Projet : Classification de chansons à l'aide de motifs émergents

L'objectif de ce projet est d'utiliser la fouille de motifs séquentiels émergents pour construire des classifieurs permettant d'attribuer une chanson à son artiste interprète.

Constitution des corpus bruts : choix des chansons de 3 artistes

Dans ce projet, dans la phrase d'évaluation, on souhaite en fait retrouver les chansons d'un artiste principal parmi un ensemble de chansons de deux ou trois artistes.

Vous devez ainsi choisir un premier artiste (l'artiste principal) puis un deuxième artiste dont le style musical sera proche du premier et un troisième artiste dont le style musical sera plus éloigné du premier artiste. Un exemple consisterait à choisir les Beatles comme artiste principal, Oasis comme deuxième artiste et Metallica comme troisième artiste.

Les contraintes pour le choix des artistes sont les suivantes : vous devez disposer des paroles de quatre albums pour chaque artiste et les paroles doivent être dans la même langue pour tous les artistes.

(a) La première étape de cette phase consiste à récupérer les paroles de quatre albums pour chaque artiste sur des sites mettant à disposition des paroles.

(b) La deuxième étape consiste à créer les corpus d'apprentissage et de test pour chaque artiste, ainsi que des fichiers de références pour l'artiste principal, le tout de la façon suivante :
- le corpus d'apprentissage de chaque artiste correspond à un fichier dans lequel ont été fusionnés les fichiers des chansons de deux albums choisis parmi les quatre ainsi que la moitié des chansons d'un troisième album ; on obtient ainsi trois corpus d'apprentissage.
- le corpus de test de chaque artiste correspond à un répertoire contenant les chansons des albums restants (la moitié du troisième album et le quatrième album) ; on obtient ainsi un répertoire par artiste.
- le premier fichier de références contient la liste des chansons de test de l'artiste principal, le deuxième fichier de références contient la liste des chansons de test du troisième album de l'artiste principal et le troisième fichier de références contient la liste des chansons du quatrième album de l'artiste principal.

Apprentissage des classifieurs de chansons

L'objectif de cette phase est d'apprendre un classifieur à base de motifs qui prend, en entrée, une chanson et qui donne, en sortie, l'artiste interprète de cette chanson. Différents paramétrages pourront être utilisés pour construire les classifieurs.

(a) **Pré-traitement des corpus d'apprentissage**

La première étape de cette phase consiste à appliquer des pré-traitements sur les corpus d'apprentissage. Ces pré-traitements doivent tout d'abord intégrer un découpage des

corpus en séquences se terminant par un point (une séquence par ligne, en ajoutant éventuellement un point à chaque fin de ligne si les points étaient absents des corpus bruts). Les pré-traitements peuvent ensuite intégrer la suppression des mots outils, un étiquetage morpho-syntaxique... Pour cela, vous pouvez par exemple utiliser `nltk` et/ou `TreeTagger`.

À l'issue de cette étape, les corpus doivent être soit au format texte avec une séquence par ligne, soit au format `TreeTagger`.

(b) **Extraction des motifs séquentiels des corpus d'apprentissage**

La deuxième étape de cette phase consiste à extraire les motifs séquentiels de chaque corpus pré-traité. Pour cela, vous utiliserez le site `SDMC` [1], comme utilisé au premier TP. Vous pouvez alors faire varier différents paramètres lors de l'extraction des motifs, à savoir le support minimal, les longueurs minimale et maximale des motifs et les valeurs minimale et maximale du *gap*.

À l'issue de cette étape, vous obtiendrez trois fichiers de motifs extraits (un par corpus) contenant, sur chaque ligne, un motif suivi de son support dans le corpus.

(c) **Sélection des motifs émergents des corpus d'apprentissage**

La dernière étape de cette phase consiste à extraire les motifs émergents du corpus d'apprentissage de l'artiste principal par rapport aux corpus d'apprentissage des deux autres artistes pour obtenir 7 fichiers de motifs émergents à l'issue de cette étape :

1. les motifs émergents de l'artiste principal par rapport à ceux du deuxième artiste ;
2. les motifs émergents du deuxième artiste par rapport à ceux de l'artiste principal ;
3. les motifs émergents de l'artiste principal par rapport à ceux du troisième artiste ;
4. les motifs émergents du troisième artiste par rapport à ceux de l'artiste principal ;
5. les motifs émergents de l'artiste principal par rapport à ceux des deux autres artistes après fusion de leurs motifs ;
6. les motifs émergents du deuxième artiste par rapport à ceux des deux autres artistes après fusion de leurs motifs ;
7. les motifs émergents du troisième artiste par rapport à ceux des deux autres artistes après fusion de leurs motifs ;

Pour sélectionner les motifs émergents de chaque fichier, vous pouvez faire varier le seuil d'émergence utilisé (en considérant, par exemple, uniquement les émergents infinis) et/ou ne conserver qu'une partie des motifs émergents (pour en avoir à peu près autant dans chaque fichier de motifs émergents et pour en avoir un nombre raisonnable à chaque fois).

