Enrichissements du corpus avec des analyseurs en dépendances syntaxiques

Mot d'introduction

La semaine dernière, nous avons demandé une sortie TSV plus détaillée afin de préparer le calcul de l'information mutuelle, une mesure de dépendance entre deux variables aléatoires.

Cette semaine, le but sera de calculer l'information mutuelle évoquée dans le TP précédent afin de reproduire les résultat des voisins de Le Monde. Il faudra également lire de la documentation pour préparer la visualisation des données.

Pour rappel

— un nouveau groupe gitlab vous a été attribué aléatoirement pour la semaine. Vous y trouverez un dépôt à cloner.
— la branche **main** sert à l'avancée du projet et des exercices, mais elle ne doit contenir que du code finalisé. il faut y pousser le moins souvent possible.
— une branche **doc** sert au rendu du journal de bord (un fichier markdown *différent* par semaine),
— chaque semaine, vous devrez créer des branches individuelles réservées au travail de chaque membre du groupe.
— un tag xxx-fin doit être utilisé pour indiquer qu'un exercice et terminé et un tag xxx-relu indiquera qu'il a été relu par un tiers et est prêt à être fusionné (**merge**). Un dernier tag indiquera que le travail sur la branche **main** est terminé.
— pour rappel, les **xxx** d'un tag seront à remplacer par **xy-sTrN**, où xy sont vos initiales, T le numéro de la séance et N votre rôle.

Exercice 1 Information mutuelle

L'information mutuelle mesure le degré de dépendance entre deux variables aléatoires. Il s'agit d'une mesure proche de l'entropie en théorie de l'information (voir le cours sur les arbres de décision d'introduction à la fouille de texte). Dans la section "à propos" des *voisins de le monde*[1], l'information mutuelle est utilisée dans le calcul de la proximité entre entre des prédicats ou des arguments. Tel qu'indiqué sur la page :

> « À chaque triplet extrait <recteur, relation, régi> correspond un couple <prédicat, argument> : le prédicat est constitué du recteur auquel on « accole » la relation, et l'argument est le régi. Pour chaque couple (cooccurrent) syntaxique, on calcule son information mutuelle. »

Le calcul de l'information mutuelle se fera donc sur la base de calculs de cooccurrences (en considérant une relation de voisinage direct dans un arbre en dépendances syntaxiques).

Plus précisément, nous calculerons la *pointwise mutual information* dont la formule est la suivante :

$$PMI(x, y) = log_2 \frac{p(x, y)}{p(x)p(y)} \tag{1}$$

Où :
— $p(x, y)$ est la probabilité jointe de x et y. Autrement dit, la probabilité que x et y soient observés simultanément ;
— $p(x)$ et $p(y)$ sont les probabilités de x et de y respectivement ;
— nous utiliserons comme référence pour obtenir une probabilité le nombre de tokens du corpus.

Supposons que le corpus du TP précédent était le suivant :
— je mange une pomme
— tu manges une pomme
— vous mangez des poires
— nous voyons les poires

Notre corpus fait donc 16 *tokens*. L'information mutuelle entre les lemmes "manger" et "pomme" sera donc :

1. ici : `http://redac.univ-tlse2.fr/voisinsdelemonde/infos/apropos.jsp`

$$\begin{aligned} PMI(manger\_obj, pomme) &= log_2 \frac{p(manger\_obj, pomme)}{p(manger\_obj)p(pomme)} \\ &= log_2 \frac{\frac{2}{16}}{\frac{3}{16} \times \frac{2}{16}} \\ &= log_2(5.33...) \approx 2.42 \end{aligned}$$

La sortie tabulaire du TP précédent devra être enrichie pour intégrer le calcul de l'information mutuelle :

prédicat			argument		mesures	
catégorie	lemme	relation	catégorie	lemme	fréquence	IM
V	manger	obj	N	pomme	2	2.42
V	manger	obj	N	poire	1	1.42

Ce travail est individuel. Chaque membre du groupe devra écrire sa propre version du code, une mise en commun sera effectuée à la fin pour merge avec la branche main.

Exercice 2 Mise en production

Comme pour les semaines précédentes, validez le code d'un(e) de vos camarade et ajoutez un tag **xxx-relu** quand le résultat est satisfaisant.

Finalement, fusionnez vos travaux et proposez une version finale combinant les différentes contributions sur la branche **main**.

Indiquez que le travail du groupe est terminé au moyen d'un tag.

Exercice 3 Mise à jour du journal de bord

Pour ce travail, chaque membre renseigne sa partie du journal de bord, qui sera hébergé sur la branche **doc**. Commentez :

1. vos difficultés
2. vos solutions
3. les choix lors des *merges*

Exercice 4 Préparation à la visualisation

Dans les prochaines séances nous allons effectuer de la visualisation graphique (dans les deux sens de graphique). Pour réaliser cela, nous utiliserons le format XML de GEPHI [2] appelé GEXF. [3] À ce titre, pour la prochaine séance, il vous est demandé de lire à minima le document *primer* de la spécification disponible à l'URL suivante : `https://gexf.net/primer.html`.

2. URL : `https://gephi.org`
3. URL : `https://gexf.net`