Unstructured Data Analysis

HAZIM FITRI

2025-04-13

Contents

Part 1: Extract Text from Files

```
eg1<-read.table("Attachment 1/GC.txt",fill=T,header=F) #Data CG.txt
eg1[1,]
```

```
##         V1   V2  V3    V4   V5 V6         V7 V8 V9 V10 V11 V12 V13 V14 V15 V16
## 1 Gertrude Cox, The First Lady Of Statistics
##   V17 V18 V19 V20 V21 V22 V23 V24 V25
## 1
```

```
eg2<-read.csv("Attachment 1/GC.csv",header=F) #Data GC.csv
eg2[1,]
```

```
## [1] "Gertrude Cox, The First Lady Of Statistics"
```

```
#Using tm package
library(tm)# text mining
```

```
## Warning: package 'tm' was built under R version 4.4.3
```

```
## Loading required package: NLP
```

```
## Warning: package 'NLP' was built under R version 4.4.2
```

```
eg3<-c("Hi!","Welcome to STQD6114","Tuesday, 11-1pm")
mytext<-VectorSource(eg3)
mycorpus<-VCorpus(mytext)
inspect(mycorpus)
```

```
## <<VCorpus>>
## Metadata:  corpus specific: 0, document level (indexed): 0
## Content:  documents: 3
##
## [[1]]
## <<PlainTextDocument>>
```

```
## Metadata:  7
## Content:  chars: 3
##
## [[2]]
## <<PlainTextDocument>>
## Metadata:  7
## Content:  chars: 19
##
## [[3]]
## <<PlainTextDocument>>
## Metadata:  7
## Content:  chars: 15
```

```
as.character(mycorpus[[1]])
```

```
## [1] "Hi!"
```

```
#Example using VectorSource
eg4<-t(eg1) #From example 1
a<-sapply(1:7,function(x) # 1:7 can be changed into "ncol(eg4)"
  trimws(paste(eg4[,x],collapse=" "),"right"))
mytext<-VectorSource(a)
mycorpus<-VCorpus(mytext)
inspect(mycorpus)
```

```
## <<VCorpus>>
## Metadata:  corpus specific: 0, document level (indexed): 0
## Content:  documents: 7
##
## [[1]]
## <<PlainTextDocument>>
## Metadata:  7
## Content:  chars: 42
##
## [[2]]
## <<PlainTextDocument>>
## Metadata:  7
## Content:  chars: 52
##
## [[3]]
## <<PlainTextDocument>>
## Metadata:  7
## Content:  chars: 118
##
## [[4]]
## <<PlainTextDocument>>
## Metadata:  7
## Content:  chars: 153
##
## [[5]]
## <<PlainTextDocument>>
## Metadata:  7
## Content:  chars: 100
##
## [[6]]
## <<PlainTextDocument>>
## Metadata:  7
## Content:  chars: 119
##
## [[7]]
## <<PlainTextDocument>>
## Metadata:  7
## Content:  chars: 147
```

```r
as.character(mycorpus[[1]])
```

```
## [1] "Gertrude Cox, The First Lady Of Statistics"
```

```r
#Example using DirSource
mytext<-DirSource("movies")
mycorpus<-VCorpus(mytext)
inspect(mycorpus)
```

```
## <<VCorpus>>
## Metadata:  corpus specific: 0, document level (indexed): 0
## Content:  documents: 5
##
## [[1]]
## <<PlainTextDocument>>
## Metadata:  7
## Content:  chars: 502
##
## [[2]]
## <<PlainTextDocument>>
## Metadata:  7
## Content:  chars: 400
##
## [[3]]
## <<PlainTextDocument>>
## Metadata:  7
## Content:  chars: 506
##
## [[4]]
## <<PlainTextDocument>>
## Metadata:  7
## Content:  chars: 387
##
## [[5]]
## <<PlainTextDocument>>
## Metadata:  7
## Content:  chars: 385
```

```r
as.character(mycorpus[[1]])
```

```
## [1] "Spider-Man: Homecoming"
## [2] ""
## [3] "Thrilled by his experience with the Avengers, young Peter Parker returns home to live with his Aunt May. U
```

```r
#Example using DataFrameSource
eg5<-read.csv("Attachment 1/Doc6.csv",header=F) #Using doc6.csv
docs<-data.frame(doc_id=c("doc_1","doc_2"),
                 text=c(as.character(eg5[1,]),
                        as.character(eg5[2,])),
                 dmeta1=1:2,dmeta2=letters[1:2],
                 stringsAsFactors=F)
mytext<-DataframeSource(docs)
mycorpus<-VCorpus(mytext)
inspect(mycorpus)
```

```
## <<VCorpus>>
## Metadata:  corpus specific: 0, document level (indexed): 2
## Content:  documents: 2
##
## [[1]]
```

```
## <<PlainTextDocument>>
## Metadata:  7
## Content:  chars: 15
##
## [[2]]
## <<PlainTextDocument>>
## Metadata:  7
## Content:  chars: 413
```

```
as.character(mycorpus[[1]])
```

```
## [1] "Despicable Me 3"
```

```
eg5b = read.csv("Attachment 1/GC.csv", header=F)
```

Part 2: Web Scrapping

HTML Cheat Sheet

```
eg6<-readLines("https://en.wikipedia.org/wiki/Data_science")
```

```
## Warning in readLines("https://en.wikipedia.org/wiki/Data_science"): incomplete
## final line found on 'https://en.wikipedia.org/wiki/Data_science'
```

```
eg6[grep("\\h2",eg6)]
```

```
## [1] "\t<h2 class=\"vector-pinnable-header-label\">Contents</h2>"
## [2] "<div class=\"mw-heading mw-heading2\"><h2 id=\"Foundations\">Foundations</h2><span class=\"mw-editsection\
## [3] "<div class=\"mw-heading mw-heading2\"><h2 id=\"Etymology\">Etymology</h2><span class=\"mw-editsection\"><sp
## [4] "<div class=\"mw-heading mw-heading2\"><h2 id=\"Data_science_and_data_analysis\">Data science and data analy
## [5] "<div class=\"mw-heading mw-heading2\"><h2 id=\"Cloud_computing_for_data_science\">Cloud computing for data
## [6] "<div class=\"mw-heading mw-heading2\"><h2 id=\"Ethical_consideration_in_data_science\">Ethical consideratio
## [7] "<div class=\"mw-heading mw-heading2\"><h2 id=\"See_also\">See also</h2><span class=\"mw-editsection\"><spa
## [8] "<div class=\"mw-heading mw-heading2\"><h2 id=\"References\">References</h2><span class=\"mw-editsection\">
```

```
eg6[grep("\\p",eg6)] #paragraph
```

```
##   [1] "<html class=\"client-nojs vector-feature-language-in-header-enabled vector-feature-language-in-main-page
##   [2] "<title>Data science - Wikipedia</title>"
##   [3] "<script>(function(){var className=\"client-js vector-feature-language-in-header-enabled vector-feature-la
##   [4] "RLSTATE={\"ext.globalCssJs.user.styles\":\"ready\",\"site.styles\":\"ready\",\"user.styles\":\"ready\",\
##   [5] "<script>(RLQ=window.RLQ||[]).push(function(){mw.loader.impl(function(){return[\"user.options@12s5i\",func
##   [6] "}];});});</script>"
##   [7] "<link rel=\"stylesheet\" href=\"/w/load.php?lang=en&amp;modules=ext.cite.styles%7Cext.uls.interlanguage%7
##   [8] "<script async=\"\" src=\"/w/load.php?lang=en&amp;modules=startup&amp;only=scripts&amp;raw=1&amp;skin=vect
##   [9] "<link rel=\"stylesheet\" href=\"/w/load.php?lang=en&amp;modules=site.styles&amp;only=styles&amp;skin=vect
##  [10] "<meta name=\"robots\" content=\"max-image-preview:standard\">"
##  [11] "<meta name=\"format-detection\" content=\"telephone=no\">"
##  [12] "<meta property=\"og:image\" content=\"https://upload.wikimedia.org/wikipedia/commons/thumb/4/45/PIA23792-
##  [13] "<meta property=\"og:image:width\" content=\"1200\">"
##  [14] "<meta property=\"og:image:height\" content=\"900\">"
##  [15] "<meta property=\"og:image\" content=\"https://upload.wikimedia.org/wikipedia/commons/thumb/4/45/PIA23792-
##  [16] "<meta property=\"og:image:width\" content=\"800\">"
##  [17] "<meta property=\"og:image:height\" content=\"600\">"
##  [18] "<meta property=\"og:image:width\" content=\"640\">"
##  [19] "<meta property=\"og:image:height\" content=\"480\">"
##  [20] "<meta name=\"viewport\" content=\"width=1120\">"
##  [21] "<meta property=\"og:title\" content=\"Data science - Wikipedia\">"
##  [22] "<meta property=\"og:type\" content=\"website\">"
```