À l'issue de cette phase, on désigne par *classifieur*, l'ensemble des 2 ou 3 fichiers obtenus et qui seront utilisés pour réaliser la classification des chansons de test comme définie au point « Réalisation de la classification des chansons de test », ci-après (les fichiers de motifs émergents 1 et 2 constituent ainsi un premier classifieur, les fichiers 3 et 4 un deuxième classifieur et les fichiers 5, 6 et 7 un troisième classifieur).

1. `https://sdmc.greyc.fr`

Utilisation et évaluation des classifieurs de chansons

L'objectif de cette phase est de réaliser la classification des chansons de test et d'évaluer les résultats obtenus pour le premier artiste, selon les cas de test considérés et selon les paramétrages choisis pour la construction et l'utilisation des classifieurs de chansons.

(a) **Pré-traitement des chansons de test**

Pour réaliser le pré-traitement des chansons de test, on utilise de nouveau le script défini au point précédent « Pré-traitement des corpus d'apprentissage », en choisissant le paramètrage utilisé pour construire les 3 classifieurs considérés ici.

(b) **Extraction des motifs séquentiels des chansons de test**

Pour extraire les motifs séquentiels des chansons de test, on utilise de nouveau `SDMC`, en choisissant le paramètrage utilisé pour construire les 3 classifieurs considérés ici.

(c) **Réalisation de la classification des chansons de test**

Pour effectuer la classification d'une chanson de test (c'est-à-dire identifier son artiste interprète), on choisit l'artiste avec lequel la chanson partagent le plus de motifs communs (en comparant les motifs extrait des paroles de la chanson aux motifs émergents contenu dans chacun des fichiers du classifieur utilisé).

Pour effectuer la classification de l'ensemble des chansons de test, on distingue en fait trois cas de test, chacun utilisant un des trois classifieurs considérés ici :

1. on cherche à retrouver les chansons de l'artiste principal parmi les chansons de test de l'artiste principal, mélangées à celles du deuxième artiste, c'est-à-dire qu'on considère les fichiers de test de l'artiste principal et ceux du deuxième artiste mais sans fusionner les fichiers.
2. on cherche à retrouver les chansons de l'artiste principal parmi les chansons de test de l'artiste principal, mélangées à celles du troisième artiste, c'est-à-dire qu'on considère les fichiers de test de l'artiste principal et ceux du troisième artiste mais sans fusionner les fichiers.
3. on cherche à retrouver les chansons de l'artiste principal parmi les chansons de test de l'artiste principal, mélangées à celles des deux autres artistes, c'est-à-dire qu'on considère les fichiers de test de l'artiste principal et ceux des deux autres artiste mais sans fusionner les fichiers.

À l'issue de chaque cas de test, on doit obtenir un fichier contenant les chansons de test identifiées comme étant celles de l'artiste principal, parmi les fichiers de test considérés. (même format que celui des fichiers de références constitués dans la première phase).

(d) **Évaluation de la classification des chansons de test de l'artiste 1**

L'évaluation de la classification obtenu avec un ensemble de trois classifieurs s'effectue en calculant la précision, le rappel et la F-mesure sur l'ensemble des chansons de test de l'artiste principal, pour chacun des trois cas de test.

On calcule également le rappel en utilisant le fichier de références des chansons de test du troisième album puis le rappel en utilisant le fichier de références des chansons de test du quatrième album (à faire pour chacun des trois cas de test).

Travail à rendre

Vous aurez ainsi à comparer l'influence des pré-traitements ainsi que du choix des paramètres utilisés pour l'extraction des motifs, en construisant un ensemble de trois classifieurs pour chaque paramétrage choisi.

Vous rendrez un rapport contenant des statistiques sur les données choisies (découpées en corpus d'apprentissage et en corpus de test), en termes de nombre de séquences, taille des séquences, nombre de motifs extraits et nombre de motifs émergents sélectionnés (pour chacun des paramétrages évalués). Vous y présenterez également les résultats obtenus pour les 3 cas de test définis précédemment, en comparant différents classifieurs construits (chaque classifieur correspond à un paramétrage des pré-traitements et de l'extraction de motifs).

Ce rapport est à déposer sur Madoc pour le **19 décembre 2014**.