<head>` tag
 - Container for metadata about the document like its title, styles, scripts, etc.
 - In HTML5, it can be omitted.
- `<body>` tag
 - Contains all the contents of the document like text, images, tables, etc.

*Data about
the data*



- Step 1) Get Web URL
 - `rvest::read_html("url")`
- Step 2) Choose specific tag
 - `rvest::html_nodes("tag or tag.class or tag#id")`
- Step 3) Get text
 - `rvest::html_text()` → get text
 - `rvest::html_attr("name of attribute")` → used when to get hyperlink