```
##  [23] "<link rel=\"preconnect\" href=\"//upload.wikimedia.org\">"
##  [24] "<link rel=\"alternate\" media=\"only screen and (max-width: 640px)\" href=\"//en.m.wikipedia.org/wiki/Dat
##  [25] "<link rel=\"alternate\" type=\"application/x-wiki\" title=\"Edit this page\" href=\"/w/index.php?title=Da
##  [26] "<link rel=\"apple-touch-icon\" href=\"/static/apple-touch/wikipedia.png\">"
##  [27] "<link rel=\"icon\" href=\"/static/favicon/wikipedia.ico\">"
##  [28] "<link rel=\"search\" type=\"application/opensearchdescription+xml\" href=\"/w/rest.php/v1/search\" title=
##  [29] "<link rel=\"EditURI\" type=\"application/rsd+xml\" href=\"//en.wikipedia.org/w/api.php?action=rsd\">"
##  [30] "<link rel=\"canonical\" href=\"https://en.wikipedia.org/wiki/Data_science\">"
##  [31] "<link rel=\"license\" href=\"https://creativecommons.org/licenses/by-sa/4.0/deed.en\">"
##  [32] "<link rel=\"alternate\" type=\"application/atom+xml\" title=\"Wikipedia Atom feed\" href=\"/w/index.php?t
##  [33] "<link rel=\"dns-prefetch\" href=\"//meta.wikimedia.org\" />"
##  [34] "<link rel=\"dns-prefetch\" href=\"auth.wikimedia.org\">"
##  [35] "<body class=\"skin--responsive skin-vector skin-vector-search-vue mediawiki ltr sitedir-ltr mw-hide-empty
##  [36] "<div id=\"vector-main-menu-dropdown\" class=\"vector-dropdown vector-main-menu-dropdown vector-button-flu
##  [37] "\t<input type=\"checkbox\" id=\"vector-main-menu-dropdown-checkbox\" role=\"button\" aria-haspopup=\"true
##  [38] "\t<label id=\"vector-main-menu-dropdown-label\" for=\"vector-main-menu-dropdown-checkbox\" class=\"vector
##  [39] "<span class=\"vector-dropdown-label-text\">Main menu</span>"
##  [40] "\t<div class=\"vector-dropdown-content\">"
##  [41] "\t\t\t\t<div id=\"vector-main-menu-unpinned-container\" class=\"vector-unpinned-container\">"
##  [42] "<div id=\"vector-main-menu\" class=\"vector-main-menu vector-pinnable-element\">"
##  [43] "\tclass=\"vector-pinnable-header vector-main-menu-pinnable-header vector-pinnable-header-unpinned\""
##  [44] "\tdata-feature-name=\"main-menu-pinned\""
##  [45] "\tdata-pinnable-element-id=\"vector-main-menu\""
##  [46] "\tdata-pinned-container-id=\"vector-main-menu-pinned-container\""
##  [47] "\tdata-unpinned-container-id=\"vector-main-menu-unpinned-container\""
##  [48] "\t<div class=\"vector-pinnable-header-label\">Main menu</div>"
##  [49] "\t<button class=\"vector-pinnable-header-toggle-button vector-pinnable-header-pin-button\" data-event-nam
##  [50] "\t<button class=\"vector-pinnable-header-toggle-button vector-pinnable-header-unpin-button\" data-event-n
##  [51] "<div id=\"p-navigation\" class=\"vector-menu mw-portlet mw-portlet-navigation\"  >"
##  [52] "\t\t\t<li id=\"n-mainpage-description\" class=\"mw-list-item\"><a href=\"/wiki/Main_Page\" title=\"Visit
##  [53] "<div id=\"p-interaction\" class=\"vector-menu mw-portlet mw-portlet-interaction\"  >"
##  [54] "\t\t\t<li id=\"n-help\" class=\"mw-list-item\"><a href=\"/wiki/Help:Contents\" title=\"Guidance on how to
##  [55] "\t<img class=\"mw-logo-icon\" src=\"/static/images/icons/wikipedia.png\" alt=\"\" aria-hidden=\"true\" he
##  [56] "\t<span class=\"mw-logo-container skin-invert\">"
##  [57] "\t\t<img class=\"mw-logo-wordmark\" alt=\"Wikipedia\" src=\"/static/images/mobile/copyright/wikipedia-wor
##  [58] "\t\t<img class=\"mw-logo-tagline\" alt=\"The Free Encyclopedia\" src=\"/static/images/mobile/copyright/wi
##  [59] "\t</span>"
##  [60] "<div id=\"p-search\" role=\"search\" class=\"vector-search-box-vue  vector-search-box-collapses vector-se
##  [61] "\t<a href=\"/wiki/Special:Search\" class=\"cdx-button cdx-button--fake-button cdx-button--fake-button--en
##  [62] "<span>Search</span>"
##  [63] "\t<div class=\"vector-typeahead-search-container\">"
##  [64] "\t\t<div class=\"cdx-typeahead-search cdx-typeahead-search--show-thumbnail cdx-typeahead-search--auto-exp
##  [65] "\t\t\t<form action=\"/w/index.php\" id=\"searchform\" class=\"cdx-search-input cdx-search-input--has-end-
##  [66] "\t\t\t\t<div id=\"simpleSearch\" class=\"cdx-search-input__input-wrapper\"  data-search-loc=\"header-move
##  [67] "\t\t\t\t\t<div class=\"cdx-text-input cdx-text-input--has-start-icon\">"
##  [68] "\t\t\t\t\t\t<input"
##  [69] "\t\t\t\t\t\tclass=\"cdx-text-input__input\""
##  [70] "\t\t\t\t\t\t type=\"search\" name=\"search\" placeholder=\"Search Wikipedia\" aria-label=\"Search Wikip
##  [71] "\t\t\t\t\t<span class=\"cdx-text-input__icon cdx-text-input__start-icon\"></span>"
##  [72] "\t\t\t\t<input type=\"hidden\" name=\"title\" value=\"Special:Search\">"
##  [73] "\t\t\t<button class=\"cdx-button cdx-search-input__end-button\">Search</button>"
##  [74] "<div id=\"p-vector-user-menu-preferences\" class=\"vector-menu mw-portlet emptyPortlet\"  >"
##  [75] "<div id=\"p-vector-user-menu-userpage\" class=\"vector-menu mw-portlet emptyPortlet\"  >"
##  [76] "\t<nav class=\"vector-appearance-landmark\" aria-label=\"Appearance\">"
##  [77] "<div id=\"vector-appearance-dropdown\" class=\"vector-dropdown \"  title=\"Change the appearance of the p
##  [78] "\t<input type=\"checkbox\" id=\"vector-appearance-dropdown-checkbox\" role=\"button\" aria-haspopup=\"tru
##  [79] "\t<label id=\"vector-appearance-dropdown-label\" for=\"vector-appearance-dropdown-checkbox\" class=\"vect
##  [80] "<span class=\"vector-dropdown-label-text\">Appearance</span>"
##  [81] "\t<div class=\"vector-dropdown-content\">"
##  [82] "\t\t\t<div id=\"vector-appearance-unpinned-container\" class=\"vector-unpinned-container\">"
##  [83] "<div id=\"p-vector-user-menu-notifications\" class=\"vector-menu mw-portlet emptyPortlet\"  >"
##  [84] "<div id=\"p-vector-user-menu-overflow\" class=\"vector-menu mw-portlet\"  >"
##  [85] "\t\t\t<li id=\"pt-sitesupport-2\" class=\"user-links-collapsible-item mw-list-item user-links-collapsible
##  [86] "<li id=\"pt-createaccount-2\" class=\"user-links-collapsible-item mw-list-item user-links-collapsible-ite
```

