Master 2 ATAL

Fouille de textes et recherche d'informations

Classification de chansons à l'aide de motifs émergents

Étudiante :
Coraline Marie

Encadrant :
Solen Quiniou

12 janvier 2015

UNIVERSITÉ DE NANTES

Introduction

La fouille de données est une application au coeur de l'actualité. En effet, de plus en plus de données textuelles sont produites chaque jour, et sont le plus souvent non structurées et non ordonnées. C'est là tout l'intéret des applications de RI, qui permettent de sélectionner et d'extraire des informations pertinentes, dans des flux de données en éternelle extension.

Dans le cadre du cours *Fouille de textes et recherche d'informations*, dispensé au Master 2 ATAL de l'Université de Nantes, nous avions pour objectif d'utiliser des techniques de RI sur des corpus constitués de chansons. Pour cela, nous devions d'abord construire et prétraiter des corpus bien précis, puis coder un programme dont le but est d'essayer de déterminer à quel artiste appartient une chanson.

Dans ce rapport, nous détaillerons les différents choix et démarches adoptés pour ce projet. Nous étudierons donc d'abord les artistes et la construction des corpus. Puis nous verrons quels ont été les prétraitements et les méthodes utilisées pour permettre l'extraction de motifs émergents, avant de finir sur les résultats produit par le programme.

1 Construction des corpus

1.1 Choix des artistes

De nos jours, la musique est présente partout, dans toutes les langues et dans de nombreux styles musicaux différents. La première démarche de ce projet fût donc des choisir trois artistes, dont les oeuvres soient à la fois nombreuses, et dans la même langue.

Le premier artiste que j'ai choisi est Linkin Park. Ce groupe à déjà produit une quinzaine d'albums, tous en anglais, dont la plupart dans un style proche du hard rock, et du néo métal. J'ai donc retennu quatre albums : The Hunting Party (2014), Hybrid Theory (2000), Meteora (2003) et Linving Things (2012).

Le second artiste que j'ai sélectionné ressemble relativement à Linkin Park de part son style musical (hard rock et néo métal), son style de paroles et sa langue. Il s'agit du groupe Breaking Benjamin, auteur de sept albums, dont quatre uilisés pour ce projet : Dear Agony (2009), We Are Not Alone (2004), Phobia (2006) et Saturate (2002).

Le dernier artiste que j'ai choisi diffère complètement des deux précédents : Daft Punk. Bien que ce groupe français produise des oeuvres en Anglais, leur style musical est plus électronique et groovy. Pour cet artiste, sélectionné les quatre albums suivants : Random Access Memories (2013), Homework (1997), Discovery (2001) et Human after all (2005).

1.2 Création de corpus

Pour chaque artiste, j'ai créé un corpus d'apprentissage, et un corpus de test. Les corpus d'apprentissage serviront à l'entrainement des données, alors que les corpus de test seront utilisés pour essayer de trouver à quel artistes ils appartiennent.

1.2.1 Corpus d'entrainement

Pour ce projet, trois corpus d'entrainement ont été créés : un pour chaque artiste. Ces corpus sont contitués de deux albums et demi par artiste, ce qui signifie que deux albums entiers et la moitié d'un troisième ont été utilisés pour construire ces corpus. Les albums n'ont pas été sélectionné aléatoirement, car à chaque fois, il y a un vieil et un récent album dans le corpus d'apprentissage. Ce choix a été fait intentionnellement pour que l'impact de l'évolution de langage d'un artiste dans le temps, ait le moins d'influence possible sur les motifs.

Ainsi, le corpus d'apprentissage de Linkin Park contient les albums "The Hunting Party" (2014), "Hybrid Theory" (2000), et la moitié des chansons de "Meteora" (2003). Ceci représente au total 28 chansons, soit 1535 vers pour un peu moins de 10 000 mots.

Le corpus d'entrainement de Breaking Benjamin est composé des albums "Dear Agony" (2009), "We Are Not Alone" (2004) et de la moitié de "Phobia" (2006). Ces chansons sont au nombre de 28, ce qui fait 1204 vers, pour environ 6000 mots.

Pour finir, le corpus d'apprentissage de Daft Punk a été formé à partir des albums "Random Access Memories" (2013), "Homework" (1997) et la moitié de "Discovery" (2001). Ce corpus contient 21 chansons, c'est à dire 877 vers et environ 4800 mots.

1.2.2 Corpus de test

Les corpus de tests créés sont légèrement différents des corpus d'apprentissage. En effet, ceux-ci doivent être analysé par le programme qui doit déterminer ensuite à quel artiste appartient chaque chansons. Ainsi, il existe un corpus de test par artiste, et chacun contient un fichier par chanson.

Le corpus de test de Linkin Park contient la moitié des chansons de l'album "Meteora" (2003), et toutes les chansons de "Linving Things" (2012). Ce qui représente au total 14 chansons de 567 vers et environ 3700 mots.

