Análisis de letras de canciones en función de su país de origen

Minería de datos y el paradigma Big Data
Curso 2018–2019

Autor
Alejandro Díaz Román
Jaime Del Rey García
Paula Munoz Lago

Facultad de Informática
Universidad Complutense de Madrid

Índice

Capítulo 1

Introducción

"Frase célebre dicha por alguien inteligente"
— Autor

El auge de la producción musical en los últimos años ha propiciado un aumento en el volumen de los datos, de los cuales se puede extraer el contexto social y cultural del país de procedencia del autor.

Capítulo 2

Objetivos

En esta práctica, nuestro objetivo es extraer, además de las palabras más utilizadas de cada autor, y de cada país, es extraer el sentimiento que prima en dicho país. Partimos de la hipótesis de que la temática principal de las letras de las canciones españolas es el amor, y lo comprobaremos como parte de esta práctica. El alcance de nuestro trabajo lo hemos establecido en función al tiempo del que disponemos, por lo que no pretendemos cubrir la totalidad de los países, sino los más significativos o de los que dispongamos la información suficiente.

Capítulo 3

Metodología

Para realizar la investigación hemos barajado las distintas metodologías y hemos seleccionado KDD (knowledge discovery in databases). Este modelo está directamente relacionado con la extracción de información y conocimiento de los datos disponibles. Sigue un proceso iterativo que, explorando los datos, extrae las distintas relaciones. http://www.webmining.cl/2011/01/proceso-de-extraccion-de-conocimiento/

Capítulo 4

Implementación

4.1. Desarrollo de la aplicación

Dentro del campo de procesamiento de textos, el análisis de letras de canciones no es el más popular, sin embargo existen algunos trabajos que nos han servido de guía para establecer los límites y punto de comienzo de nuestro proyecto. Sin duda el proyecto existente que más se ajusta a nuestra primera idea es *Data, data*, que hace un estudio exahustivo de las canciones del cantante Uruguayo Jorge Drexler, tanto de la letra como de la música.

Como hemos comentado en el Capítulo 2, nuestros objetivos son obtener los sentimientos que predominan en diferentes países, y comprobar si la hipótesis que hemos establecido indicando que el tema dominante en España es el amor.

4.2. Creación del conjunto de datos

Nuestro punto de partida son los datos que encontramos en varios datasets de Kaggle [1] [2]. Cuyo formato, en el primer caso, es el que se aprecia en la Figura 4.1. Se componede 4 columnas, la primera contiene el nombre del grupo, la segunda el título de la canción, seguida en la siguiente columna de un link que unido a http://www.lyricsfreak.com/ nos redirige a una página donde encontramos más información de la canción, como el año de publicación entre otros, y por último la letra. En el caso del segundo dataset, el formato es el que se aprecia en la Figura 4.2. En él encontramos el nombre del grupo, seguido del año de publicación de la canción, cuyo nombre aparece en la tercera columna, y finalmente la letra.

Un dato relevante en el desarrollo de ésta práctica del cual no disponemos a través de Kaggle es el país de procedencia de cada autor(a). Para obtenerlo

[1] `https://www.kaggle.com/mousehead/songlyrics`
[2] `https://www.kaggle.com/gyani95/380000-lyrics-from-metrolyrics`

ABBA	Ahe's My Kind Of Girl	/a/abba/ahes+my+kind+of+girl_20598417.html	Look at her face, it's a wonderful face And it means something special to me Look at the way that she smiles when she sees me How lucky can one fellow be? She's just my kind of girl, she ma...

Figura 4.1

ABBA	Ahe's My Kind Of Girl	/a/abba/ahes+my+kind+of+girl_20598417.html	Look at her face, it's a wonderful face And it means something special to me Look at the way that she smiles when she sees me How lucky can one fellow be? She's just my kind of girl, she ma...

Figura 4.2

utilizaremos técnicas de minería web para extraer dicha información a través de Wikipedia u otras fuentes. Una vez obtenida la información (país, autor, letras de sus canciones), limpiaremos los datos escogiendo sólo las palabras más relevantes, siendo éstas los verbos, sustantivos y adjetivos, eliminando artículos, conjunciones.

4.3. Limpieza y preprocesamiento de los datos

La limpiza de las letras es un paso crucial en el desarrollo de éste proyecto, puesto que muchas letras están compuestas de palabras que no aportan información relevante a la hora de realizar el análisis, como son los artículos, conjunciones, etc. Por ello, una vez obtenidos todos los datos, el paso siguiente será eliminar las llamadas *stopwords*, tanto del inglés como del castellano. Éste conjunto de palabras está compuesto por los tipos comentados anteriormente, es decir, palabras que no aportan información al análisis. Para ello, hemos desarrollado una función en R que, dado un texto, extrae únicamente las palabras que se encuentran en el diccionario, tanto en el de inglés como en el de castellano.