```
##  [87] "<li id=\"pt-login-2\" class=\"user-links-collapsible-item mw-list-item user-links-collapsible-item\"><a c
##  [88] "<div id=\"vector-user-links-dropdown\" class=\"vector-dropdown vector-user-menu vector-button-flush-right
##  [89] "\t<input type=\"checkbox\" id=\"vector-user-links-dropdown-checkbox\" role=\"button\" aria-haspopup=\"tru
##  [90] "\t<label id=\"vector-user-links-dropdown-label\" for=\"vector-user-links-dropdown-checkbox\" class=\"vect
##  [91] "<span class=\"vector-dropdown-label-text\">Personal tools</span>"
##  [92] "\t<div class=\"vector-dropdown-content\">"
##  [93] "<div id=\"p-personal\" class=\"vector-menu mw-portlet mw-portlet-personal user-links-collapsible-item\"
##  [94] "\t\t\t<li id=\"pt-sitesupport\" class=\"user-links-collapsible-item mw-list-item\"><a href=\"https://dona
##  [95] "<div id=\"p-user-menu-anon-editor\" class=\"vector-menu mw-portlet mw-portlet-user-menu-anon-editor\"  >"
##  [96] "\t\tPages for logged out editors <a href=\"/wiki/Help:Introduction\" aria-label=\"Learn more about editi
##  [97] "\t\t\t<li id=\"pt-anoncontribs\" class=\"mw-list-item\"><a href=\"/wiki/Special:MyContributions\" title=\
##  [98] "<div class=\"mw-page-container\">"
##  [99] "\t<div class=\"mw-page-container-inner\">"
## [100] "\t\t\t<nav id=\"mw-panel\" class=\"vector-main-menu-landmark\" aria-label=\"Site\">"
## [101] "\t\t\t\t<div id=\"vector-main-menu-pinned-container\" class=\"vector-pinned-container\">"
## [102] "\t<div class=\"vector-sticky-pinned-container\">"
## [103] "\t\t\t\t\t<nav id=\"mw-panel-toc\" aria-label=\"Contents\" data-event-name=\"ui.sidebar-toc\" class=\"mw-ta
## [104] "\t\t\t\t\t\t<div id=\"vector-toc-pinned-container\" class=\"vector-pinned-container\">"
## [105] "\t\t\t\t\t\t<div id=\"vector-toc\" class=\"vector-toc vector-pinnable-element\">"
## [106] "\tclass=\"vector-pinnable-header vector-toc-pinnable-header vector-pinnable-header-pinned\""
## [107] "\tdata-feature-name=\"toc-pinned\""
## [108] "\tdata-pinnable-element-id=\"vector-toc\""
## [109] "\t<h2 class=\"vector-pinnable-header-label\">Contents</h2>"
## [110] "\t<button class=\"vector-pinnable-header-toggle-button vector-pinnable-header-pin-button\" data-event-nam
## [111] "\t<button class=\"vector-pinnable-header-toggle-button vector-pinnable-header-unpin-button\" data-event-n
## [112] "\t<ul class=\"vector-toc-contents\" id=\"mw-panel-toc-list\">"
## [113] "\t\t\t\t<div class=\"vector-toc-text\">(Top)</div>"
## [114] "\t\tclass=\"vector-toc-list-item vector-toc-level-1 vector-toc-list-item-expanded\">"
## [115] "\t\t\t\t<span class=\"vector-toc-numb\">1</span>"
## [116] "\t\t\t\t<span>Foundations</span>"
## [117] "\t\tclass=\"vector-toc-list-item vector-toc-level-1 vector-toc-list-item-expanded\">"
## [118] "\t\t\t\t<span class=\"vector-toc-numb\">2</span>"
## [119] "\t\t\t\t<span>Etymology</span>"
## [120] "\t\t\t\t<span class=\"vector-icon mw-ui-icon-wikimedia-expand\"></span>"
## [121] "\t\t\t\t<span>Toggle Etymology subsection</span>"
## [122] "\t\t\t\t\t<span class=\"vector-toc-numb\">2.1</span>"
## [123] "\t\t\t\t\t<span>Early usage</span>"
## [124] "\t\t\t\t\t<span class=\"vector-toc-numb\">2.2</span>"
## [125] "\t\t\t\t\t<span>Modern usage</span>"
## [126] "\t\tclass=\"vector-toc-list-item vector-toc-level-1 vector-toc-list-item-expanded\">"
## [127] "\t\t\t\t<span class=\"vector-toc-numb\">3</span>"
## [128] "\t\t\t\t<span>Data science and data analysis</span>"
## [129] "\t<li id=\"toc-Cloud_computing_for_data_science\""
## [130] "\t\tclass=\"vector-toc-list-item vector-toc-level-1 vector-toc-list-item-expanded\">"
## [131] "\t\t<a class=\"vector-toc-link\" href=\"#Cloud_computing_for_data_science\">"
## [132] "\t\t\t\t<span class=\"vector-toc-numb\">4</span>"
## [133] "\t\t\t\t<span>Cloud computing for data science</span>"
## [134] "\t\t<ul id=\"toc-Cloud_computing_for_data_science-sublist\" class=\"vector-toc-list\">"
## [135] "\t\tclass=\"vector-toc-list-item vector-toc-level-1 vector-toc-list-item-expanded\">"
## [136] "\t\t\t\t<span class=\"vector-toc-numb\">5</span>"
## [137] "\t\t\t\t<span>Ethical consideration in data science</span>"
## [138] "\t\tclass=\"vector-toc-list-item vector-toc-level-1 vector-toc-list-item-expanded\">"
## [139] "\t\t\t\t<span class=\"vector-toc-numb\">6</span>"
## [140] "\t\t\t\t<span>See also</span>"
## [141] "\t\tclass=\"vector-toc-list-item vector-toc-level-1 vector-toc-list-item-expanded\">"
## [142] "\t\t\t\t<span class=\"vector-toc-numb\">7</span>"
## [143] "\t\t\t\t<span>References</span>"
## [144] "\t\t\t\t<header class=\"mw-body-header vector-page-titlebar\">"
## [145] "<div id=\"vector-page-titlebar-toc\" class=\"vector-dropdown vector-page-titlebar-toc vector-button-flush
## [146] "\t<input type=\"checkbox\" id=\"vector-page-titlebar-toc-checkbox\" role=\"button\" aria-haspopup=\"true\
## [147] "\t<label id=\"vector-page-titlebar-toc-label\" for=\"vector-page-titlebar-toc-checkbox\" class=\"vector-d
## [148] "<span class=\"vector-dropdown-label-text\">Toggle the table of contents</span>"
## [149] "\t<div class=\"vector-dropdown-content\">"
## [150] "\t\t\t\t\t\t\t<div id=\"vector-page-titlebar-toc-unpinned-container\" class=\"vector-unpinned-container\
```

```
## [151] "\t\t\t\t\t<h1 id=\"firstHeading\" class=\"firstHeading mw-first-heading\"><span class=\"mw-page-title-ma
## [152] "<div id=\"p-lang-btn\" class=\"vector-dropdown mw-portlet mw-portlet-lang\"  >"
## [153] "\t<input type=\"checkbox\" id=\"p-lang-btn-checkbox\" role=\"button\" aria-haspopup=\"true\" data-event-n
## [154] "\t<label id=\"p-lang-btn-label\" for=\"p-lang-btn-checkbox\" class=\"vector-dropdown-label cdx-button cdx
## [155] "<span class=\"vector-dropdown-label-text\">50 languages</span>"
## [156] "\t<div class=\"vector-dropdown-content\">"
## [157] "\t\t\t\t<li class=\"interlanguage-link interwiki-ar mw-list-item\"><a href=\"https://ar.wikipedia.org/wik
Arabic\" lang=\"ar\" hreflang=\"ar\" data-title=\"   "\     data-language-autonym=\"    \" data-language-local-name
