Analisi di un testo inglese (gennaio 2016)

Mario Puppi

Un testo in *Wolfram Language* è una *stringa* ossia un sequenza di caratteri.

La funzione **DeleteStpwords** prende un testo e restituisce la sequenza di parole significative di cui esso è formato.

La funzione **WordCloud** prende una sequenza di parole e ne fornisce un wordcloud.

Prendiamo un testo inlgese su *Wikipedia* che parla di Dracula:

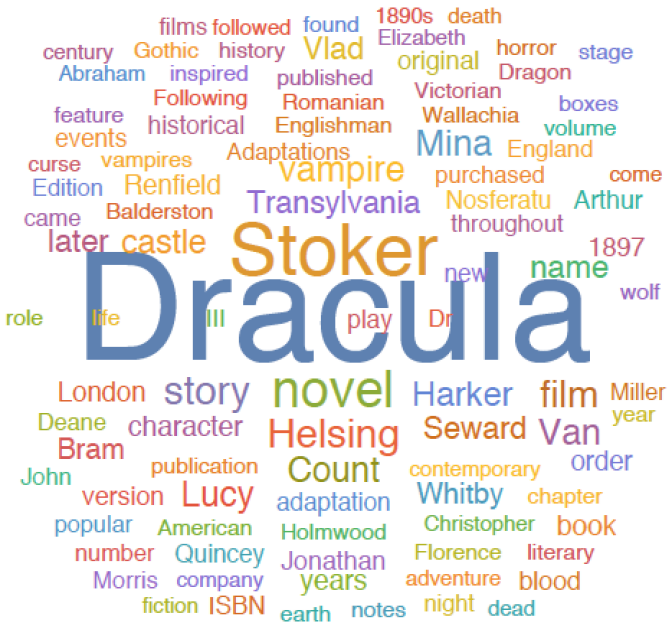
```
testo = WikipediaData["Dracula"] ;
```

Vogliamo ricavare la sequenza di parole significative presenti nel testo, eliminando segni di punteggiatura, gli articoli, le congiunzioni e le preposizioni quali "the", "with", "and", ... (parole dette *stopwords*)

```
parole = DeleteStopwords[testo];
```

E' divertente fare un *wordcloud* del testo che mette in evidenza le parole più frequenti:

```
WordCloud[parole]
```



StringCount[S, w] prende una sequenza **S** di parole e una parola **w** e dà come risultato la frequenza di **w** in **S**.

Possiamo misurare la frequenza nel testo di una parola, come "vampire" ad esempio:

```
StringCount[parole, "vampire"]
```

20

Definiamo la frequenza di una parola **w** nel testo dato:

```
fr[w] := StringCount[parole, w]
```

Definiamo un piccolo insieme di parole, un vocabolario:

```
Vocabolario = {"death", "blood", "vampire" };
```

e andiamo a misurare la loro frequenza nel testo:

```
Table[fr[w], {w,vocabolario}]
```

{3, 9, 20}

meglio mettere assieme la parola **w** con la frequenza **fr[w]**, formando la coppia **{w, fr[w]}**

```
Table[{w, fr[w]}, {w,vocabolario}]
```

{ {death, 3}, {blood, 9}, {vampire, 20} }

Dato un testo, cioè una stringa di caratteri, la funzione **TextWords** fornisce la sequenza di tutte le parole del testo.

Ad esempio, otteniamo le parole presenti nel testo "abc cd abc dc cvf"

```
TextWords["abc cd abc dc cvf"]
```

{abc, cd, abc, dc, cvf}

La funzione **Union** prende una lista di oggetti e fornisce l'insieme degli elementi della lista senza ripetizioni.

Usiamo **Union** per ottenere togliere le parole ripetute:

```
Union[TextWords["abc cd abc dc cvf"]]
```

{abc, cd, cvf, dc}

Possiamo così ottenere tutte le parole presenti nel testo senza ripetizioni:

```
Vocabolario = Union[TextWords[parole]] ;
```

La funzione **Length** conta il numero di elementi di una lista.

Contiamo le parole del vocabolario:

```
Length[Vocabolario]
```

1653

Esercizio 1. Prendere un testo inglese su *Wikipedia* che parla dei *Beatles* e ricavarne un testo significativo eleminando le stopwords (punteggiatura, articoli, preposizioni, avverbi, ecc..)

1.1 Fare un wordlcoud del testo significativo e misurare la frequenza delle parole "Lennon", "McCartney", "Starr", "Harrison", "rock", "music".

