Rapport de compréhension

LTAL : TP4

MANGNAN Valentin 21100233@etu.unicaen.fr
GOTTSTEIN Cyprien 21205784@etu.unicaen.fr

1 Introduction

L'objet de ce TP est le rapprochement de documents comparables. Ce procédé peut être utilisé en détection de plagiat ou de clone. Voici comment nous procédons :

1. On calcule les distances par compression entre chaque document du corpus.
2. On fabrique un *dendrogramme* en fusionnant deux à deux les documents, puis les groupes de documents (selon la distance *moyenne*, *minimale* et *maximale* entre les documents).
3. On sort une matrice (représentée par un tableau en HTML grâce aux groupes formés par le dendrogramme).
4. On colorie les valeurs (à l'aide de moyennes emboitées) du tableau pour rendre le résultat interprétable.

À travers nos tests nous chercherons à répondre aux questions suivantes :
- La relation de traduction conserve-t-elle les similitudes entre les documents ?
- Le choix d'une méthode entre *moy*, *min*, *max* est-il significatif ? Change-t-il les groupes formés ? Si oui, pourquoi ?

2 Observations sur les ressources constituées

Nous traiterons les corpus suivants :
- un corpus en langage naturel
  - 8 documents traduits dans 5 langues (en, fr, de, el, it) issus du corpus `europa_2008`
- un corpus de code
  - les sources du `tp1-2` concaténées
  - les sources du `tp3` concaténées

Suivent quelques observations sur le corpus constitué et nos résultats. D'une part, la taille de deux documents compressés influe leur distance. Et pour cause, `Zlib` (module de compression sous python) renvoie des résultats incorrects à partir de 32 ko. D'autre part, les documents du corpus `europa_2008` sont des pages HTML, allégées (le `head` n'est pas présent). Or c'est présisement dans ce qui a été enlevé qu'il se trouve le plus de similarité entre les documents. On peut alors considérer que toute distance comprise dans $\{0, 1\}$ est réellement significative.

3 Compte-rendu du résultat

3.1 Interprétation du tableau HTML

Suite aux exigences de l'énoncé nous avons colorié notre tableau. Les couleurs rouges indiquent des documents très proches, *a contrario* les couleurs vertes indiquent des documents très éloignés.

3.2 Influence de la méthode choisie

On remarque que le choix de la méthode n'est pas significatif pour le premier document. En effet, lors de la première itération du programme, on considère simplement la plus petite distance entre les documents. On observe une "cassure" (voir Figure 1 page 2) dans la courbe pour les corpus de code. On peut considérer que c'est dû à la présence d'un document *basique* plus ou moins exploité selon des groupes. Ainsi les quelques groupes très proches de la base se démarquent des autres qui s'en éloignent davantage.

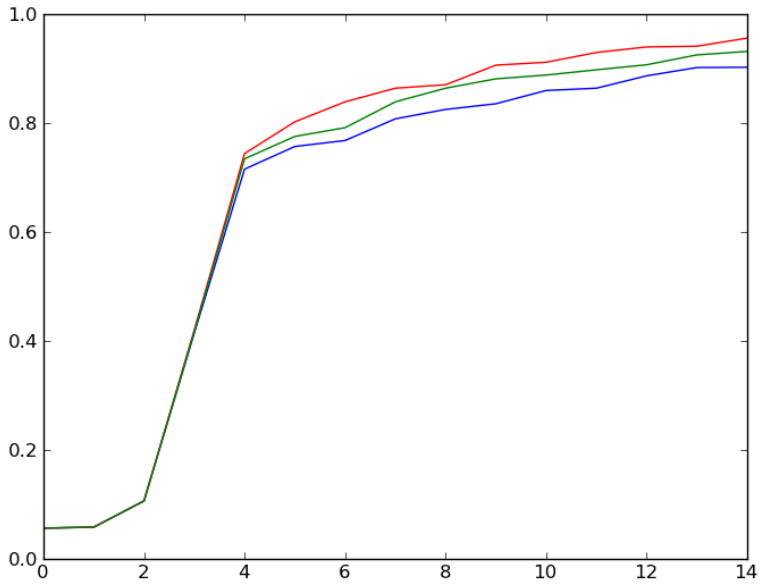

FIGURE 1 – Distance entre les groupes à chaque itération du programme (on voit une cassure à la 4e itération)