Azerbaijani\" lang=\"az\" hreflang=\"az\" data-title=\"Verilənlər elmi\" data-language-autonym=\"Azərbaycanca\" dat
Bangla\" lang=\"bn\" hreflang=\"bn\" data-title=\"       \" data-language-autonym=\"   \" data-language-local-name=
Bulgarian\" lang=\"bg\" hreflang=\"bg\" data-title=\"          \" data-language-autonym=\"     \" data-language-loc
Catalan\" lang=\"ca\" hreflang=\"ca\" data-title=\"Ciència de les dades\" data-language-autonym=\"Català\" data-lan
Czech\" lang=\"cs\" hreflang=\"cs\" data-title=\"Data science\" data-language-autonym=\"Čeština\" data-language-loc
German\" lang=\"de\" hreflang=\"de\" data-title=\"Data Science\" data-language-autonym=\"Deutsch\" data-language-lo
Estonian\" lang=\"et\" hreflang=\"et\" data-title=\"Andmeteadus\" data-language-autonym=\"Eesti\" data-language-loc
Greek\" lang=\"el\" hreflang=\"el\" data-title=\"E          \" data-language-autonym=\"E    \" data-language-local-
Spanish\" lang=\"es\" hreflang=\"es\" data-title=\"Ciencia de datos\" data-language-autonym=\"Español\" data-langua
Esperanto\" lang=\"eo\" hreflang=\"eo\" data-title=\"Datum-scienco\" data-language-autonym=\"Esperanto\" data-langu
Basque\" lang=\"eu\" hreflang=\"eu\" data-title=\"Datu zientzia\" data-language-autonym=\"Euskara\" data-language-l
Persian\" lang=\"fa\" hreflang=\"fa\" data-title=\"   "\   data-language-autonym=\"   \" data-language-local-name=\
French\" lang=\"fr\" hreflang=\"fr\" data-title=\"Science des données\" data-language-autonym=\"Français\" data-lan
Galician\" lang=\"gl\" hreflang=\"gl\" data-title=\"Ciencia de datos\" data-language-autonym=\"Galego\" data-langua
Korean\" lang=\"ko\" hreflang=\"ko\" data-title=\"     \" data-language-autonym=\"  \" data-language-local-name=\"K
Armenian\" lang=\"hy\" hreflang=\"hy\" data-title=\"           \" data-language-autonym=\"    \" data-language-loca
Hindi\" lang=\"hi\" hreflang=\"hi\" data-title=\"      \" data-language-autonym=\"   \" data-language-local-name=\"
Ido\" lang=\"io\" hreflang=\"io\" data-title=\"Cienco di datumi\" data-language-autonym=\"Ido\" data-language-local
Indonesian\" lang=\"id\" hreflang=\"id\" data-title=\"Ilmu data\" data-language-autonym=\"Bahasa Indonesia\" data-l
Zulu\" lang=\"zu\" hreflang=\"zu\" data-title=\"INzululwazi yeMininingo\" data-language-autonym=\"IsiZulu\" data-la
Italian\" lang=\"it\" hreflang=\"it\" data-title=\"Scienza dei dati\" data-language-autonym=\"Italiano\" data-langu
Hebrew\" lang=\"he\" hreflang=\"he\" data-title=\"   "\     data-language-autonym=\"   \" data-language-local-name=\
Kannada\" lang=\"kn\" hreflang=\"kn\" data-title=\"       \" data-language-autonym=\"  \" data-language-local-name=
Kazakh\" lang=\"kk\" hreflang=\"kk\" data-title=\"            \" data-language-autonym=\"    \" data-language-local
Latvian\" lang=\"lv\" hreflang=\"lv\" data-title=\"Datu mācība\" data-language-autonym=\"Latviešu\" data-language-l
Macedonian\" lang=\"mk\" hreflang=\"mk\" data-title=\"           \" data-language-autonym=\"     \" data-language-
Malay\" lang=\"ms\" hreflang=\"ms\" data-title=\"Sains data\" data-language-autonym=\"Bahasa Melayu\" data-language
Burmese\" lang=\"my\" hreflang=\"my\" data-title=\"       \" data-language-autonym=\"     \" data-language-local-na
Dutch\" lang=\"nl\" hreflang=\"nl\" data-title=\"Datawetenschap\" data-language-autonym=\"Nederlands\" data-languag
Japanese\" lang=\"ja\" hreflang=\"ja\" data-title=\"    \" data-language-autonym=\"  \" data-language-local-name=\"
Norwegian Bokmål\" lang=\"nb\" hreflang=\"nb\" data-title=\"Datavitenskap\" data-language-autonym=\"Norsk bokmål\"
Punjabi\" lang=\"pa\" hreflang=\"pa\" data-title=\"     \" data-language-autonym=\"   \" data-language-local-name=\
Polish\" lang=\"pl\" hreflang=\"pl\" data-title=\"Danologia\" data-language-autonym=\"Polski\" data-language-local-
Portuguese\" lang=\"pt\" hreflang=\"pt\" data-title=\"Ciência de dados\" data-language-autonym=\"Português\" data-l
Quechua\" lang=\"qu\" hreflang=\"qu\" data-title=\"Willakuy hamut&#039;ay\" data-language-autonym=\"Runa Simi\" dat
Russian\" lang=\"ru\" hreflang=\"ru\" data-title=\"        \" data-language-autonym=\"    \" data-language-local-na
Albanian\" lang=\"sq\" hreflang=\"sq\" data-title=\"Shkenca e të dhënave\" data-language-autonym=\"Shqip\" data-lan
Simple English\" lang=\"en-simple\" hreflang=\"en-simple\" data-title=\"Data science\" data-language-autonym=\"Simp
Sindhi\" lang=\"sd\" hreflang=\"sd\" data-title=\"   "\    data-language-autonym=\"  \" data-language-local-name=\"
Serbian\" lang=\"sr\" hreflang=\"sr\" data-title=\"         \" data-language-autonym=\"    / srpski\" data-languag
Finnish\" lang=\"fi\" hreflang=\"fi\" data-title=\"Datatiede\" data-language-autonym=\"Suomi\" data-language-local-
Tamil\" lang=\"ta\" hreflang=\"ta\" data-title=\"       \" data-language-autonym=\"  \" data-language-local-name=\"
Thai\" lang=\"th\" hreflang=\"th\" data-title=\"      \" data-language-autonym=\"  \" data-language-local-name=\"Th
Turkish\" lang=\"tr\" hreflang=\"tr\" data-title=\"Veri bilimi\" data-language-autonym=\"Türkçe\" data-language-loc
Ukrainian\" lang=\"uk\" hreflang=\"uk\" data-title=\"        \" data-language-autonym=\"     \" data-language-loca
Urdu\" lang=\"ur\" hreflang=\"ur\" data-title=\"   "\    data-language-autonym=\"  \" data-language-local-name=\"Ur
Vietnamese\" lang=\"vi\" hreflang=\"vi\" data-title=\"Khoa học dữ liệu\" data-language-autonym=\"Tiếng Việt\" data-
Cantonese\" lang=\"yue\" hreflang=\"yue\" data-title=\"  \" data-language-autonym=\" \" data-language-local-name=\
Chinese\" lang=\"zh\" hreflang=\"zh\" data-title=\"  \" data-language-autonym=\" \" data-language-local-name=\"Chi