En ce qui concerne le corpus de test de Breaking Benjamin, celui-ci contient la moitié de l'album "Phobia" (2006) et toutes les chansons de "Saturate" (2002). Ce qui donne 18 chansons de 659 vers et d'environ 3350 mots.

Le corpus de test de Daft Punk est quant à lui composé de la moitié des chansons de l'album "Discovery" (2001) et de toutes celles de "Human after all" (2005). On dénombre donc 9 chansons avec 343 vers et environ 2100 mots.

2 Extraction de motifs

2.1 Prétraitement des corpus

Après avoir fini de construire les corpus, nous devions extraires un certain nombre de motifs émergents. Cependant, les corpus sont extrèmement bruités, il a donc fallu au préalable les prétraiter. Pour cela, j'ai commencé par transformer toutes les majuscules en minuscules, et tous les élément de ponctuations comme les '!' et les '?' en '.'. En analysant les corpus, j'ai également remarqué que certaines phrases, sont composée de plusieurs vers, mais qu'aucune ponctuation n'y a été placée. J'ai donc rajouté des '.' aux endroits qu'il me paraissait être le mieux, pour donner un sens aux groupes de vers. J'ai également supprimé les smileys, et les précision de chants (chorus, single, x2, ...).

Dans un second temps, j'ai également voulu supprimer les stopwords, cependant je ne l'ai finalement pas fait pour deux raisons. Tout d'abord car cela diminue énormément la taille de mes corpus (plus de 80% des mots sont supprimés dans le corpus d'apprentissage de Linkin Park), et ensuite car je pense que les stopwords peuvent être représentatif du style d'un artiste. En effet, certains artistes emploient beaucoup plus de stopwords que d'autres, et cela peut être un moyen de différencier le style de parole lors de la reconnaissance des chansons.

2.2 Extraction de motifs automatique

Une fois que tous les corpus ont été prétraités, il faut en extraire des motifs. Pour cela, je me suis servie de Greyc sur le site SDMC [1]. Sur cet outils plusieurs paramêtres peuvent être modifié, pour l'extraction de motif automatique. Pour ma part, j'ai utilisé 6 combinaisons différentes de paramêtres :

Config	gap	taille	support
n° 1	[0 ;5]	10	10
n° 2	[0 ;0]	10	10
n° 3	[0 ;5]	10	20
n° 4	[0 ;0]	10	20
n° 5	[0 ;5]	10	5
n° 6	[0 ;0]	10	5
n° 7	[0 ;5]	10	3
n° 8	[0 ;0]	10	3

Table 1 – Paramêtres utilisés sur Greyc

J'ai également voulu faire varier le paramêtre concernant la taille maximale du motif. Cependant la majeur partie des mes motifs sont au maximum de taille 3, donc faire varier ce paramêtre n'influe quasiment pas sur le nombre de motif produit.

1. https ://sdmc.greyc.fr

Config	Linkin Park	Breaking Benjamin	Daft Punk
n° 1	633	284	239
n° 2	297	169	164
n° 3	197	102	97
n° 4	117	74	74
n° 5	1813	3777	452
n° 6	654	385	271
n° 7	10155	14257	1123
n° 8	1002	713	373

Table 2 – Nombre de motifs obtenus en fonction des paramêtres choisis

En ce qui concerne les corpus de tests, j'ai utilisé les mêmes paramêtres à l'exception du dernier, car dans une chanson d'en moyenne 30 vers, il est difficile d'obtenir des motifs qui apparaissent au moins 20 fois.

2.3 Extraction de motifs émergents

L'extraction de motifs automatique par l'intermédiaire de Greyc, m'a permis d'obtenir de nouveaux fichiers d'entrainement. Cependant, certains motifs sont communs à plusieurs de ces fichiers, il faut donc les supprimer. C'est pourquoi nous avons créé trois "classifieurs" différents. Le premier va permettre de comparer Linkin Park à Breaking Benjamin, le second permet de comparer Linkin Park à Daft Punk, et le dernier compare Linkin Park au deux autres artistes.

Ainsi, pour éviter au mieux au programme toutes confusions entre les artistes, nous avons extraits les motifs émergents de chaque artiste, en en fonction du classifieur où il se trouve. Pour ce faire, j'ai codé un script qui compare tous les motifs des artistes et qui ne conserve que les motifs émergents.

3 Résultats

L'évaluation se fait par l'intermédiaire d'un script qui évalue la précision, le rappel et la f-mesure pour chaque classifieur. Bien que les résultats ne soient pas très satisfaisant, je vais les détailler ci-après.

Conclusion

Config	Précision	Rappel	F-mesure
n° 1	633	284	239
n° 2	297	169	164
n° 3	197	102	97
n° 4	117	74	74
n° 5	1813	3777	452
n° 6	654	385	271
n° 7	10155	14257	1123
n° 8	1002	713	373

Table 3 – Résultat obtenus pour le premier classifieur en fonction de la configuration des paramêtres