min (en bleu sur la Figure 1) : tendance à associer un élément avec un groupe (les Figure 1 des annexes **"Influence de la méthode sur la matrice des distances"** montrent un grand groupe auquel on a ajouté petit à petit des éléments seuls)

max (en rouge sur la Figure 1) : tendance à associer deux groupes ensemble (les Figure 2 ididem montrent plutôt des groupes indépendants réunis vers la fin)

moy (en vert sur la Figure 1) : tendance à associer un élément avec un groupe (les Figure 3 *ibidem* montrent une tendance similaire à la méthode *min*)

4 Documents en relation de traduction

Il est certes intéressant de rapprocher des documents d'une même langue. Néanmoins, il va de soit que certaines subtilités d'une langue peuvent altérer les résultats. Par exemple, deux textes peuvent parler d'un chose semblable, mais pour cela utiliser des mots différents. Comment nous affranchir de cette faiblesse ?

En composant notre corpus, nous avons choisi des documents en relation de traduction. En théorie, si a est très proche de c alors b sera très proche de d (avec a traduit par b, et c par d). Nos résultats vont-ils dans ce sens ? Considérez par exemple ce qui se passe pour la méthode *min* dans les 5 langues. (Annexe **"Comparaison des matrices de documents en relation de traduction"**). Vous remarquerez que les documents 1019, 1025 et 103 sont toujours dans le même groupe. Il est clair que la similarité est conservée dans ces 5 langues. De plus, la distance est aussi conservée. En effet, le document 100, qui est éloigné de tous les autres, l'est dans chacune des langues.

Qu'en conclure ? D'abord, la similarité et la distance sont conservées par la traduction. Il est donc possible de confirmer un rapprochement de documents en passant par la traduction. En outre, sachant cela, il devient possible d'évaluer si un document est la traduction d'un autre, et ce simplement en lisant les matrices des distances.

5 Synthèse des résultats

Le passage à la traduction conserve-t-il les similarités ? Oui, nous l'avons vu à travers quelques matrices, les documents conservent ces propriétés. On peut imaginer différentes applications de ce phénomène, comme l'alignement de documents avec leur traduction.

Qu'en est-il maintenant des différentes méthodes pour la construction du dendrogramme (*min*, *max*, *moy*) ? Y a-t-il une méthode meilleure que les autres ? Rappelons quelques spécificités de nos méthodes :

min : homogénéité au sein d'un groupe, rapprochement des points communs entre les documents

max : séparation des documents au profit de la distance la plus faible

moy : résultats proches du min

Les méthodes se valent. Selon l'utilisation qu'on en fera, il sera plus utile de choisir telle ou telle méthode. Par exemple, si l'on cherche à retrouver *une* base commune dans un corpus, le *min* sera déterminant. En revanche, si l'on cherche à grouper des documents par thèmes différents, le *max* sera préférable. Enfin, la *moy* n'est pas - à l'échelle de notre étude - particulièrement propice à un développement individuel.

Annexes : Influence de la méthode sur la matrice des distances (tp1-2)

Figure 1 – Matrices des distances (selon la méthode "min"). On observe principalement un gros groupe en haut à gauche, ainsi que quelques petits. Les groupes plutôt homogènes.

Figure 2 – Matrices des distances (selon la méthode "max"). On voit très nettement deux groupes. Ils sont séparés même s'ils ont des caractéristiques communes.

Figure 3 – Matrices des distances (selon la méthode "moy"). On est presque identique à la méthode "min".