## [158] "\t\t\t<div class=\"after-portlet after-portlet-lang\"><span class=\"wb-langlinks-edit wb-langlinks-link\"
## [159] "\t\t\t\t<div class=\"vector-page-toolbar\">"
## [160] "\t\t\t\t\t<div class=\"vector-page-toolbar-container\">"
## [161] "\t\t\t\t\t\t\t<nav aria-label=\"Namespaces\">"
## [162] "<div id=\"p-associated-pages\" class=\"vector-menu vector-menu-tabs mw-portlet mw-portlet-associated-page
## [163] "\t\t\t<li id=\"ca-nstab-main\" class=\"selected vector-tab-noicon mw-list-item\"><a href=\"/wiki/Data_sci
## [164] "<div id=\"vector-variants-dropdown\" class=\"vector-dropdown emptyPortlet\"  >"
```

```
## [165] "\t<input type=\"checkbox\" id=\"vector-variants-dropdown-checkbox\" role=\"button\" aria-haspopup=\"true\
## [166] "\t<label id=\"vector-variants-dropdown-label\" for=\"vector-variants-dropdown-checkbox\" class=\"vector-
## [167] "\t<div class=\"vector-dropdown-content\">"
## [168] "<div id=\"p-variants\" class=\"vector-menu mw-portlet mw-portlet-variants emptyPortlet\"  >"
## [169] "\t\t\t\t\t\t<div id=\"right-navigation\" class=\"vector-collapsible\">"
## [170] "<div id=\"p-views\" class=\"vector-menu vector-menu-tabs mw-portlet mw-portlet-views\"  >"
## [171] "\t\t\t<li id=\"ca-view\" class=\"selected vector-tab-noicon mw-list-item\"><a href=\"/wiki/Data_science\
## [172] "\t\t\t\t\t\t\t<nav class=\"vector-page-tools-landmark\" aria-label=\"Page tools\">"
## [173] "<div id=\"vector-page-tools-dropdown\" class=\"vector-dropdown vector-page-tools-dropdown\"  >"
## [174] "\t<input type=\"checkbox\" id=\"vector-page-tools-dropdown-checkbox\" role=\"button\" aria-haspopup=\"tru
## [175] "\t<label id=\"vector-page-tools-dropdown-label\" for=\"vector-page-tools-dropdown-checkbox\" class=\"vect
## [176] "\t<div class=\"vector-dropdown-content\">"
## [177] "\t\t\t\t\t\t\t\t\t<div id=\"vector-page-tools-unpinned-container\" class=\"vector-unpinned-container\">"
## [178] "<div id=\"vector-page-tools\" class=\"vector-page-tools vector-pinnable-element\">"
## [179] "\tclass=\"vector-pinnable-header vector-page-tools-pinnable-header vector-pinnable-header-unpinned\""
## [180] "\tdata-feature-name=\"page-tools-pinned\""
## [181] "\tdata-pinnable-element-id=\"vector-page-tools\""
## [182] "\tdata-pinned-container-id=\"vector-page-tools-pinned-container\""
## [183] "\tdata-unpinned-container-id=\"vector-page-tools-unpinned-container\""
## [184] "\t<div class=\"vector-pinnable-header-label\">Tools</div>"
## [185] "\t<button class=\"vector-pinnable-header-toggle-button vector-pinnable-header-pin-button\" data-event-nam
## [186] "\t<button class=\"vector-pinnable-header-toggle-button vector-pinnable-header-unpin-button\" data-event-n
## [187] "<div id=\"p-cactions\" class=\"vector-menu mw-portlet mw-portlet-cactions emptyPortlet vector-has-collaps
## [188] "\t\t\t<li id=\"ca-more-view\" class=\"selected vector-more-collapsible-item mw-list-item\"><a href=\"/wik
## [189] "<div id=\"p-tb\" class=\"vector-menu mw-portlet mw-portlet-tb\"  >"
## [190] "\t\t\t<li id=\"t-whatlinkshere\" class=\"mw-list-item\"><a href=\"/wiki/Special:WhatLinksHere/Data_scienc
## [191] "<div id=\"p-coll-print_export\" class=\"vector-menu mw-portlet mw-portlet-coll-print_export\"  >"
## [192] "\t\tPrint/export"
## [193] "\t\t\t<li id=\"coll-download-as-rl\" class=\"mw-list-item\"><a href=\"/w/index.php?title=Special:Downloa
## [194] "<div id=\"p-wikibase-otherprojects\" class=\"vector-menu mw-portlet mw-portlet-wikibase-otherprojects\"
## [195] "\t\tIn other projects"
## [196] "\t\t\t<li class=\"wb-otherproject-link wb-otherproject-commons mw-list-item\"><a href=\"https://commons.w
## [197] "\t\t\t\t\t<div class=\"vector-sticky-pinned-container\">"
## [198] "\t\t\t\t\t\t<nav class=\"vector-page-tools-landmark\" aria-label=\"Page tools\">"
## [199] "\t\t\t\t\t\t\t<div id=\"vector-page-tools-pinned-container\" class=\"vector-pinned-container\">"
## [200] "\t\t\t\t\t\t<nav class=\"vector-appearance-landmark\" aria-label=\"Appearance\">"
## [201] "\t\t\t\t\t\t\t<div id=\"vector-appearance-pinned-container\" class=\"vector-pinned-container\">"
## [202] "\t\t\t\t<div id=\"vector-appearance\" class=\"vector-appearance vector-pinnable-element\">"
## [203] "\tclass=\"vector-pinnable-header vector-appearance-pinnable-header vector-pinnable-header-pinned\""
## [204] "\tdata-feature-name=\"appearance-pinned\""
## [205] "\tdata-pinnable-element-id=\"vector-appearance\""
## [206] "\tdata-pinned-container-id=\"vector-appearance-pinned-container\""
## [207] "\tdata-unpinned-container-id=\"vector-appearance-unpinned-container\""
## [208] "\t<div class=\"vector-pinnable-header-label\">Appearance</div>"
## [209] "\t<button class=\"vector-pinnable-header-toggle-button vector-pinnable-header-pin-button\" data-event-nam
## [210] "\t<button class=\"vector-pinnable-header-toggle-button vector-pinnable-header-unpin-button\" data-event-n
## [211] "\t\t\t\t\t\t<div id=\"siteSub\" class=\"noprint\">From Wikipedia, the free encyclopedia</div>"
## [212] "\t\t\t\t\t<div id=\"mw-content-text\" class=\"mw-body-content\"><div class=\"mw-content-ltr mw-parser-out
## [213] "<style data-mw-deduplicate=\"TemplateStyles:r1236090951\">.mw-parser-output .hatnote{font-style:italic}.m
## [214] "<p class=\"mw-empty-elt\">"
## [215] "</p>"
## [216] "<figure class=\"mw-default-size\" typeof=\"mw:File/Thumb\"><a href=\"/wiki/File:PIA23792-1600x1200(1).jpg
## [217] "<p><b>Data science</b> is an <a href=\"/wiki/Interdisciplinary\" class=\"mw-redirect\" title=\"Interdisci
## [218] "</p><p>Data science also integrates domain knowledge from the underlying application domain (e.g., natura
## [219] "</p><p>Data science is \"a concept to unify <a href=\"/wiki/Statistics\" title=\"Statistics\">statistics<
## [220] "</p><p>A <b>data scientist</b> is a professional who creates programming code and combines it with statis
## [221] "</p>"
## [222] "<meta property=\"mw:PageProp/toc\" />"
## [223] "<div class=\"mw-heading mw-heading2\"><h2 id=\"Foundations\">Foundations</h2><span class=\"mw-editsection
## [224] "<p>Data science is an <a href=\"/wiki/Interdisciplinarity\" title=\"Interdisciplinarity\">interdisciplina
## [225] "</p><p><a href=\"/wiki/Vasant_Dhar\" title=\"Vasant Dhar\">Vasant Dhar</a> writes that statistics emphasi
## [226] "</p>"
## [227] "<div class=\"mw-heading mw-heading2\"><h2 id=\"Etymology\">Etymology</h2><span class=\"mw-editsection\"><
## [228] "<div class=\"mw-heading mw-heading3\"><h3 id=\"Early_usage\">Early usage</h3><span class=\"mw-editsection
```

