R Notebook

Christophe@pallier.org

Contents

On télécharge d'abord la table Lexique3:

```
require(rjson)
```

```
## Loading required package: rjson
```

```
source('https://raw.githubusercontent.com/chrplr/openlexicon/master/datasets-info/fetch_datasets.R')
```

```
## Loading required package: tools
```

```
lexique <- get_lexique383()
```

```
## Warning in fetch_dataset("Lexique383", format = "rds"): You already have
## the file /home/cp983411/openlexicon_datasets/Lexique383.rds which seems up
## to date.
```

Puis, supprimons les items de très basse fréquence (inférieure à 0.5 par million):

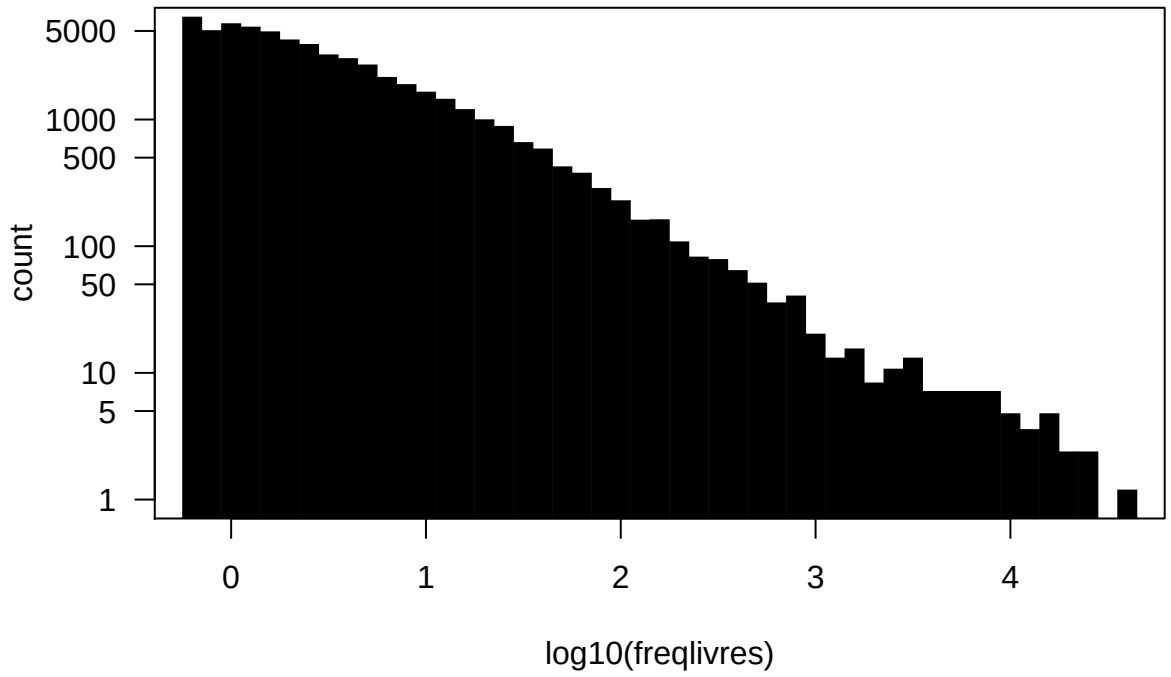
```
lexique1 = subset(lexique, freqlivres > 0.5)
lexique1$logfreq <- log10(lexique1$freqlivres)
```

Distribution des fréquences lexicales

Calculons l'histogramme des fréquences par million dans le corpus des livres, et affichons sur un graphique avec des axes logarithmiques:

```
with(lexique1, {
  histdata <- hist(logfreq, plot=FALSE, nclass=50)
  plot(histdata$breaks[-1], histdata$count, log="y", type='h', lwd=10, lend=2, las=1, xlab='log10(freql
  }
)
```

```
## Warning in xy.coords(x, y, xlabel, ylabel, log): 1 y value <= 0 omitted
## from logarithmic plot
```



Note: On retrouve la loi de Zipf, c'est à dire une relation à peu près linéaire sur ce graphique log-log, qui reflète une distribution en loi de puissance.