1.2 Usando le funzioni **TextWords, Union** determinare un vocabolario delle parole significative presenti nel testo.

Riprendiamo il testo su Dracula. Vogliamo fare una classifica delle parole del vocabolario in base alla loro frequenza.

Definiamo una relazione d'ordine **PiùFrequente** sulle parole del vocabolario:

```
PiùFrequente[x_, y_] := fr[x] ¥ fr[y]
```

La funzione **Sort[X, R]** ordina una lista di oggetti **X** usando la relazione d'ordine **R**.

Ordiniamo il vocabolario secondo la relazione **PiùFrequente**

```
classifica = Sort[Vocabolario, PiùFrequente];
```

Prendiamo le prime 100 parole in classifica:

```
classifica = Take[classifica, 100]
```

Possiamo determinare la posizione in classifica di ogni parola:

```
Position[classifica, "Lucy"]
```

{{27}}

Position fornisce una lista **{{n}}** formata da un unico elemento, la lista **{n}**.
Per estrarre la posizione **n** usiamo il selettore **[[1,1]]**

Definiamo una funzione **rng** che determina la posizione in classifica di ogni parola (*rango* della parola):

```
rng[w_] := Position[classifica, w][[1, 1]]
```

Definiamo una funzione **S** che calcola assieme il rango e la frequenza di ogni parola:

```
S[w_]:= {rng[w], fr[w]}
```

Ad esempio,

```
S["vampire"]
```

{20, 20}

S determina una relazione, data dall'insieme delle coppie {rng[w], fr[w]}, per dtutte le parole w del vocabolario.

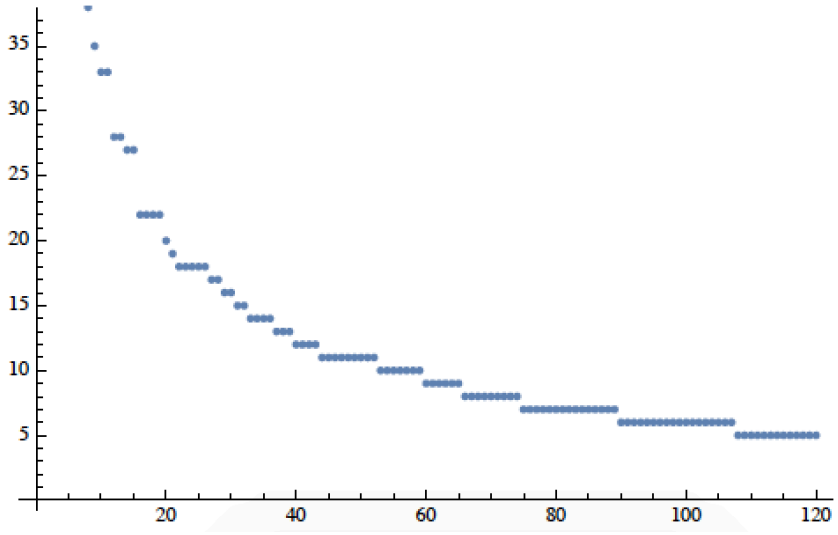
Vogliamo indagare sulla relazione tra il rango e frequenza delle parole e a questo scopo faremo un grafico cartesiano della relazione **S**.

Data una lista **L** e un numero **n**, **Take[L, n]** fornisce la lista dei primi **n** elementi di **L**.

Se **S** è una lista di coppie di numeri, **ListPlot[S]** disegna tutti gli elementi di **S** nel piano cartesiano.

Ci limitamo a fare il grafico della relazione S ristretta all'insieme **V** delle prime 120 parole in classifica del vocabolario.

```
V = Take[classifica, 120];
ListPlot[Table[S[w], {w, V}]]
```



Il grafico della relazione **S** sembra il grafico di una relazione di proporzionalità inversa. Vogliamo verificare se è vero che esiste una costante **k** tale che la frequenza **f** di una parola sia legata al suo rango **r** dalla relazione **f = k/r.**

```
V = Take[classifica, 120];
ListPlot[Table[S[w], {w, V}]]
```

Determiniamo in modo statistico **k** come media del prodotto **f r** calcolato sulle prime 120 parole del testo:

```
k = Sum[rng[w]/fr[w], {w, V}] / 120.0
```

520.758

Verifichiamo la buona approssimazione data dal modello proporzionale inverso. Confrontiamo il grafico della relazione **S** con quello della relazione di proporzionalità inversa di costante **k** = 120.

```
Show[ListPlot[Table[S[w], {w, V}]] ,
     Plot[k / x, {x, 5, 120}, PlotStyle → Red]]
```