```
## [229] "<p>In 1962, <a href=\"/wiki/John_Tukey\" title=\"John Tukey\">John Tukey</a> described a field he called
## [230] "</p><p>The term \"data science\" has been traced back to 1974, when <a href=\"/wiki/Peter_Naur\" title=\
## [231] "</p>"
## [232] "<div class=\"mw-heading mw-heading3\"><h3 id=\"Modern_usage\">Modern usage</h3><span class=\"mw-editsecti
## [233] "<p>In 2012, technologists <a href=\"/wiki/Thomas_H._Davenport\" title=\"Thomas H. Davenport\">Thomas H.
## [234] "</p><p>The modern conception of data science as an independent discipline is sometimes attributed to <a
## [235] "</p><p>The professional title of \"data scientist\" has been attributed to <a href=\"/wiki/DJ_Patil\" ti
## [236] "</p>"
## [237] "<div class=\"mw-heading mw-heading2\"><h2 id=\"Data_science_and_data_analysis\">Data science and data ana
## [238] "<figure class=\"mw-default-size\" typeof=\"mw:File/Thumb\"><a href=\"/wiki/File:EDA_example_-_Always_plot
## [239] "<figure class=\"mw-default-size\" typeof=\"mw:File/Thumb\"><a href=\"/wiki/File:Data_Science.png\" class=
## [240] "<p>Data analysis typically involves working with structured datasets to answer specific questions or sol
## [241] "</p><p>Data science involves working with larger datasets that often require advanced computational and s
## [242] "</p>"
## [243] "<div class=\"mw-heading mw-heading2\"><h2 id=\"Cloud_computing_for_data_science\">Cloud computing for dat
## [244] "<figure class=\"mw-default-size\" typeof=\"mw:File/Thumb\"><a href=\"/wiki/File:Cloud_computing_in_enabli
## [245] "<p><a href=\"/wiki/Cloud_computing\" title=\"Cloud computing\">Cloud computing</a> can offer access to la
## [246] "</p><p>Some distributed computing frameworks are designed to handle big data workloads. These frameworks
## [247] "</p>"
## [248] "<div class=\"mw-heading mw-heading2\"><h2 id=\"Ethical_consideration_in_data_science\">Ethical considerat
## [249] "<p>Data science involves collecting, processing, and analyzing data which often includes personal and ser
## [250] "</p><p>Machine learning models can amplify existing biases present in training data, leading to discrimi
## [251] "</p>"
## [252] "<div class=\"mw-heading mw-heading2\"><h2 id=\"See_also\">See also</h2><span class=\"mw-editsection\"><sp
## [253] "<ul><li><a href=\"/wiki/Python_(programming_language)\" title=\"Python (programming language)\">Python (p
## [254] "<li><a href=\"/wiki/R_(programming_language)\" title=\"R (programming language)\">R (programming language
## [255] "<li><a href=\"/wiki/Topological_data_analysis\" title=\"Topological data analysis\">Topological data anal
## [256] "<li><a href=\"/wiki/List_of_free_and_open-source_software_packages#Data_science\" title=\"List of free an
## [257] "<div class=\"mw-heading mw-heading2\"><h2 id=\"References\">References</h2><span class=\"mw-editsection\"
## [258] "<style data-mw-deduplicate=\"TemplateStyles:r1239543626\">.mw-parser-output .reflist{margin-bottom:0.5em
## [259] "<div class=\"mw-references-wrap mw-references-columns\"><ol class=\"references\">"
## [260] "<li id=\"cite_note-1\"><span class=\"mw-cite-backlink\"><b><a href=\"#cite_ref-1\">^</a></b></span> <span
</span>766. <a href=\"/wiki/Doi_(identifier)\" class=\"mw-redirect\" title=\"Doi (identifier)\">doi</a>:<span class
## [261] "<li id=\"cite_note-2\"><span class=\"mw-cite-backlink\"><b><a href=\"#cite_ref-2\">^</a></b></span> <span
</span>73. <a href=\"/wiki/Doi_(identifier)\" class=\"mw-redirect\" title=\"Doi (identifier)\">doi</a>:<a rel=\"no
## [262] "<li id=\"cite_note-3\"><span class=\"mw-cite-backlink\"><b><a href=\"#cite_ref-3\">^</a></b></span> <span
## [263] "<li id=\"cite_note-4\"><span class=\"mw-cite-backlink\"><b><a href=\"#cite_ref-4\">^</a></b></span> <span
</span>13. <a href=\"/wiki/Doi_(identifier)\" class=\"mw-redirect\" title=\"Doi (identifier)\">doi</a>:<span class=
## [264] "<li id=\"cite_note-5\"><span class=\"mw-cite-backlink\"><b><a href=\"#cite_ref-5\">^</a></b></span> <span
</span>51. <a href=\"/wiki/Doi_(identifier)\" class=\"mw-redirect\" title=\"Doi (identifier)\">doi</a>:<a rel=\"no
## [265] "<li id=\"cite_note-:2-6\"><span class=\"mw-cite-backlink\">^ <a href=\"#cite_ref-:2_6-0\"><sup><i><b>a</b
43:42. <a href=\"/wiki/ArXiv_(identifier)\" class=\"mw-redirect\" title=\"ArXiv (identifier)\">arXiv</a>:<span clas
## [266] "<li id=\"cite_note-TansleyTolle2009-7\"><span class=\"mw-cite-backlink\"><b><a href=\"#cite_ref-TansleyTo
## [267] "<li id=\"cite_note-BellHey2009-8\"><span class=\"mw-cite-backlink\"><b><a href=\"#cite_ref-BellHey2009_8-
</span>1298. <a href=\"/wiki/Doi_(identifier)\" class=\"mw-redirect\" title=\"Doi (identifier)\">doi</a>:<a rel=\"n
## [268] "<li id=\"cite_note-9\"><span class=\"mw-cite-backlink\"><b><a href=\"#cite_ref-9\">^</a></b></span> <span
</span>76, 128. <a href=\"/wiki/PMID_(identifier)\" class=\"mw-redirect\" title=\"PMID (identifier)\">PMID</a> 
## [269] "<li id=\"cite_note-10\"><span class=\"mw-cite-backlink\"><b><a href=\"#cite_ref-10\">^</a></b></span> <sp
</span>251. <a href=\"/wiki/Doi_(identifier)\" class=\"mw-redirect\" title=\"Doi (identifier)\">doi</a>:<span class
## [270] "<li id=\"cite_note-11\"><span class=\"mw-cite-backlink\"><b><a href=\"#cite_ref-11\">^</a></b></span> <sp
## [271] "<li id=\"cite_note-12\"><span class=\"mw-cite-backlink\"><b><a href=\"#cite_ref-12\">^</a></b></span> <sp
</span>73. <a href=\"/wiki/Doi_(identifier)\" class=\"mw-redirect\" title=\"Doi (identifier)\">doi</a>:<a rel=\"no
## [272] "<li id=\"cite_note-13\"><span class=\"mw-cite-backlink\"><b><a href=\"#cite_ref-13\">^</a></b></span> <sp
## [273] "<li id=\"cite_note-:7-14\"><span class=\"mw-cite-backlink\">^ <a href=\"#cite_ref-:7_14-0\"><sup><i><b>a<
## [274] "<li id=\"cite_note-15\"><span class=\"mw-cite-backlink\"><b><a href=\"#cite_ref-15\">^</a></b></span> <sp
</span>7<span class=\"reference-accessdate\">. Retrieved <span class=\"nowrap\">29 November</span> 2020</span>.</c
## [275] "<li id=\"cite_note-16\"><span class=\"mw-cite-backlink\"><b><a href=\"#cite_ref-16\">^</a></b></span> <sp
## [276] "<li id=\"cite_note-Murtagh_2018_14-17\"><span class=\"mw-cite-backlink\">^ <a href=\"#cite_ref-Murtagh_20
## [277] "<li id=\"cite_note-18\"><span class=\"mw-cite-backlink\"><b><a href=\"#cite_ref-18\">^</a></b></span> <sp
## [278] "<li id=\"cite_note-19\"><span class=\"mw-cite-backlink\"><b><a href=\"#cite_ref-19\">^</a></b></span> <sp
## [279] "<li id=\"cite_note-20\"><span class=\"mw-cite-backlink\"><b><a href=\"#cite_ref-20\">^</a></b></span> <sp
## [280] "<li id=\"cite_note-21\"><span class=\"mw-cite-backlink\"><b><a href=\"#cite_ref-21\">^</a></b></span> <sp
## [281] "<li id=\"cite_note-22\"><span class=\"mw-cite-backlink\"><b><a href=\"#cite_ref-22\">^</a></b></span> <sp
## [282] "<li id=\"cite_note-23\"><span class=\"mw-cite-backlink\"><b><a href=\"#cite_ref-23\">^</a></b></span> <sp
```