Exemples de noms tirés dans différentes bandes de fréquence:

```
b1 <- subset(lexique1, ((logfreq < 0.1) & (cgram == 'NOM') & (islem==1)), c('ortho',  'freqlivres', 'log
b1[sample(1:nrow(b1), 10),]
```

```
##               ortho freqlivres      logfreq
## 59380       fripier       0.61 -0.214670165
## 4368           alfa       0.88 -0.055517328
## 22316  centre-ville       0.54 -0.267606240
## 21241         carie       0.61 -0.214670165
## 64492     gueuleton       0.74 -0.130768280
## 63696       griffon       1.01  0.004321374
## 78396      lynchage       0.68 -0.167491087
## 119147        scout       0.95 -0.022276395
## 69184    impuissant       0.95 -0.022276395
## 53342      excision       0.81 -0.091514981
```

```
b2 <- subset(lexique1, ((logfreq > 1) & (logfreq < 1.1) & (cgram == 'NOM') & (islem==1)), c('ortho',  'f
b2[sample(1:nrow(b2), 10),]
```

```
##             ortho freqlivres  logfreq
## 122970  splendeur      10.95 1.039414
## 4323       alerte      10.47 1.019947
## 85677    médecine      12.36 1.092018
## 99769  prisonnier      11.69 1.067815
## 15270      blessé      11.49 1.060320
## 86154      mérite      11.62 1.065206
## 138446  éclairage      10.74 1.031004
```

```
## 93507       pause      10.14 1.006038
## 77565  locomotive      10.61 1.025715
## 30186   continent      12.16 1.084934
```

```
b3 <- subset(lexique1, ((logfreq > 1.5) & (logfreq < 1.6) & (cgram == 'NOM') & (islem==1)), c('ortho',
b3[sample(1:nrow(b3), 10),]
```

```
##               ortho freqlivres  logfreq
## 70108         index      32.43 1.510947
## 98538        portée      33.18 1.520876
## 31348         corde      31.76 1.501880
## 83162           mme      33.24 1.521661
## 57240        fleuve      39.32 1.594614
## 18124    brouillard      32.84 1.516403
## 132981        tâche      35.95 1.555699
## 80781      maréchal      38.24 1.582518
## 14418  bibliothèque      36.82 1.566084
## 51012  enthousiasme      33.72 1.527888
```

```
b4 <- subset(lexique1, ((logfreq > 2.0) & (logfreq < 2.1) & (cgram == 'NOM') & (islem==1)), c('ortho',
b4[sample(1:nrow(b4), 10),]
```

```
##            ortho freqlivres  logfreq
## 119686 sentiment     106.42 2.027023
## 39002     droite     116.69 2.067034
## 95584     pierre     119.39 2.076968
## 68197      image     119.39 2.076968
## 75959     langue     103.78 2.016114
## 135641   village     118.24 2.072764
## 1156      accord     124.66 2.095727
## 74703    journal     124.32 2.094541
## 76942      ligne     101.01 2.004364
## 81330     maître     125.74 2.099473
```

```
b5 <- subset(lexique1, ((logfreq > 2.5) & (logfreq < 2.6) & (cgram == 'NOM') & (islem==1)), c('ortho',
b5[sample(1:nrow(b5), 10),]
```

```
##          ortho freqlivres  logfreq
## 22699  chambre     380.07 2.579864
## 81072    matin     376.89 2.576215
## 93757    peine     388.24 2.589100
## 55015     fait     325.27 2.512244
## 121269  soleil     328.78 2.516905
## 49443   enfant     381.96 2.582018
## 38703    doute     341.35 2.533200
## 87559      nom     326.89 2.514402
## 126699   table     341.08 2.532856
## 84292     mort     373.99 2.572860
```

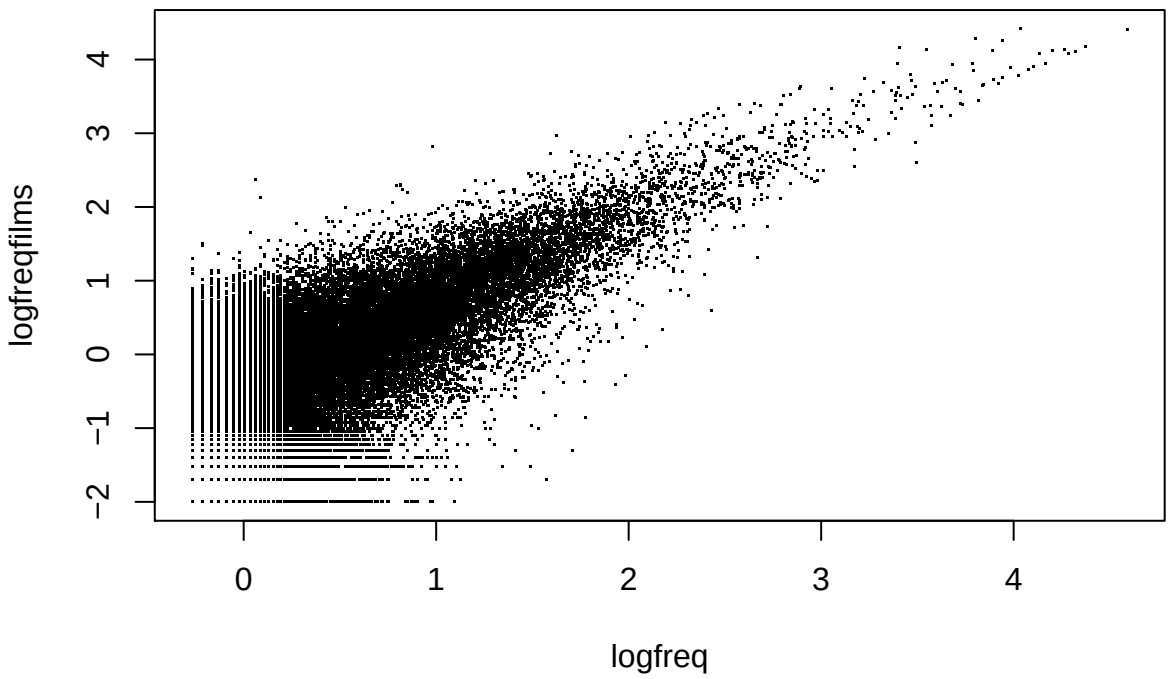
Comparaison d'estimations de fréquences

Lexique3 fournit également des fréquences lexicales estimées à partir d'un corpus de sous-titres de films (`freqfilms2`).

Examinons la relation entre les fréquences estimées sur les libvres (corpus Frantext) et celles estimées à partir

des sous-titres.

```
lexique1$logfreqfilms = log10(lexique1$freqfilms2)
with(lexique1, plot(logfreq, logfreqfilms, pch='.'))
```



Pour sélectionner des mots peu fréquents, pour lesquels les fréquences peuvent être mal estimées, il peut être une bonne idée de combiner les critères sur les deux type de fréquences.