Fouille de textes et recherche d'information

Who Wrote This Song ?

Marie-Charlotte Daureu

December 14, 2014

Master 2 ATAL

Contents

1 Introduction

L'objectif de ce projet est de pouvoir reconnaître l'artiste qui a écrit une chanson, à partir de classifieurs. Ces classifieurs sont créés à partir de techniques de fouille de motifs séquentiels émergents.

Trois artistes sont séléctionnés. Ici, il s'agit d'Avenged Sevenfold, de Three Days Grace, et d'Alestorm, des artistes composant tous leurs textes en anglais. Les deux premiers ont des thématiques relativement similaires, parlant de la vie de tous les jours et des émotions humaines, tandis que le troisième est très différent car il s'agit d'un groupe de métal pirate, utilisant un vocabulaire très spécifique.

Ce rapport présentera premièrement la création des corpus d'entraînements et de tests, avec les pré-traitements réalisés. Ensuite, il expliquera la procédure utilisée pour créer les classifieurs, enfin, il évaluera les performances de ces derniers selon différents paramètres pris en compte.

2 Constitution et pré-traitements des corpus

Afin que l'évaluation soit cohérente, les corpus d'entraînements et de tests sont créés et pré-traités de façon similaire. Pour chaque artiste, les paroles de quatre albums sont récupérées depuis internet.

Les corpus d'entraînements sont constitués de la fusion des deux premiers albums pour chaque artiste ainsi que de la première moitié du troisième. Les chansons restantes sont concervées séparément dans un dossier, et consituent ainsi le corpus de test.

Pour utiliser ces corpus, des pré-traitements sont nécessaires. Un script python a été réalisé pour cela.

Premièrement, le bruit inséré dans les paroles de chansons tels que les symboles `[*2]` indiquant une répétition de la phrase ont été remplacés par la répétition de la phrase en question. De même pour les informations de type `Chorus` qui ont été remplacées par le texte des refrains concernés. Toutes les phrases ont été converties en minuscules.

Ensuite, il a été décidé ici de supprimer les mots outils pour l'extraction des motifs émergents. Ils sont nombreux dans les chansons mais pas forcément porteurs de sens ni d'information caractéristique. Il existe dans les corpus bruts extraits d'internet, de nombreuses abréviations anglaises communes. Il n'est pas pertinent de les concerver sous leur forme abrégée, car il s'agit des "fans" qui écrivent les chansons de la sorte sur internet, et non pas une façon spécifique d'écriture d'un groupe de musique. Il s'agit le plus souvent de mots outils, or, nous ne pouvons les reconnaitre de façon automatique avec la bibliothèque nltk sous leur formes abrégées. Les abréviations ont donc été remplacées par leur mot complet avant la suppression automatique des mots outils.

```
Exemple : 're -> are
```

Finalement, les corpus ont été formatés de façon à ne garder qu'une phrase par ligne, en ajoutant un point à la fin de chaque ligne. En effet sur le site où les textes ont été trouvés, il n'y avait pas de ponctuation. Ces corpus pourront alors enfin être traités par la suite pour les expérimentations.

3 Création des classifieurs

3.1 Extraction des motifs séquentiels

Pour chaque corpus maintenant pré-traité, le site SDMC est utilisé afin d'extraire les motifs séquentiels. Cela est réalisé pour les corpus d'entraînements et de tests avec les mêmes paramètres.

Un motif de taille trop grande serait inutile, étant donné nos données. En effet les phrases des chansons, pour les trois artistes, sont souvent assez courtes, surtout après suppression des mots outils. Vu que la taille du motif n'est pas très très grande, il semble peu utile de choisir un gap trop grand. Un support de 1 serait peu approprié, mais un support trop grand n'est pas possible. En effet, comme les corpus sont de petite taille, trop peu de motifs seraient retournés.

Trois paramétrages seront ici testés, et détaillés dans la partie de ce rapport consacrée aux expérimentations.