Como parte del preprocesamiento de los datos y con el fin de obtener una descripción más visual del set de datos obtenido, hemos realizado algunas pruebas. En primer lugar, hemos obtenido las palabras que aparecen en las canciones más conocidas del grupo Queen.

Finalmente, tras estudiar los gráficos obtenidos y ver que tiene sentido que las palabras más utilizadas en dichas canciones son las que se muestran, procedemos a generalizar un poco más nuestra exploración, para ver cuales son las palabras más utilizadas por el grupo británico. Como se muestra en la figura 4.8, la palabra *amor* aparece más de 400 veces en las canciones de

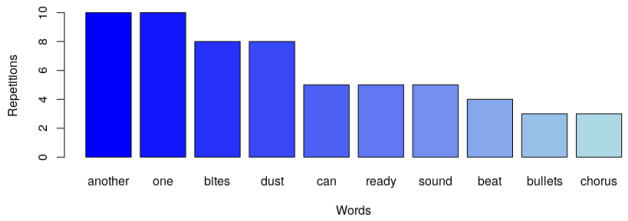

Figura 4.3: *Another One Bites the Dust* - Queen

las que disponemos de dicha banda.

4.4. Transformación de los datos

4.5. Elección de la tarea de minado adecuada

4.6. Eleección del algoritmo de minería de datos

4.7. Aplicación del algoritmo elegido

4.8. Interpretación de los resultados

4.9. Aplicación de los resultados obtenidos

