Projet sur Scilab (LM206) — Matilde RODRIGUES
Linguistique : fréquence des mots — Anne REGNIER

Dans un premier temps, nous allons présenter l'objectif de notre projet et les outils utilisés pour le réaliser.

Le thème de notre projet est le calcul de la fréquence des mots dans un texte. Contre toute attente, ce type de calcul a des utilités intéressantes. En effet, grâce à la comparaison de la fréquence des champs lexicaux dans deux textes, on a pu émettre l'hypothèse que Corneille serait l'auteur de certains textes voire d'œuvres complètes de Molière.

Cependant, les qualités de ce calcul peuvent se mesurer dans d'autres domaines, notamment dans l'étude de textes politiques. Nous en avons fait l'expérience au cours de notre projet en nous appuyant sur les discours pour l'investiture socialiste de Ségolène Royal, Dominique Strauss-Kahn et Laurent Fabius. Nous avons pu remarquer que les mots apparaissant le plus souvent illustrent les personnalités de chacun des orateurs et donnent les idées directrices de leurs propos.

Au cours de ce projet, divers problèmes sont survenus dont celui-ci : comment travailler avec des chaînes de caractères sur Scilab alors que jusqu'alors nous n'avions utilisé que des fonctions relatives aux nombres ?

Nous avons résolu notre problème en transformant notre texte en matrice colonne. Voici les fonctions traitant les matrices qui ont concourues à la réalisation de notre projet :

- size : fonction utilisée pour le calcul de la taille la matrice.

- length : fonction permettant de déterminer la longueur des différents éléments d'une matrice. Ici, elle permet de calculer le nombre de lettres par mot.

- sort : fonction qui classe les éléments de la matrice par ordre alphabétique.

- tabul : fonction qui crée une matrice avec les éléments de la matrice de départ associés à leurs fréquences d'apparition dans le texte.

- find : fonction qui donne les indices des éléments satisfaisant une condition qu'on définit.

Cette dernière fonction nous a permis d'affiner notre projet en sélectionnant les mots les plus importants, c'est-à-dire ceux qui ne répondent pas aux critères suivants : articles, prépositions, conjonctions. Nous éliminons aussi les mots dont la fréquence n'est pas comprise entre 2 et 9 car il nous semble que ces derniers ne peuvent être considérés comme représentatif de la façon de s'exprimer des trois Hommes politiques. Ainsi, la matrice finale obtenue ne comprend qu'une dizaine de mots caractéristiques, on l'espère, de la personnalité des candidats.

Dans un second temps, nous allons présenter la retranscription exacte de notre projet en terme de Scilab. Pour ce faire, nous allons nous appuyer sur une question posée aux trois socialistes et nous allons nous intéresser à la fréquence des mots dans leur réponse.

La question que nous avons choisie concerne leur opinion et agissements face au conflit israélo-palestiniens : « Dans le conflit israélo-palestinien, quelle sera la position de votre gouvernement ? »
(Les différentes réponses sont extraites de « l'hebdo des socialistes » sur le site du PS).

Nous allons présenter ici le travail effectué en s'appuyant sur le texte de Ségolène Royal et nous donnerons par la suite les résultats obtenus pour Laurent Fabius et Dominique Strauss-Kahn.

Au préalable, nous avons du enregistrer le texte réponse de Madame Royal sur le bureau afin de l'intégrer à Scilab et le placer dans une matrice puis le concaténer et faire en sorte que la matrice est autant de lignes que le nombre de mots et un mot par ligne. Nous avons procédé de la manière suivante :

```
-->sego = read ("sego.txt",-1,1,'(a)')

-->sego1 = strcat (sego)

-->sego2 = tokens (sego1)
```

A la matrice ainsi obtenue nous avons ensuite retiré les mots ne satisfaisant pas les conditions qui nous ont semblées intéressantes. Dans un premier temps nous avons retiré les mots de plus de 3 lettres qui sont souvent des mots de liaisons, notamment des prépositions. Nous avons utilisé la méthode suivante :

```
-->tmp = length (sego2)

-->tmp2 = find (tmp>3)

-->sego3 = sego2 (tmp2)
```

Cette troisième matrice ne comprend que les mots significatifs, nous pouvons donc calculer les fréquences des mots dans cette dernière :

```
-->sego4 = tabul (sego3)
```

Nous allons enfin extraire les fréquences comprises entre 3 et 8 dans le texte, fréquences que nous pouvons désigner comme étant celles des mots les plus importants du texte, hormis les mots de liaisons :

```
-->tmp3 = find (sego4(2)>1 & sego4(2)<9)

-->sego5 = sego4(1)(tmp3)
```

"sego5" est la matrice finale, voici son contenu:

```
!État     !
!         !
!sécurité  !
```

```
!        !
!sont     !
!        !
!région    !
!        !
!pour     !
!        !
!plus     !
!        !
!paix     !
!        !
!majorité  !
!        !
!l'Europe  !
!        !
!juste    !
!        !
!droit    !
!        !
!dans     !
```

Pour finir, nous pouvons ajouter le dernier programme que nous avons utilisé afin de retrouver les fréquences de ces mots. Cette dernière manipulation peut servir si l'on s'intéresse par la suite à la construction graphique :

```
-->size(sego5)
-->frq = zeros(12,1)

-->for i = 1:12
-->frq(i) = length ( find(sego2==sego5(i) ))
-->end

-->frq
 frq  =

   2.
   2.
   2.
   2.
   4.
   2.
   4.
   2.
   2.
   2.
   2.
   2.
```

A titre indicatif, nous pouvons remarquer que les fréquences significatives sont finalement comprises entre 2 et 4.

En complément, voici la matrice obtenue pour Dominique Strauss-Kahn :

!sécurité !
! !
!sans !
! !
!retrait !
! !
!pour !
! !
!parvenir !
! !
!négociation, !
! !
!entre !
! !
!enfin !
! !
!d'un !
! !
!droit !
! !
!dans !
! !
!conditions !
! !
!cette !
! !
!avec !
! !
!agir !
! !
!Liban !

Voici celle de Laurent Fabius :

!tant !
! !
!reconnaissance !
! !
!pour !
! !
!d'un !
! !
!droit !
! !
!dialogue !
! !
!deux !
! !
!dans !
! !
!aura !

! !
!Palestiniens !
! !
!Israéliens !

Nous pouvons faire un bilan de cette étude. Il est intéressant de comparer les trois matrices car elles reflètent les traits de caractère souvent évoqués à propos des trois socialistes lors de la campagne, au sujet de l'international. Dominique Strauss-Kahn est le spécialiste du sujet ; il utilise des mots d'action et il a déjà bien pensé le sujet. Laurent Fabius a une méthode plus conciliante, il cherche le dialogue et la reconnaissance entre les états. Quant à Ségolène Royal, plus en retrait sur les questions internationales, elle ne se positionne pas vraiment et utilise de mots plus généraux.

Pour conclure quant à l'utilisation de Scilab, nous pouvons dire que notre projet a été compliqué par le fait que Scilab ne propose que peu de fonctions relatives aux matrices de chaînes de caractères. Nous avons passé beaucoup de temps à les trouver. Néanmoins, le résultat nous a relativement satisfaits car nous avons tout de même pu extraire des informations, à l'issue de nos calculs, concernant le caractère des trois socialistes.