3.2 Selection des motifs émergents

Un artiste principal est choisit : ici, il s'agit de Three Days Grace. Un second artiste, le similaire, est Avenged Sevenfold. Le troisième artiste est ici Alestorm, le plus différent. Une expérience réalisée par la suite inversera le premier et le second artiste, afin de voir l'impact sur les résultats.

Cette étape est réalisée uniquement pour les corpus de tests. Il s'agit de retrouver les motifs émergents, en comparant les artistes les uns aux autres selon les méthodes suivantes :

Le premier artiste est comparé au second, et vice versa. Cela créé deux fichiers d'émergences, pour la création d'un premier classifieur.

Le premier artiste est comparé au troisième, et vice versa. Cela permettra la création d'un second classifieur.

Le premier artiste est comparé aux deux autres après leur fusion, puis le deuxième est comparé aux autres, puis le troisième. Cela permettra la création du troisième classifieur. Il faut faire attention à correctement fusionner les motifs des deux artistes dans ces cas là. Un simple copier coller ne suffit pas dans le cas où un même motif est présent dans deux fichiers d'artistes différents.

Afin de créer ces fichiers d'émergences, qui constitueront les classifieurs, un script python a été réalisé. Ce script prend en paramètre le seuil d'émergence choisi. Par convention ici, si ce seuil est à 1000, cela équivaut à un seuil d'émergence infini. Le script prend deux fichiers en paramètres, ceux des fichiers de motifs séquentiels des artistes à comparer.

Si le seuil d'émergence choisi est infini, alors seuls les motifs apparaissant avec le premier artiste et pas le second seront extraits. Si le seuil d'émergence n'est pas infini, alors le taux de croissance pour chaque motif entre les deux artistes est calculé. Si il est supérieur au paramètre choisi, on garde le motif, sinon, non.

Les fichiers constituant les classifieurs sont donc des fichiers contenant les motifs émergents des artistes par rapport aux autres. Plusieurs paramétrages pourront être alors facilement testés grâce à ce script, en faisant varier le seuil d'émergence.

4 Utilisation des classifieurs

Afin de pouvoir utiliser ces classifieurs sur nos données, nous créons trois scripts pythons, un pour tester chaqun des classifieurs indépendament.

Pour le premier classifieur, le script prend en paramètres les deux fichiers constituant le classifieur. Il prend également en paramètre le dossier contenant les chansons de test pour le cas de test 1, c'est à dire dans le cas ou les chansons du premier et du second artiste ont étées mélangées. Il va écrire, pour chaque chanson du dossier, l'artiste correspondant. Pour cela, les motifs séquentiels extraits de la chanson de test vont être comparés aux motifs émergents des fichiers classifieurs. Si il y a plus de similitudes avec l'artiste 1, alors l'artiste 1 va être écrit après le nom de la chanson. Inversement s'il y a plus de similitudes avec l'artiste 2. Si il y a égalité, alors aucune décision ne sera prise. Cela fera baisser le rappel pour le nombre de chansons bien reconnues, mais cela évitera de faire baisser la précision des résultats.

5 Expérimentations et Résultats

Différentes expériences ont été ici réalisées, comme mentionné plus haut. Nous allons les récapituler et les nommer. Des statistiques sont réalisées pour chaque expérience, et le rappel, la précision et la F-mesure sont calculés.

5.1 Rappels et Statistiques globales sur les corpus

Mesures utilisées pour comparer les résultats :

Rappel (R): Chansons correctement attribués à l'artiste 1 sur les Chansons appartenant à l'artiste 1

Précision (P): Chansons correctement attribués à l'artiste 1 sur les Chansons attribués à l'artiste 1

F-mesure : 2PR/P+R

A noter que dans les tableaux de statistiques, le nombre de mots par séquence est une moyenne. Ces statistiques sont calculées sur les corpus nettoyés, ce qui signifie que les mots qui ne sont pas des mots outils seulement sont comptés.