```
## [283] "<li id=\"cite_note-24\"><span class=\"mw-cite-backlink\"><b><a href=\"#cite_ref-24\">^</a></b></span> <sp
## [284] "<li id=\"cite_note-25\"><span class=\"mw-cite-backlink\"><b><a href=\"#cite_ref-25\">^</a></b></span> <sp
## [285] "<li id=\"cite_note-26\"><span class=\"mw-cite-backlink\"><b><a href=\"#cite_ref-26\">^</a></b></span> <sp
NSB-05-40, Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century\"</a>. <i>www.ns
## [286] "<li id=\"cite_note-:8-27\"><span class=\"mw-cite-backlink\"><b><a href=\"#cite_ref-:8_27-0\">^</a></b></s
## [287] "<li id=\"cite_note-:9-28\"><span class=\"mw-cite-backlink\"><b><a href=\"#cite_ref-:9_28-0\">^</a></b></s
## [288] "<li id=\"cite_note-:10-29\"><span class=\"mw-cite-backlink\"><b><a href=\"#cite_ref-:10_29-0\">^</a></b><
## [289] "<li id=\"cite_note-30\"><span class=\"mw-cite-backlink\"><b><a href=\"#cite_ref-30\">^</a></b></span> <sp
</span>115. <a href=\"/wiki/Doi_(identifier)\" class=\"mw-redirect\" title=\"Doi (identifier)\">doi</a>:<a rel=\"no
## [290] "<li id=\"cite_note-:0-31\"><span class=\"mw-cite-backlink\"><b><a href=\"#cite_ref-:0_31-0\">^</a></b></s
## [291] "<li id=\"cite_note-:3-32\"><span class=\"mw-cite-backlink\"><b><a href=\"#cite_ref-:3_32-0\">^</a></b></s
</span>1394. <a href=\"/wiki/Doi_(identifier)\" class=\"mw-redirect\" title=\"Doi (identifier)\">doi</a>:<a rel=\"n
## [292] "<li id=\"cite_note-:4-33\"><span class=\"mw-cite-backlink\"><b><a href=\"#cite_ref-:4_33-0\">^</a></b></s
## [293] "<li id=\"cite_note-:11-34\"><span class=\"mw-cite-backlink\"><b><a href=\"#cite_ref-:11_34-0\">^</a></b><
</span>341. <a href=\"/wiki/Doi_(identifier)\" class=\"mw-redirect\" title=\"Doi (identifier)\">doi</a>:<a rel=\"no
## [294] "<li id=\"cite_note-35\"><span class=\"mw-cite-backlink\"><b><a href=\"#cite_ref-35\">^</a></b></span> <sp
## [295] "<li id=\"cite_note-36\"><span class=\"mw-cite-backlink\"><b><a href=\"#cite_ref-36\">^</a></b></span> <sp
</span>186. <a href=\"/wiki/ArXiv_(identifier)\" class=\"mw-redirect\" title=\"ArXiv (identifier)\">arXiv</a>:<span
## [296] "<div class=\"navbox-styles\"><style data-mw-deduplicate=\"TemplateStyles:r1129693374\">.mw-parser-output
## [297] "<li><a href=\"/wiki/Data_compression\" title=\"Data compression\">Compression</a></li>"
## [298] "<li><a href=\"/wiki/Data_corruption\" title=\"Data corruption\">Corruption</a></li>"
## [299] "<li><a href=\"/wiki/Data_deduplication\" title=\"Data deduplication\">Deduplication</a></li>"
## [300] "<li><a href=\"/wiki/Data_exploration\" title=\"Data exploration\">Exploration</a></li>"
## [301] "<ul><li><a href=\"/wiki/Data_cooperative\" title=\"Data cooperative\">Cooperatives</a></li></ul></li>"
## [302] "<li><a href=\"/wiki/Data_philanthropy\" title=\"Data philanthropy\">Philanthropy</a></li>"
## [303] "<li><a href=\"/wiki/Data_pre-processing\" class=\"mw-redirect\" title=\"Data pre-processing\">Pre-process
## [304] "<li><a href=\"/wiki/Data_preservation\" title=\"Data preservation\">Preservation</a></li>"
## [305] "<li><a href=\"/wiki/Data_processing\" title=\"Data processing\">Processing</a></li>"
## [306] "<li><a href=\"/wiki/Information_privacy\" title=\"Information privacy\">Protection (privacy)</a></li>"
## [307] "<li><a href=\"/wiki/Data_publishing\" title=\"Data publishing\">Publishing</a>"
## [308] "<ul><li><a href=\"/wiki/Open_data\" title=\"Open data\">Open data</a></li></ul></li>"
## [309] "<li><a href=\"/wiki/Data_scraping\" title=\"Data scraping\">Scraping</a></li>"
## [310] "<li><a href=\"/wiki/Data_steward\" title=\"Data steward\">Stewardship</a></li>"
## [311] "<li><a href=\"/wiki/Topological_data_analysis\" title=\"Topological data analysis\">Topological data anal
## [312] "<li><a href=\"/wiki/Data_type\" title=\"Data type\">Type</a></li>"
## [313] "NewPP limit report"
## [314] "Cache expiry: 2592000"
## [315] "Reduced expiry: false"
## [316] "Complications: [vary-revision-sha1, show-toc]"
## [317] "Preprocessor visited node count: 2302/1000000"
## [318] "Post-expand include size: 85027/2097152 bytes"
## [319] "Template argument size: 780/2097152 bytes"
## [320] "Highest expansion depth: 12/100"
## [321] "Expensive parser function count: 4/500"
## [322] "Unstrip recursion depth: 1/20"
## [323] "Unstrip post-expand size: 142478/5000000 bytes"
## [324] "Transclusion expansion time report (%,ms,calls,template)"
## [325] " 65.45%  361.134      1 Template:Reflist"
## [326] " 32.63%  180.049     16 Template:Cite_journal"
## [327] " 13.97%   77.061      1 Template:Data"
## [328] " 13.23%   73.001      1 Template:Navbox"
## [329] " 11.32%   62.486      1 Template:Short_description"
## [330] "  7.58%   41.816      6 Template:Cite_book"
## [331] "  7.07%   39.038      7 Template:Cite_web"
## [332] "  6.38%   35.213      4 Template:Cite_news"
## [333] "  6.16%   34.012      2 Template:Pagetype"
## [334] "<!-- Saved in parser cache with key enwiki:pcache:35458904:|#|:idhash:canonical and timestamp 20250412160
## [335] "</div><!--esi <esi:include src=\"/esitest-fa8a495983347898/content\" /> --><noscript><img src=\"https://a
## [336] "<div class=\"printfooter\" data-nosnippet=\"\">Retrieved from \"<a dir=\"ltr\" href=\"https://en.wikipedi
## [337] "\t\t\t\t\t<div id=\"catlinks\" class=\"catlinks\" data-mw=\"interface\"><div id=\"mw-normal-catlinks\" cl
## [338] "\t<li id=\"footer-info-lastmod\"> This page was last edited on 17 March 2025, at 07:41<span class=\"anony
## [339] "\t<li id=\"footer-info-copyright\">Text is available under the <a href=\"/wiki/Wikipedia:Text_of_the_Crea
## [340] "additional terms may apply. By using this site, you agree to the <a href=\"https://foundation.wikimedia.o
## [341] "\t<ul id=\"footer-places\">"
```

```
## [342] "\t<li id=\"footer-places-privacy\"><a href=\"https://foundation.wikimedia.org/wiki/Special:MyLanguage/Pol
## [343] "\t<li id=\"footer-places-about\"><a href=\"/wiki/Wikipedia:About\">About Wikipedia</a></li>"
## [344] "\t<li id=\"footer-places-disclaimers\"><a href=\"/wiki/Wikipedia:General_disclaimer\">Disclaimers</a></li
## [345] "\t<li id=\"footer-places-contact\"><a href=\"//en.wikipedia.org/wiki/Wikipedia:Contact_us\">Contact Wikip
## [346] "\t<li id=\"footer-places-wm-codeofconduct\"><a href=\"https://foundation.wikimedia.org/wiki/Special:MyLan
## [347] "\t<li id=\"footer-places-developers\"><a href=\"https://developer.wikimedia.org\">Developers</a></li>"
## [348] "\t<li id=\"footer-places-statslink\"><a href=\"https://stats.wikimedia.org/#/en.wikipedia.org\">Statistic
## [349] "\t<li id=\"footer-places-cookiestatement\"><a href=\"https://foundation.wikimedia.org/wiki/Special:MyLang
## [350] "\t<li id=\"footer-places-mobileview\"><a href=\"//en.m.wikipedia.org/w/index.php?title=Data_science&amp;m
## [351] "\t<ul id=\"footer-icons\" class=\"noprint\">"
## [352] "\t<li id=\"footer-copyrightico\"><a