

STQD6114

UNSTRUCTURED TEXT MINING



NOR HAMIZAH MISWAN

Retrieving Textual Data

- ❖ Text mining (also called text data mining or text analytics) is at its simplest, a method for drawing out content based on meaning and context from a large bodies of text.
- ❖ It is a method for gathering structured information from unstructured text.
- ❖ Text data is everywhere. According to estimates, 80% of the world's data is in “unstructured text format”.
- ❖ We need methods to extract, summarize, and analyze useful information from unstructured/text data.
- ❖ Text mining seeks to automatically discover useful knowledge from the massive amount of data.
- ❖ Use of computational techniques to extract high quality information from text.



Retrieving Textual Data

- ❖ Unstructured text is present in various forms, and in huge and ever-increasing quantities:
 - books
 - financial and other business reports
 - various kinds of business and administrative documents
 - new articles
 - blog posts
 - wiki
 - messages/posts on social networking and social media sites



Extract Text from Files

- ❖ You may use the usual `read.table()` or `read.csv()` function.
- ❖ However, by using `read.table()` or `read.csv()` function, one can only run text operations on one document at a time.
- ❖ A common method for retrieving textual data in R is by using the `tm` package.
- ❖ The main data structure in the package is a corpus.
- ❖ A corpus is a collection of text documents.
- ❖ The benefit of using corpus is that it provides the ability to run text operations on the entire corpus.



Extract Text from Files

- ❖ There are two types of corpus: Vcorpus and Pcorpus.
- ❖ Vcorpus is the default implementation in R, which is short for volatile corpus. It is volatile since once the R object is destroyed, the whole corpus is gone.
- ❖ Meanwhile, the Pcorpus is short for permanent corpus since the documents are physically stored outside of R.
- ❖ There are predefined sources in the package:
 - DirSource – handle a directory
 - VectorSource – handle a vector
 - DataframeSource – handle data rframe like structures



Web Scraping

- ❖ There is a massive amount of data available on the web.
- ❖ However, getting the data into an analyzable format is difficult.
- ❖ Assessing online data of this sort is called “web scraping”.
- ❖ A key challenge is finding a way to unpack the data from a web page full of other elements.
- ❖ You may use the `readLines()` function that does not require any package. This function reads some or all text lines from a connection. It allows for simple access to webpage source.
- ❖ You may also use the `httr` package. Within this package, there is a function `GET()` which access the web.



Web Scrapping

- ❖ Reading from the web will give you a messy result. The result sometimes needs to be restructured and parsed. The `htm_treeParse()` function from the XML package is tailored for this task.
- ❖ Another common package is the rvest package. In this package, there is a function `read_html()` which reads url into memory similar to the way we read csv files using `read.csv()` function. Then, we can extract pieces out of the HTML documents using Xpath and css selectors (selector gadget). By using css selectors to scrap, web scraping can be performed for targeted content since not all content on the web page is gold.
- ❖ We can also perform web scraping on multiple pages.



Extract through API

- ❖ API is an Application Programming Interface. It is a software that allows two applications to talk to each other. It is a set of programming code that enables data transmission between one software product to another.
- ❖ API is also referred to as sets of protocol that govern interactions between sites and users. For public APIs, many sites only allow authorized users; eg. Twitter, Facebook, Instagram, Github, etc.



Extract through API

Twitter APIs

- ❖ Twitter API enables programmatic access to Twitter in unique and advanced ways. It can be used to analyze and interact with Tweets. You must sign and log in to Twitter Developer to obtain consumer key, consumer secret, access token and access secret. In order to sign in to Twitter Developer, you need to already have a Twitter account.
- ❖ When you sign in, you need to provide a Callback URL. This must be exact. You may use the following Callback URL: <http://127.0.0.1:1410>



Extract through API

Twitter APIs

❖ Example of consumer key and consumer secret:

##these are fake

```
consumer_key<-"XYznzPFOFZR2a39FwWKN1Jp41"
```

```
consumer_secret<-"CtkGEWmSevZqJuKI6HHrBxbCybXl1xGLqrD5ynPd9jG0SoHZbD"
```

- ❖ Twitter provides two types of API: Twitter's REST API and Twitter's STREAM API.
- ❖ Every day many thousands of developers make requests to the Twitter API. To help manage the sheer volume of these requests, limits are placed on the number of requests that can be made. The maximum number of requests that are allowed is based on a time interval, some specific period or window of time. The most common request limit interval is fifteen minutes. If an endpoint has a rate limit of 900 requests/15 minutes, then up to 900 requests over any 15-minute interval is allowed. Rate limits are applied based on which authentication method you are using.