Statistiques globales sur les corpus d'entraînements nettoyés :

Artiste	Nb de séquences	Nb de mots/séquences
Three Days Grace	1375	2
Avenged Sevenfold	1119	3
Alestorm	851	3

Dans ces corpus d'entraînements, l'artiste Three Days Grace possède un plus grand nombre de séquences, mais avec moins de mots par séquence. En effet, les phrases des chansons de ce groupe sont plus courtes en moyenne que celles de Avenged Sevenfold.

Pour chaque ensemble de test, on a 15 chansons. 5 chansons sont extraites de l'album 3, et 10 chansons de l'album 4.

Statistiques globales sur les corpus de tests nettoyés, par artiste :

Artiste	Nb de séquences	Nb de mots/séquences
Three Days Grace	788	2
Avenged Sevenfold	869	2
Alestorm	725	3

Les trois cas de tests présentés sont :
1) pour les chansons de tests de l'artiste 1 et 2 mélangées
2) pour les chansons de tests de l'artiste 1 et 3 mélangées
3) pour les chansons de tests de l'artiste 1 et 2 et 3 mélangées

5.2 Expériences : Statistiques et Résultats

Les abréviations correspondent aux artistes suivants : TDG pour Three Days Grace, A7X pour Avenged Sevenfold.

Pour les trois premières expériences, Three Days Grace est considéré en tant qu'artiste 1.

5.2.1 Première Expérience

```
Sup>2, Gap[0:2], 2<=taille<=5, lemme uniquement.
```

Le tableau ci dessous représente le nombre des motifs extraits pour cette expérience pour chaque artiste ou groupement d'artiste, en vue de réaliser les différents cas de tests.

TDG	A7X	Alestorm	TDG et A7X	TDG et Alestorm	A7X et Alestorm
300	502	283	786	582	783

Ces résultats montrent que pour l'artiste Three Days Grace, moins de motifs ont été extraits du corpus d'entraînement. Cet artiste possédait pourtant le plus grand nombre de séquences, mais comme les séquences étaient plus courtes, cela réduit sûrement le nombre de motifs différents extraits. De plus, en regardant les données, le vocabulaire utilisé par cet artiste semble plus limité que le vocabulaire de Avenged Sevenfold. Cela explique le plus petit nombre de motifs extraits.

Le tableau ci dessous représente les statistiques pour les motifs émergents, en nombre de motifs, selon le taux d'émergence choisi. La notation "1 sur 2" signifie qu'il s'agit des motifs émergents de l'artiste 1 par rapport à l'artiste 2. Cette notation sera la même pour les expériences suivantes.

Taux	1 sur 2	2 sur 1	1 sur 3	3 sur 1	1 sur 2 et 3	2 sur 1 et 3	3 sur 1 et 2
1	284	486	299	282	284	485	281
Infini	284	486	299	282	284	485	281

Que le taux d'émergence soit à 1 ou à l'infini, les motifs émergents extraits sont les mêmes. Les résultats seront donc les mêmes pour les deux cas. Il faut

noter également que le nombre de motifs émergents de l'artiste 2 comparé à l'artiste 1 est nettement supérieur à l'inverse. Il y a en effet à la base un plus grand nombre de motifs extraits pour l'artiste Avenged Sevenfold, ce qui peut en partie expliquer ces résultats. Ce tableau montre que ces motifs supplémentaires ne sont pas communs avec ceux de Three Days Grace.

Résultats:

Méthode	Précision	Rappel	F-mesure	Rappel 3ème album	Rappel 4ème album
Class1	66.67%	26.67%	38.10%	40.00%	20.00%
Class2	83.33%	33.33%	47.61%	40.00%	30.00%
Class3	66.67%	26.67%	38.10%	40.00%	20.00%

Suite à ces résultats, il est visible que le premier artiste est plus difficile à différencier du second que du troisième. Cela est évident puisqu'ils se ressemblent plus. A remarquer également que les chansons du troisième album sont mieux retrouvées que celles du quatrième, que ce soit pour le cas de test 1 ou 2. Bien que la précision soit au dessus de 50% pour chaque cas de test, le rappel lui est très faible.