href=\"https://www.wikimedia.org/\" class=\"cdx-button cdx-button--fa
## [353] "\t<li id=\"footer-poweredbyico\"><a href=\"https://www.mediawiki.org/\" class=\"cdx-button cdx-button--fa
## [354] "\t\t\t\t<button class=\"cdx-button cdx-button--weight-quiet cdx-button--icon-only vector-sticky-header-se
## [355] "<span>Search</span>"
## [356] "\t\t\t<div class=\"vector-typeahead-search-container\">"
## [357] "\t\t\t\t<div class=\"cdx-typeahead-search cdx-typeahead-search--show-thumbnail\">"
## [358] "\t\t\t\t\t<form action=\"/w/index.php\" id=\"vector-sticky-search-form\" class=\"cdx-search-input cdx-sea
## [359] "\t\t\t\t\t\t<div  class=\"cdx-search-input__input-wrapper\"  data-search-loc=\"header-moved\">"
## [360] "\t\t\t\t\t\t\t<div class=\"cdx-text-input cdx-text-input--has-start-icon\">"
## [361] "\t\t\t\t\t\t\t\t<input"
## [362] "\t\t\t\t\t\t\t\tclass=\"cdx-text-input__input\""
## [363] "\t\t\t\t\t\t\t\ttype=\"search\" name=\"search\" placeholder=\"Search Wikipedia\">"
## [364] "\t\t\t\t\t\t\t<span class=\"cdx-text-input__icon cdx-text-input__start-icon\"></span>"
## [365] "\t\t\t\t\t\t\t<input type=\"hidden\" name=\"title\" value=\"Special:Search\">"
## [366] "\t\t\t\t\t\t<button class=\"cdx-button cdx-search-input__end-button\">Search</button>"
## [367] "\t\t\t\t\t<div id=\"vector-sticky-header-toc\" class=\"vector-dropdown mw-portlet mw-portlet-sticky-heade
## [368] "\t\t\t\t\t<input type=\"checkbox\" id=\"vector-sticky-header-toc-checkbox\" role=\"button\" aria-haspop
## [369] "\t\t\t\t\t<label id=\"vector-sticky-header-toc-label\" for=\"vector-sticky-header-toc-checkbox\" class=
## [370] "<span class=\"vector-dropdown-label-text\">Toggle the table of contents</span>"
## [371] "\t\t\t\t\t<div class=\"vector-dropdown-content\">"
## [372] "\t\t\t\t\t<div id=\"vector-sticky-header-toc-unpinned-container\" class=\"vector-unpinned-container\">"
## [373] "\t\t\t\t<div class=\"vector-sticky-header-context-bar-primary\" aria-hidden=\"true\" ><span class=\"mw-pa
## [374] "\t\t\t\t<a href=\"#\" class=\"cdx-button cdx-button--fake-button cdx-button--fake-button--enabled cdx-but
## [375] "<span></span>"
## [376] "\t\t\t<a href=\"#\" class=\"cdx-button cdx-button--fake-button cdx-button--fake-button--enabled cdx-butto
## [377] "<span></span>"
## [378] "\t\t\t<a href=\"#\" class=\"cdx-button cdx-button--fake-button cdx-button--fake-button--enabled cdx-butto
## [379] "<span></span>"
## [380] "\t\t\t<a href=\"#\" class=\"cdx-button cdx-button--fake-button cdx-button--fake-button--enabled cdx-butto
## [381] "<span></span>"
## [382] "\t\t\t<a href=\"#\" class=\"cdx-button cdx-button--fake-button cdx-button--fake-button--enabled cdx-butto
## [383] "<span></span>"
## [384] "\t\t\t<a href=\"#\" class=\"cdx-button cdx-button--fake-button cdx-button--fake-button--enabled cdx-butto
## [385] "<span></span>"
## [386] "\t\t\t<a href=\"#\" class=\"cdx-button cdx-button--fake-button cdx-button--fake-button--enabled cdx-butto
## [387] "<span></span>"
## [388] "\t\t\t\t<button class=\"cdx-button cdx-button--weight-quiet mw-interlanguage-selector\" id=\"p-lang-btn-s
## [389] "<span>50 languages</span>"
## [390] "\t\t\t<a href=\"#\" class=\"cdx-button cdx-button--fake-button cdx-button--fake-button--enabled cdx-butto
## [391] "<span>Add topic</span>"
## [392] "<div class=\"mw-portlet mw-portlet-dock-bottom emptyPortlet\" id=\"p-dock-bottom\">"
## [393] "<script>(RLQ=window.RLQ||[]).push(function(){mw.config.set({\"wgHostname\":\"mw-web.codfw.main-658b85fd8c
## [394] "<script type=\"application/ld+json\">{\"@context\":\"https:\\/\\/schema.org\",\"@type\":\"Article\",\"nam
```

```
#Using library XML
library(XML)
```

```
## Warning: package 'XML' was built under R version 4.4.3
```

```
doc<-htmlParse(eg6)
doc.text<-unlist(xpathApply(doc,'//p',xmlValue))
unlist(xpathApply(doc,'//h2',xmlValue))
```

```
## [1] "Contents"
## [2] "Foundations"
## [3] "Etymology"
## [4] "Data science and data analysis"
## [5] "Cloud computing for data science"
## [6] "Ethical consideration in data science"
## [7] "See also"
## [8] "References"
```

```
#Using library httr
library(httr)
```

```
## Warning: package 'httr' was built under R version 4.4.2
```

```
##
## Attaching package: 'httr'
```

```
## The following object is masked from 'package:NLP':
##
##     content
```

```
eg7<-GET("https://www.edureka.co/blog/what-is-data-science/")
doc<-htmlParse(eg7)
doc.text<-unlist(xpathApply(doc,'//p',xmlValue))
```

```
#Using library rvest
library(rvest)
```

```
## Warning: package 'rvest' was built under R version 4.4.2
```

```
eg8<-read_html("https://www.edureka.co/blog/what-is-data-science/")
nodesh<-html_nodes(eg8,'.col-lg-9 :nth-child(1)')
nodes<-html_nodes(eg8,'.color-4a div span , .btn-become-profesional-link+ p')
texts<-html_text(nodes)
#Selecting multiple pages
pages<-paste0('https://www.amazon.co.jp/s?k=skincare&crid=28HIW1TYLV9UM&sprefix=skincare%2Caps%2C268&ref=nb_sb_noss
pages<-paste0('https://www.amazon.com/s?k=condominium&crid=1ZNT0FW4JKRC3&sprefix=condominiu%2Caps%2C312&ref=nb_sb_
# &page=
# read the first 3 pages
eg10<-read_html(pages[1])
nodes<-html_nodes(eg10,'.a-price-whole')
texts<-html_text(nodes)
Price<-function(page){
  url<-read_html(page)
  nodes<-html_nodes(url ,'.a-price-whole ')
  html_text(nodes)}
sapply(pages,Price)
```

```
## $`https://www.amazon.com/s?k=condominium&crid=1ZNT0FW4JKRC3&sprefix=condominiu%2Caps%2C312&ref=nb_sb_noss_2&page
##  [1] "0."   "15."  "12."  "26."  "23."  "24."  "23."  "16."  "15."  "98."
## [11] "14."  "9."   "0."   "5."   "18."  "16."  "16."  "9."   "9."   "21."
## [21] "69."  "7."   "0."   "9."   "29."  "49."  "7."   "19."  "0."   "179."
## [31] "177." "0."   "0."   "0."   "0."
##
## $`https://www.amazon.com/s?k=condominium&crid=1ZNT0FW4JKRC3&sprefix=condominiu%2Caps%2C312&ref=nb_sb_noss_2&page
##  [1] "0."  "15." "12." "26." "23." "16." "15." "24." "23." "7."  "0."  "0."
## [13] "5."  "9."  "16." "9."  "18." "16." "9."  "21." "69." "98." "14." "9."
## [25] "7."  "19." "0."  "7."  "0."  "18." "10." "0."  "0."  "0."  "0."
##
## $`https://www.amazon.com/s?k=condominium&crid=1ZNT0FW4JKRC3&sprefix=condominiu%2Caps%2C312&ref=nb_sb_noss_2&page
## [1] "0."  "0."  "0."  "9."  "29." "49." "9."  "0."  "12."
```

```r
do.call("c",lapply(pages,Price))
```

```
##  [1] "0."   "15."  "12."  "26."  "23."  "16."  "15."  "24."  "23."  "7."
## [11] "0."   "0."   "5."   "9."   "16."  "9."   "18."  "16."  "9."   "21."
## [21] "69."  "98."  "14."  "9."   "7."   "19."  "0."   "7."   "0."   "18."
## [31] "10."  "0."   "0."   "0."   "0."   "0."   "15."  "12."  "26."  "23."
## [41] "24."  "23."  "16."  "15."  "98."  "14."  "9."   "0."   "5."   "18."
## [51] "16."  "16."  "9."   "9."   "21."  "69."  "7."   "0."   "9."   "29."
## [61] "49."  "7."   "19."  "0."   "179." "177." "0."   "0."   "0."   "0."
## [71] "0."   "0."   "0."   "18."  "10."  "29."  "49."  "9."   "0."   "9."
```