Importazione dei testi

Importazione di file di testo in R, in formati adatti all'uso con principali pacchetti per il text mining e l'analisi testuale.

I **dati degli esempi** sono disponibili a questo link. (in aggiornamento).

Vedi anche: *Strumenti per l'analisi testuale e il text mining con R*

Importazione di un solo testo

Con il pacchetto *readtext*, il file di testo viene importato come dataframe (funzione `readtext()`).

```
library(tidyverse)
library(readtext)
```

```
# sepolcri
sepolcri <- readtext("dati/sepolcri.txt")
str(sepolcri)
```

```
Classes 'readtext' and 'data.frame':    1 obs. of  2 variables:
 $ doc_id: chr "sepolcri.txt"
 $ text  : chr "All'ombra de' cipressi e dentro l'urne\nConfortate di
pianto è forse il
sonno\nDella morte men duro? Ove più il"| __truncated__
```

```
# articolo
articolo <- readtext("dati/ansa_2020-02-05.txt")
str(articolo)
```

```
Classes 'readtext' and 'data.frame':    1 obs. of  2 variables:
 $ doc_id: chr "ansa_2020-02-05.txt"
 $ text  : chr "##TITOLO Sanremo: Fiorello al festival mai più
ospite, magari in gara\n\n##SOTTOTITOLO \"Mi
ha colpito Achille "| __truncated__
```

I nomi dei campi sono quelli riconosciuti per default dalle funzioni di Quanteda, Tidytext e tm, ovvero:

- `doc_id`: identificativo di documento.
- `text`: campo con il testo da analizzare.

È possibile importare anche testi organizzati in dataset (formati: csv, tab, tsv) e documenti di Word e di OpenOffice. Per l'esattezza, i formati di documento accettati sono: txt, odt, doc, docx, ma anche rtf,pdf, json, html, e xml.

Importare più testi organizzati in una cartella

```
canzoni <- readtext('dati/canzoni/*.txt')
canzoni
```

```
readtext object consisting of 2 documents and 0 docvars.
# Description: df [2 x 2]
  doc_id               text
  <chr>                <chr>
1 01_Achille Lauro.txt "\"Achille La\"..."
2 02_Alberto Urso.txt  "\"Alberto Ur\"..."
```

Vengono così importati tutti i file della cartella, e nel campo `doc_id` viene conservato il nome dei file.

Le variabili relative ai singoli documenti possono essere estrapolate dal nome o dal percorso del file (le variabili saranno i nomi delle cartelle: nell'esempio che segue, `docvarsfrom = "filenames"`).

```
canzoni <- readtext('dati/canzoni/*.txt',
                    docvarsfrom = "filenames",
                    docvarnames = c("ID",
                                    "cantante"))

canzoni
```

```
readtext object consisting of 2 documents and 2 docvars.
# Description: df [2 x 4]
  doc_id               text                    ID cantante
  <chr>                <chr>                <int> <chr>
1 01_Achille Lauro.txt "\"Achille La\"..."     1 Achille Lauro
2 02_Alberto Urso.txt  "\"Alberto Ur\"..."     2 Alberto Urso
```

Le variabili possono anche essere importate da un file csv (con un identificativo che associ le variabili al testo).

Il fatto che i testi siano organizzati in un dataframe permette di usare le funzioni per la gestione dei dati, dopo l'importazione. Ad esempio, potremmo voler trasformare una variabile "data" nel più appropriato formato *Date*:

```
articolo$data <- as.Date(articolo$data)

articolo
```

```
readtext object consisting of 1 document and 2 docvars.
# Description: df [1 x 4]
  doc_id              text                fonte data
  <chr>               <chr>               <chr> <date>
1 ansa_2020-02-05.txt "\"##TITOLO S\"..." ansa  2020-02-05
```

Importare un dataset con i testi

Nel caso di testi brevi o molto brevi, come le risposte aperte a un questionario, o i messaggi di Twitter, i dati sono generalmente raccolti e organizzati in un dataframe. In questo caso:

- non useremo `readtext()`, ma le funzioni del pacchetto *readr* (quindi potremo importare il dataset anche dal menu di RStudio);
- potremo strutturare le variabili anche prima dell'importazione.

Vediamo qui il dataset contenente le canzoni di Sanremo 2021, in formato tab delimited (tsv, anche se con estensione .csv):

```
sanremo21 <- read_tsv("dati/sanremo2021.csv")
head(sanremo21)
```

```
# A tibble: 6 x 3
  cantante           titolo              testo
  <chr>              <chr>               <chr>
1 Aiello             Ora                 "Ora ora ora ora\nMi parli
come allora~
2 Annalisa Scarrone  Dieci               "Cos'è che ti ho
promesso\nNon so\nNon~
3 Arisa              Potevi Fare Di Più  "Lasciarsi adesso non fa più
male non ~
4 Bugo               E Invece Sì         "Le metropolitane vanno
molto veloci\n~
5 COLAPESCEDIMARTINO Musica Leggerissima "Se fosse un'orchestra a
parlare per n~
6 Coma Cose          Fiamme Negli Occhi  "Quando ti sto vicino
sento\nChe a vol~
```

L'importazione può essere effettuata anche da RStudio (*Import Dataset*).

Organizziamo i nomi dei campi secondo gli standard visti sopra:

```
sanremo21 <- sanremo21 %>%
  # titolo come doc_id
  mutate("doc_id" = str_replace_all(titolo, " ", "_")) %>%
  # rinominiamo il campo "testo" e riorganizziamo le colonne
  select(doc_id, "text" = testo,
         cantante, titolo)
head(sanremo21)
```

```
# A tibble: 6 x 4
  doc_id              text                                  cantante
titolo
  <chr>               <chr>                                 <chr>
<chr>
1 Ora                 "Ora ora ora ora\nMi parli come a~ Aiello
Ora
2 Dieci               "Cos'è che ti ho promesso\nNon so~ Annalisa S~
Dieci
3 Potevi_Fare_Di_Più  "Lasciarsi adesso non fa più male~ Arisa
Potevi Far~
4 E_Invece_Sì         "Le metropolitane vanno molto vel~ Bugo
E Invece Sì
5 Musica_Leggerissima "Se fosse un'orchestra a parlare ~ COLAPESCED~
Musica Leg~
6 Fiamme_Negli_Occhi  "Quando ti sto vicino sento\nChe ~ Coma Cose
Fiamme Neg~
```

Script di esempio

E' possibile scaricare ed eseguire lo script dell'esempio:

import_testi.R

```
library(tidyverse)
library(readtext)

# sepolcri
sepolcri <- readtext("dati/sepolcri.txt")

# articolo
articolo <- readtext("dati/ansa_2020-02-05.txt")
articolo$data <- as.Date(articolo$data)

# canzoni
```

```
canzoni <- readtext('dati/canzoni/*.txt',
                    docvarsfrom = "filenames",
                    docvarnames = c("ID",
                                    "cantante"))

# dataset
sanremo21 <- read_tsv("dati/sanremo2021.csv")
sanremo21 <- sanremo21 %>%
  # titolo come doc_id
  mutate("doc_id" = str_replace_all(titolo, " ", "_")) %>%
  # rinominiamo il campo "testo" e riorganizziamo le colonne
  select(doc_id, "text" = testo,
         cantante, titolo)
```

Analisi testuale, Importazione, Gestione dei dati