Il y a beaucoup de décisions qui ne peuvent pas être prises par le classifieur. Cela ne peut pas être dû au fait que les motifs de la chanson sont présents pour différents artistes, puisque les émergents infinis sont traîtés ici. Cela veut donc dire que les motifs extraits de chansons ne sont pas présents dans les fichiers d'émergents, ou bien qu'ils sont présents mais qu'il y en a a peu près autant pour un artiste que pour un autre.

Une explication possible à ce fait serait que le nombre de motifs extraits n'est pas assez important pour chaque artiste dans ce cas de figure. D'où la seconde expérience.

5.2.2 Seconde Expérience

```
Sup>2, Gap[0:2], 1<=taille<=5, lemme uniquement.
```

Le tableau ci dessous représente le nombre des motifs extraits pour cette expérience pour chaque artiste ou groupement d'artiste, en vue de réaliser les différents cas de tests.

TDG	A7X	Alestorm	TDG et A7X	TDG et Alestorm	A7X et Alestorm
500	845	619	1182	1021	1325

Les observations sont les mêmes que pour l'expérience précédente.

Le tableau ci dessous représente les statistiques pour les motifs émergents, en nombre de motifs, selon le taux d'émergence choisi.

Taux	1 sur 2	2 sur 1	1 sur 3	3 sur 1	1 sur 2 et 3	2 sur 1 et 3	3 sur 1 et 2
1	337	682	402	521	326	630	469
Infini	337	682	402	521	326	630	469
Précédentes	284	486	299	282	284	485	281

Comme pour l'expérience précédente, ici le taux d'émergence n'influe pas sur les motifs émergents extraits. Cependant, autoriser les motifs de taille 1 augmente le nombre de motifs extraits pour tous les artistes, ce qui était attendu. Il faut noter tout de même que sur les 500 motifs extraits de Three Days Grace, les motifs émergents entre ce dernier et Avenged Sevenfold ne sont qu'au nombre de 337. Il y a des motifs communs entre ces deux artistes, ce qui fait baisser le score de motifs émergents.

Résultats :

Méthode	Précision	Rappel	F-mesure	Rappel 3ème album	Rappel 4ème album
Class1	50.00%	20.00%	28.57%	60.00%	0.00%
Class2	88.89%	53.33%	66.67%	100.00%	30.00%
Class3	42.86%	20.00%	27.27%	60.00%	0.00%

Ces résultats se montrent très bons pour la comparaison entre l'artiste 1 et l'artiste 3, qui utilise un vocabulaire très particulier, celui de la piraterie. Le terme "sails" qui signifie "voiles", n'apparaitra jamais dans les textes des autres artistes et deviendra alors très discriminant. Avec la concervation des motifs de taille 1, ce vocabulaire spécifique permet une bonne distinction entre les textes des artistes. Cependant nous voyons que globalement, la comparaison des trois artistes est moins bonne que dans l'expérimentation précédente. Les motifs de taille 1 ajoutent de la difficulté pour différencier l'artiste 1 de l'artiste 2.

Le vocabulaire des artistes 1 et 2 semble assez similaire d'après ces expériences. Une fois de plus cependant, l'album 3 de Three Days Grace est mieux reconnu que l'album 4. Ses textes doivent donc être un peu plus originaux. L'album 4 semble très similaire aux chansons d'Avenged Sevenfold ce qui rend difficile la classification.

La troisième expérience va tenter de mieux différencier l'artiste 1 et l'artiste 2 en ajoutant les informations sur les catégories syntaxiques des mots lors de l'extraction des motifs séquentiels.

