Lizon Claire
Sengelin Le Breton Marine

Compte-rendu

La linguistique

Projet Initiation à Scilab

Au cours de nos séances, nous nous sommes penchées sur le thème de la linguistique et nous avons plus précisément travaillé sur les vœux du président Jacques Chirac en 2001, 2002,

2003 et 2004. Pour se faire, nous avons utilisé les fichiers de données qui étaient disponibles sur http://nicolas.limare.net/tmp/. Puis nous avons dû rendre ces textes utilisables pour notre travail par le logiciel de calcul numérique SCILAB et de ce fait, nous n'avons travaillé que sur des fichiers texte brut (pas de fichier mis en page .doc), sans accents, ni ponctuation. Nous avons alors utilisé la commande READ qui transfère le fichier donné en argument dans le tableau de sortie, dont chaque élément contient une chaîne de caractères issue d'une ligne du fichier d'origine. Nous avons ensuite assemblé ces lignes par la commande STRCAT, puis il a été utile de séparer les mots par la commande TOKENS qui coupe au niveau des espaces. Avant de pouvoir commencer réellement notre travail sur chacun des textes, nous nous sommes aperçues que bien qu'ayant utilisé cette dernière commande, certains espaces manquaient entre les mots, ce qui faisait que nous avions parfois deux mots concaténés, ce qui faussait alors le contenu du texte initial. Il était donc nécessaire de créer une nouvelle fonction « ajoutant » des espaces entre chaque mot du texte, afin de pouvoir faire une étude correcte de ce dernier. Après appel de cette fonction, nous obtenions ainsi un vecteur colonne avec pour éléments les mots du texte et nous pouvions alors commencer notre travail sur SCILAB, consistant à calculer la fréquence d'utilisation des mots, affiner la méthode pour n'en conserver que les mots suffisamment significatifs, comparer à l'aide de fonctions définies des textes deux à deux et enfin évaluer l'évolution des vœux du président Jacques Chirac au cours du temps.

I) Les outils et techniques utilisés

Tout d'abord, SCILAB est un environnement agréable pour faire du calcul numérique et son utilisation comme calculette matricielle nous a été d'un grand profit puisque la totalité de notre travail se faisait à l'aide de matrices et plus précisément de vecteurs colonnes qui correspondaient aux textes initiaux modifiés. De plus, SCILAB possède une grande bibliothèque de fonctions prédéfinies et nous les avons donc utilisées autant que possible. Les fonctions nous ayant été le plus profitables sont :

-SIZE qui rend la taille d'un objet. Plus particulièrement, appliquée à une matrice x avec un seul argument en sortie, SIZE renvoie un vecteur 1*2= [Nombre de lignes, Nombre de colonnes]. Appelée avec deux arguments en sortie, SIZE renvoie le nombre de lignes et le nombre de colonnes.

-PROD qui rend le produit des termes d'une matrice. Pour un vecteur ou une matrice x, PROD(x) renvoie un scalaire égal au produit des termes de x.

Durant notre travail, nous avons utilisé PROD (SIZE (A)) où A est une matrice de chaînes de caractères pour avoir le nombre d'éléments de cette matrice. En effet, SCILAB possède une autre fonction prédéfinie LENGTH qui fournit le nombre d'éléments d'une matrice réelle ou complexe. Mais lorsqu'il s'agit de matrices de chaînes de caractères données en argument, LENGTH nous renvoie la longueur de chaque chaîne de caractères et non son nombre d'éléments et ne répondait donc pas à notre problème. Généralement, l'utilisation de PROD (SIZE (A)) servait à connaître le nombre de mots dans un texte après modification de ce dernier.

-FIND qui renvoie les indices d'un vecteur ou d'une matrice répondant à un certain critère.

Durant notre travail, FIND a été utilisé pour trouver des mots dont la fréquence était supérieure à la moyenne. Donc bien qu'elle n'est été utilisée que rarement, son rôle fut très important pour l'avancée de notre projet.

-UNIQUE qui extrait les composantes distinctes d'un vecteur. Si M est un vecteur réel ou vecteur avec des chaînes de caractères, UNIQUE (M) renvoie un vecteur contenant les valeurs distinctes présentes dans les termes de M classés par ordre croissant. Nous avons donc utilisé UNIQUE pour obtenir des matrices ne contenant qu'une seule occurrence de chaque matrice et ce par ordre alphabétique. L'appel de cette fonction nous a ainsi permis d'obtenir un texte trié, beaucoup plus intéressant pour la poursuite de notre travail.