Annexes : Influence de la méthode sur la matrice des distances (tp3)

Figure 1 – Matrices des distances (selon la méthode "min"). On observe principalement un groupe en haut à gauche, puis deux autres au milieu et en bas à droite. Les groupes sont bien homogènes.

Figure 2 – Matrices des distances (selon la méthode "max"). On voit que le groupe du haut gauche a aggloméré les petits groupes du bas (même si certains éléments communs sont très éloignés).

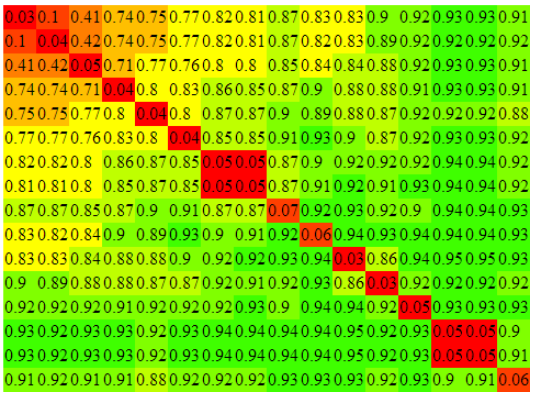

Figure 3 – Matrices des distances (selon la méthode "moy"). C'est plutôt un compromis entre les deux méthodes qu'on constate dans ce cas.

Annexes : Comparaison des matrices de documents en relation de traduction

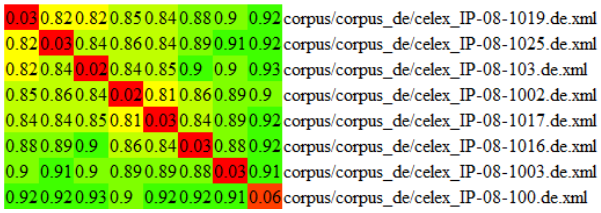

Figure 1 – Matrices des distances (selon la méthode "min") pour l'allemand.

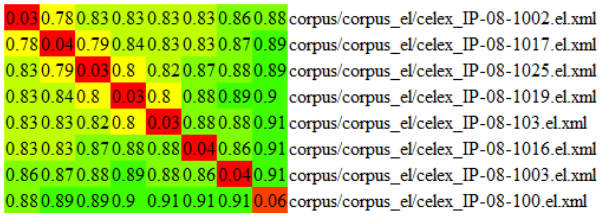

Figure 2 – Matrices des distances (selon la méthode "min") pour l'espagnol.

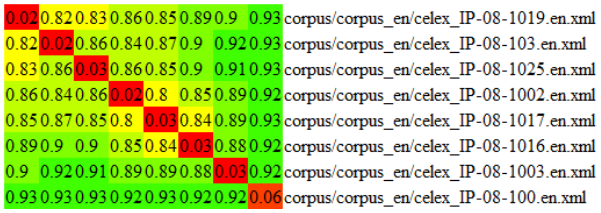

Figure 3 – Matrices des distances (selon la méthode "min") pour l'anglais.

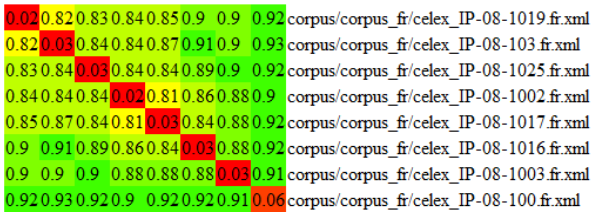

Figure 4 – Matrices des distances (selon la méthode "min") pour le français.

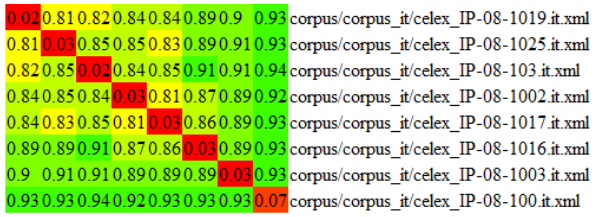

Figure 5 – Matrices des distances (selon la méthode "min") pour l'italien.