Problème pratique de statistique n° pps076

Occurrences des 26 lettres de l'alphabet dans les langues européennes

Z. Cienikovà 4BIM 2005/2006

1 Introduction

On se propose d'analyser la distribution des lettres dans différentes langues européennes. On ne s'intéressera qu'aux langues d'alphabet latin et plus particulièrement aux langues officielles de l'Union Européenne. Il y a plusieurs manières de s'y prendre.

On peut s'intéresser aux fréquences déterminées non sur l'ensemble du vocabulaire de la langue considérée, mais sur l'ensemble représentatif des textes de cette langue. Les distributions sont alors tirées des sources officielles ou, à défaut, on se sert des distributions calculées sur un échantillon important (d'ordre de 5M de caractères).

On peut aussi calculer les fréquences dans un texte d'ordre de 30K commun à toutes les langues. C'est ce qui a été fait ici. Pour éviter la variabilité due à la nature hétérogène des textes, on utilise les traductions du Traité de la Constitution Européenne, partie II. Les langues disponibles sont dans le tableaux 1.

2 Sources

On a utilisé :

http://de.wikipedia.org/wiki/Buchstabenh%C3%A4ufigkeit http://www.cs.bris.ac.uk/Teaching/Resources/COMS30124/Labs/freq.html http://nlp.fi.muni.cz/nlp/aisa/NlpCz/Frekvence_pismen_bigramu_trigramu_delka_slov.html http://www.characterfrequency.com/part1.html http://access.adobe.com/access/#form http://www.lexilogos.com/clavier/multilingue.htm http://patternmedia.com/projects/character_frequency_analyzer/index.php http://www.central.edu/homepages/LintonT/classes/spring01/cryptography/java/textanalyzer.html http://textalyser.net/index.php?lang=en#analysis http://europa.eu.int/constitution/print_de.htm http://en.wikipedia.org/wiki/European_languages

N°	Code	Langue
1	cz	tchèque
2	da	danois
3	de	allemand
4	en	anglais
5	es	espagnol
6	et	estonien
7	fi	finnois
8	fr	français
9	ga	gaélique (irlandais)
10	hu	hongrois
11	it	italien
12	lt	lithuanien
13	lv	letton
14	mt	maltais
15	nl	néerlandais
16	pl	polonais
17	pt	portugais
18	sk	slovaque
19	sl	slovène

TABLE 1 – Les 19 langues étudiées.

3 Méthodes

Pour quantifier les occurrences des lettres dans le texte du traité constitutionnel, on transforme les fichiers .pdf disponibles en ligne en texte plain. Un outil pour cette tâche est accessible sur le site de Adobe R. Les caractères spéciaux sont ensuite remplacés par ceux de l'alphabet standard. Il suffit d'un éditeur de texte classique, pour plus d'efficacité on peut écrire un script console (sous Linux). On se sert alors des analyseurs de texte, également disponibles sur des sites dédiés. On obtient en sortie un fichier de texte de 26 lignes avec des lettres et leurs occurrences en colonne.

4 Jeu de données

On en disposera par :

```
tab <- read.table("http://pbil.univ-lyon1.fr/R/donnees/pps076.txt", header = TRUE)
#tab <- read.table("pps076.txt", header = TRUE)
head(tab)
```

```
   cz   da   de   en   es   et   fi   fr   ga   hu   it   lt   lv   mt   nl   pl
A 2513 2090 1287 1512 3716 3415 5018 1497 6003 4281 3252 3236 4426 3428 2110 3271
B  448  415  379  271  408  200   18  173  612  481  293  453  869  477  584  364
C  905  106  668  825 2086    9   11  777 1665  276 1206  354  388  181  557 1426
D  877 1611 1254  722 1898 1168  335  946 1025  477 1392  680  818 1061 2055 1032
E 2233 5100 4300 2501 4049 2723 2502 3694 2111 3952 3798 2071 2156 1735 6566 2373
F   38  842  422  518  196   22    6  175  352  171  213   53   47  360  443   52
   pt   sk   sl
A 4126 3159 2655
B  342  457  406
C 1804  916  714
D 1957  925 1013
E 3660 2230 2497
F  183   36   25
```

```
head(t(tab))
```

```
      A   B    C    D    E   F    G    H    I   J    K    L   M    N    O    P   Q
cz 2513 448  905  877 2233  38   46  636 2285 397  990 1037 649 1996 2438 1091   1
da 2090 415  106 1611 5100 842 1293  441 2421 112  999 1767 819 2348 1828  495   0
de 1287 379  668 1254 4300 422  849 1024 1991  58  244  718 416 2465  562  164   2
en 1512 271  825  722 2501 518  365  951 1973  27   43  793 411 1648 1639  473  16
es 3716 408 2086 1898 4049 196  250  207 3094 106    1 2008 517 2348 3224  916 111
et 3415 200    9 1168 2723  22  725  547 3394 481 1368 1558 616 1128 1859  458   1
      R    S    T    U    V   W  X   Y   Z
cz 1272 1478 1371 1091 1259   0  4 663 661
da 2813 1722 2189  548  568   1  2 128   0
de 1849 1351 1567 1247  218 241  3   2 301
en 1360 1164 2021  501  220 187 38 291  14
es 2140 2217 1610 1168  271   0 42 308  55
et  944 2198 2221 1823  566   0  2   0   0
```

5 Questions

Le critère étudié induit une typologie des langues. Quelle est la nature de cette structure ? Comment en rendre compte ? Est-ce une question d'ACP, d'AFC ou de PCO ? Une question de classification, d'ordination, de modèlisation phylogénétique ? Un élément pour la discussion :

```
library(ade4)
w <- as.data.frame(t(tab))
coaw=dudi.coa(w,scannf=F, nf=5)
dw=dist.dudi(coaw)
plot(hclust(dw),cex=2,main="Distance du khi2",sub="")
```