Bien qu'en essayant d'utiliser au maximum des fonctions prédéfinies, nous avons été obligées de créer nos propres fonctions avec ou sans boucle. Il a donc été nécessaire de travailler de pair avec un éditeur de texte (et plus précisément le propre éditeur intégré de SCILAB: SCIPAD). Bien que les boucles soient à éviter au maximum, le tiers de nos fonctions nécessitaient leur utilisation. Après avoir traité préalablement les textes en insérant après chaque mot du vecteur colonne un espace, nous avons créé une fonction donnant la fréquence d'un mot dans un texte puis une autre renvoyant la fréquence de chaque mot du texte. Pour cette dernière, il a été nécessaire d'utiliser la fonction UNIQUE afin d'avoir un vecteur colonne ne possédant qu'une seule occurrence de chaque mot. Puis on a initialisé le vecteur résultat en une colonne de zéros de longueur celle du vecteur résultant de la fonction UNIQUE et puis pour chaque élément de ce dernier, nous avons calculé la fréquence du mot correspondant. A partir de ce résultat, nous avons calculé la fréquence moyenne qui nous a aidées à créer une fonction renvoyant les mots dont la fréquence était supérieure à la moyenne et ce grâce à la fonction FIND.

Ensuite, nous avons remarqué que certains mots « insignifiants » comme les articles étaient renvoyés par la fonction ci-dessus. Nous avons donc créé une fonction enlevant les mots de deux lettres au plus puis une autre enlevant des mots appartenant à une liste d'éléments définie auparavant. Cette liste possédait essentiellement des articles, pronoms et déterminants. Nous avons préféré ne pas créer d'emblée une fonction ôtant les mots de trois lettres au plus car certains mots comme « vie » nous paraissaient importants et nous préférions donc qu'ils puissent paraître dans les mots significatifs si leur fréquence était élevée.

Puis, nous avons créé une fonction comparant deux textes et renvoyant ainsi seulement les mots significatifs communs à eux deux. Le résultat fut assez surprenant : après comparaison des quatre textes, les termes renvoyés ne nous paraissaient pas être d'une grande importance mis à part le mot « France ». Mais ce résultat était compréhensible. En effet, le thème de chaque texte varie au cours des années. Par exemple, en 2001, Jacques Chirac évoque l'explosion de l'usine AZF située à Toulouse et l'arrivée de l'euro. En 2002, le président de la République fait référence aux élections présidentielles de la même année ainsi qu'à la menace terroriste régnant dans le monde. En 2003, Jacques Chirac évoque les tensions et la guerre en Irak, les attentats au Proche-Orient ainsi que la canicule ayant touché la France. Enfin en 2004, le président de la République fait référence au tsunami (s'étant déroulé cinq jours auparavant) ainsi qu'au référendum sur la constitution européenne. Les thèmes varient donc d'une année sur l'autre, il est alors naturel de n'avoir qu'un seul mot « significatif » au terme de notre travail et ce mot est d'autant plus en accord avec le statut de président de la République puisqu'il s'agit du nom de notre pays.

II) Les difficultés rencontrées et améliorations possibles du projet

Nous aurions aimé affiner notre travail en créant un graphique. Deux choix se présentaient à nous :
-ou bien nous représentions la fréquence de chaque mot significatif en fonction de l'année du texte. Nous aurions ainsi obtenu l'évolution d'utilisation de chaque mot au cours du temps.
-ou bien nous représentions la fréquence d'un mot dans un premier texte en fonction de sa fréquence dans un second texte.

Malheureusement, nous n'avons pas réussi à programmer une fonction renvoyant le résultat espéré. Cependant, cette recherche nous a permis de connaître une autre fonction prédéfinie de SCILAB : VARARGIN qui correspondait à un nombre aléatoire d'arguments. Nous avons ainsi observé qu'il s'agissait d'une fonction très utile car elle permettait de généraliser la fonction qu'on tentait de programmer à un nombre quelconque d'arguments. De plus nous aurions souhaité poursuivre notre travail en effectuant le même genre de calculs en tenant compte de la proximité des mots de chacun des textes. Enfin, il aurait été intéressant de

travailler sur d'autres textes de partis politiques différents afin d'étudier leurs différences et similitudes. Peut-être emploient-ils chacun les mêmes mots malgré des idées opposées ou tout simplement différentes ? D'autre part, nous aurions pu comparer les vœux présidentiels diffusés le 31 janvier à la télévision et ceux prononcés devant un autre public, comme celui de la presse afin d'étudier le changement de son discours. Enfin nous pourrions dans un autre projet élargir le thème à la littérature ou à des articles de journaux.

En conclusion, nous pensons que SCILAB est un environnement agréable pour faire du calcul numérique (et ce en général, *id est* pas seulement pour notre projet). De plus, disposant de toute une batterie d'instructions graphiques et d'un langage de programmation assez simple mais puissant, il fut plaisant de l'utiliser pour mener à bien notre projet. Cependant, le seul point négatif qu'on peut lui trouver est la nécessité d'utiliser son éditeur intégré SCIPAD pour écrire des scripts ou nos propres fonctions et le fait que les résultats apparaissent non pas dans cette fenêtre mais dans celle de départ. Il faut donc alterner entre les deux fenêtres. Mais finalement, nous gardons de SCILAB sa simplicité et la richesse de ses fonctions prédéfinies. D'autre part, il fut intéressant de remarquer que grâce à SCILAB nous avons pu traiter différents thèmes allant de la linguistique à l'imagerie en passant par la cryptographie et la finance. Ainsi à l'aide d'un logiciel basé sur les mathématiques, nous avons pu nous intéresser à d'autres sujets par son intermédiaire.