5.2.3 Troisième Expérience

`Sup>2, Gap[0:2], 1<=taille<=5, lemme et catégorie syntaxique.`

Le tableau ci dessous représente le nombre des motifs extraits pour cette expérience pour chaque artiste ou groupement d'artiste, en vue de réaliser les différents cas de tests.

TDG	A7X	Alestorm	TDG et A7X	TDG et Alestorm	A7X et Alestorm
1187	3079	2311	3783	3300	4561

En ajoutant la catégorie syntaxique, le nombre de motifs extraits de chaque corpus d'entraînement devient très grand.

Le tableau ci dessous représente les statistiques pour les motifs émergents, en nombre de motifs, selon le taux d'émergence choisi.

Taux	1 sur 2	2 sur 1	1 sur 3	3 sur 1	1 sur 2 et 3	2 sur 1 et 3	3 sur 1 et 2
1	732	2624	922	2046	695	2233	1655
Infini	732	2624	922	2046	695	2233	1655
Précédentes	337	682	402	521	326	630	469

Une fois de plus le taux d'émergence ne fait pas varier les fichiers. Aussi, même si le nombre de motifs émergents reste grand, il est tout de même globalement plus faible que le nombre de motifs de base de chaque artiste. Il y a beaucoup de recoupements entre les motifs des différents artistes.

Résultats :

Méthode	Précision	Rappel	F-mesure	Rappel 3ème album	Rappel 4ème album
Class1	0.00%	0.00%	0.00%	0.00%	0.00%
Class2	100.00%	33.33%	50.00%	40.00%	30.00%
Class3	100%	13.33%	19.99%	20.00%	10.00%

Les résultats ne sont pas du tout bons ici. La précision est certes à 100% pour le cas de test 2 et 3 mais le rappel est vraiment trop mauvais. Pour le cas de test 1, aucune chanson n'a été retournée pour l'artiste 1. L'ajout des catégories syntaxiques, bien qu'ajoutant un grand nombre de motifs émergents pour les artistes, ne permet pas une meilleure classification des textes des chansons. Le classifieur trouve qu'il y a autant de similitude pour les chansons de tests entre un artiste qu'un autre, et n'arrive pas à prendre une décision afin de retourner les chansons appartenant à l'artiste 1. Les tournures de phrases étant souvent étranges dans les chansons, pour un choix de rimes et de rythme, cela ne semble donc pas être un bon indicateur ici. Il y a beaucoup trop de motifs séquentiels différents, et rien de très caractéristique à un artiste.

Cependant en regardant plus en détails les fichiers de résultats, l'artiste 2 semble plus facilement reconnu par le système. Cela est pareil pour les deux premières expériences. Il semblerait donc qu'il possède un vocabulaire et des tournures de phrases plus originales. Cela peut être indiqué également par le tableau de motifs émergents. En effet, depuis le début des expériences, Avenged Sevenfold possède un plus grand nombre de motifs extraits et un plus grand nombre de motifs émergents en comparaison avec Three Days Grace. Cela pourait aider le classifieur à prendre plus facilement une décision.

5.2.4 Expérience 2 Bis

Après cette série de trois expériences, il a été noté que l'artiste actuellement considéré en tant que artiste 2, était apparement plus souvent reconnu par les classifieurs que l'artiste 1. Afin de tester cette hypothèse et de confirmer nos idées, l'artiste 1 et l'artiste 2 ont été inversés dans une expérience. Les paramètres sont identiques à la deuxième expérience réalisée, afin de pouvoir comparer.

Le tableau ci dessous représente le nombre des motifs extraits pour cette expérience pour chaque artiste ou groupement d'artiste, en vue de réaliser les différents cas de tests.

TDG	A7X	Alestorm	TDG et A7X	TDG et Alestorm	A7X et Alestorm
500	845	619	1182	1021	1325

Le tableau ci dessous représente les statistiques pour les motifs émergents, en nombre de motifs, selon le taux d'émergence choisi.