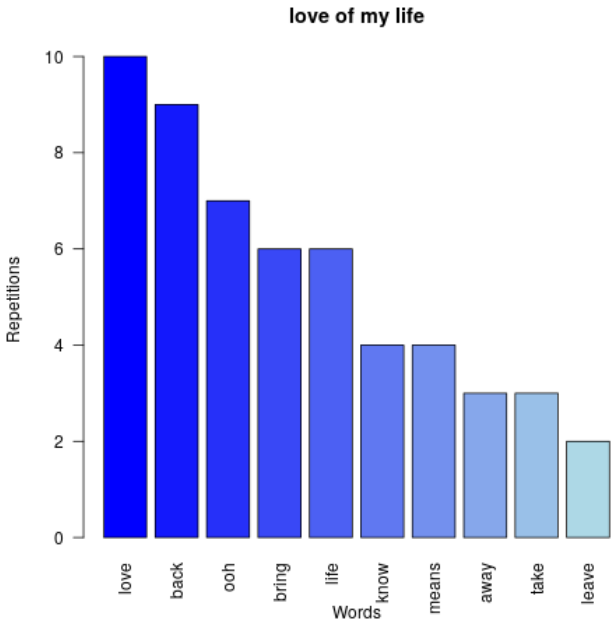

Figura 4.4: *Love Of My life* - Queen

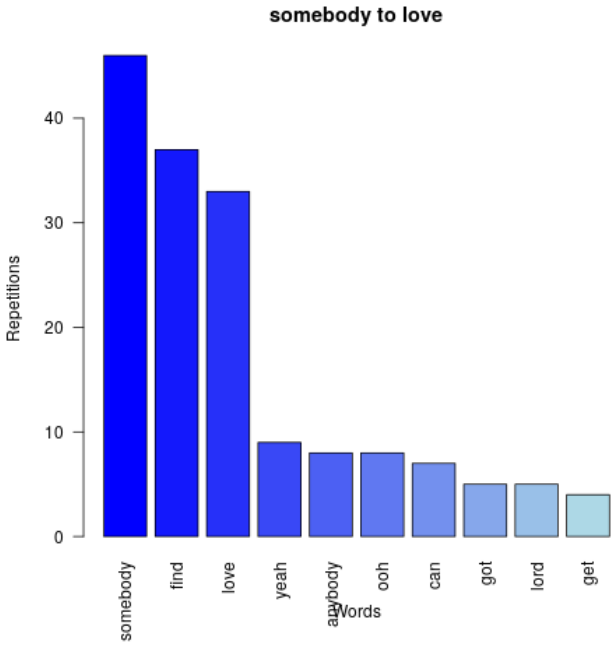

Figura 4.5: *Somebody To Love* - Queen

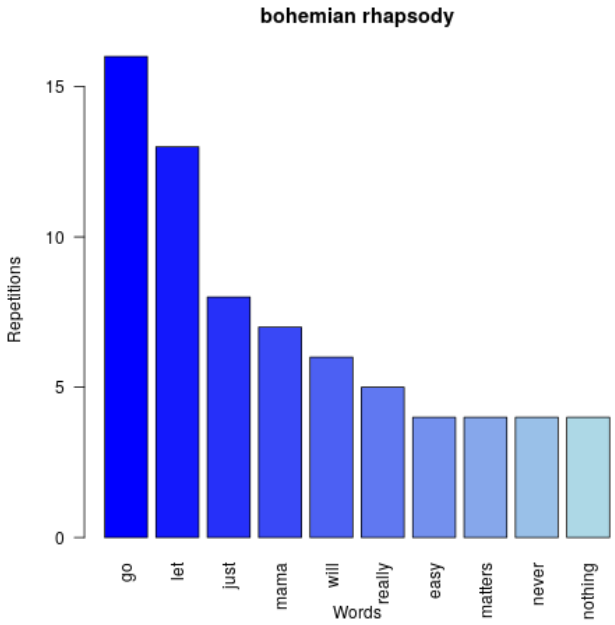

Figura 4.6: *Bohemian Rhapsody* - Queen

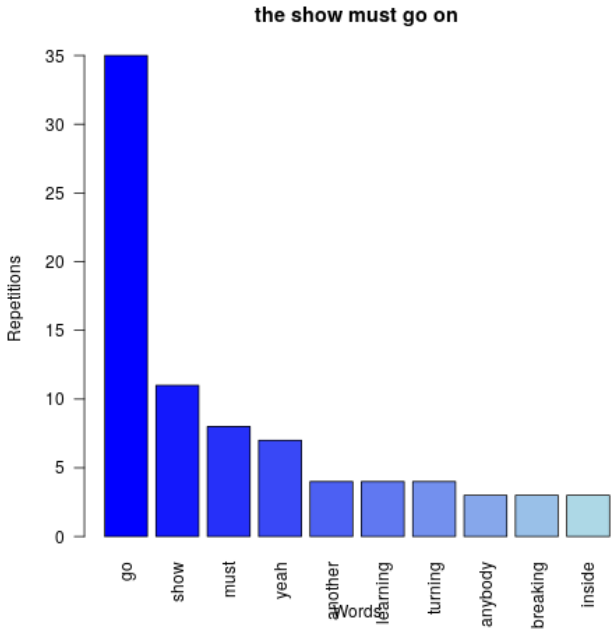

Figura 4.7: *The Show must go on* - Queen

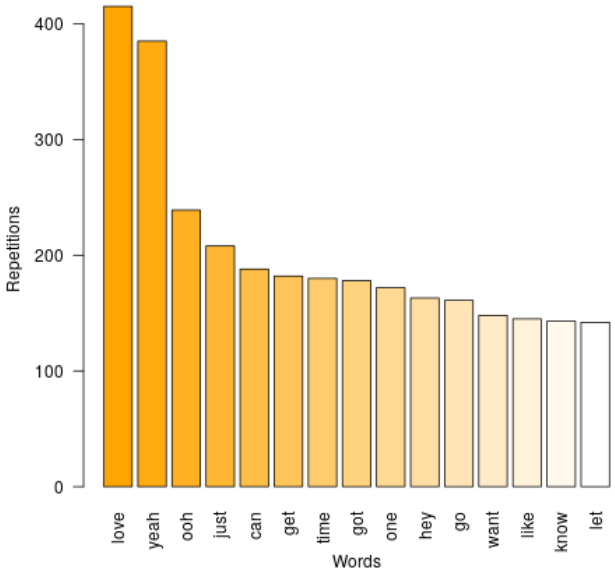

Figura 4.8: Most used words by Queen

Capítulo 5

Resultados y Evaluación

Capítulo 6

Conclusiones y Trabajo Futuro

Conclusiones del trabajo y líneas de trabajo futuro.

–¿Qué te parece desto, Sancho? – Dijo Don Quijote –
Bien podrán los encantadores quitarme la ventura,
pero el esfuerzo y el ánimo, será imposible.

Segunda parte del Ingenioso Caballero
Don Quijote de la Mancha
Miguel de Cervantes

–Buena está – dijo Sancho –; fírmela vuestra merced.
–No es menester firmarla – dijo Don Quijote–,
sino solamente poner mi rúbrica.

Primera parte del Ingenioso Caballero
Don Quijote de la Mancha
Miguel de Cervantes