Taux	1 sur 2	2 sur 1	1 sur 3	3 sur 1	1 sur 2 et 3	2 sur 1 et 3	3 sur 1 et 2
1	682	337	706	408	630	326	469
Infini	682	337	706	408	630	326	469

Comme précédement, l'artiste Avenged Sevenfold possède un plus grand nombre de motifs émergents que Three Days Grace.

Résultats :

Méthode	Précision	Rappel	F-mesure	Rappel 3ème album	Rappel 4ème album
Class1	69.23%	60.00%	64.28%	80.00%	50.00%
Class2	92.86%	86.67%	89.66%	100.00%	80.00%
Class3	60.00%	40.00%	48.00%	60.00%	30.00%
Class1 (ancien)	50.00%	20.00%	28.57%	60.00%	0.00%
Class2 (ancien)	88.89%	53.33%	66.67%	100.00%	30.00%
Class3 (ancien)	42.86%	20.00%	27.27%	60.00%	0.00%

Ces résultats confirment l'impression dégagée par les résultats des expériences précédentes : l'artiste 2 est beaucoup mieux différencié par nos classifieurs que l'artiste 1. Un plus grand nombre de motifs émergents et un vocabulaire plus spécifique semblerait être la raison à ces résultats.

6 Conclusion

Ce projet présente une réalisation simple de classifieurs automatique de textes de chansons pour trois artistes, à l'aide d'extraction de motifs séquentiels. Globalement, le système est performant pour classifier les artistes en comparaison avec Alestorm, qui a été volontairement choisi très différent des deux autres artistes. Dans la meilleure configuration, le système reconnait Avenged Sevenfold en comparaison avec Alestorm avec un taux de F-mesure de 89.66%. Cependant, lorsque les deux artistes Three Days Grace et Avenged Sevenfold sont comparés, les résultats sont bien moins bons. Les motifs extraits des chansons de tests sont pour certains déjà vus dans le corpus d'entraînement d'Avenged Sevenfold tandis que d'autres ont été vus dans le corpus d'entraînement de Three Days Grace ce qui rend souvent le choix impossible pour le classifieur.

Différentes expérimentations ont montré que concerver les motifs émergents de taille 1 permet de mieux classifier les artistes possédant un vocabulaire très spécifique. Cependant, cela peut ajouter de la difficulté à différencier deux artistes ayant du vocabulaire commun, comme ici Three Days Grace et Avenged Sevenfold.

L'utilisation de l'information sur les catégories syntaxiques ne semble pas pertinente pour le traitement de corpus de chansons. En effet, les motifs extraits sont aussi bien trouvés pour un artiste que pour un autre, et les formes varient grandement. Cela est dû au fait que dans des chansons, les artistes ne se tiennent pas à une structure de phrase très spécifique, car l'interêt est plutôt porté sur le rythme ou sur les rimes.

Enfin, ces expériences ont également montré des différences entre les albums d'un même artiste. En effet, les résultats sont meilleurs sur toutes les expériences pour l'album numéro 3 de Three Days Grace, en comparaison avec l'album numéro 4. Ce dernier se trouve en effet plus similaire aux albums d'Avenged Sevenfold ce qui créé de nombreuses erreurs de classifications.

Pour aller plus loin, il aurait été utile de tester l'hypothèse sur les mots outils, en tentant de refaire les expériences sans les supprimer. En effet, peut être que certains artistes utilisent plus souvent le mot "I" tandis que d'autres parlent peut être plus au pluriel avec "we". Aussi, l'utilisation de l'information sur le nombre de mots moyen par phrase aurait pû être utilisé pour trancher en cas de doute entre deux artistes. Le but de ce projet était surtout l'utilisation des motifs séquentiels afin de créer les classifieurs, voilà pourquoi l'ensemble des expériences ont été réalisées plutôt en ce sens ici, mais ces idées pourraient être retenues pour une amélioration future de ces classifieurs.