BENEMÉRITA UNIVERSIDAD AUTÓNOMA DE PUEBLA

FACULTAD DE CIENCIAS DE LA COMPUTACIÓN

Reconocimiento de Patrones de Repetición en Canciones usando Técnicas de Estadística Bayesiana

Tesina para obtener el título de Licenciatura en

Ingeniería en Ciencias de la Computación

PRESENTA:

Karla Karina Gómez Matías

Director de Tesis:

Dra. María Teresa Torrijos Muñoz

FEBRERO 2024

Resumen

El presente trabajo consiste en el desarrollo de una aplicación de clasificación de textos automático, que permite asignar la categoría a la oración ingresada de acuerdo con el texto de entrenamiento (letras de canciones), mediante el uso de una técnica de aprendizaje automático de la familia de los clasificadores Naive Bayes, usados para la clasificación de textos.

El teorema de Bayes, formulado por Thomas Bayes, calcula la probabilidad de que ocurra un evento en función del conocimiento previo de las condiciones relacionadas con un evento.

El clasificador Naive Bayes Simple considera la probabilidad de cada término dada la clase de forma binaria, es decir que la palabra está o no está en la clase, entonces la probabilidad condicional de la palabra, dada la clase puede no estar considerada. Con la finalidad de mejorar el clasificador se utilizó el algoritmo de clasificación Naive Bayes Multinomial, ya que considera la frecuencia de las palabras, para calcular la probabilidad condicional dada la clase, mejorando la asignación de las categorías del clasificador.

Palabras claves: clasificación de textos, aprendizaje automático, Teorema de Bayes, Naive Bayes Simple, Naive Bayes Multinomial.

Índice

Índice de tablas

Índice de Figuras

Índice de Listas

Capítulo 1: Introducción

La Ciencia de Datos analiza grandes cantidades de datos con base a diferentes modelos que aporten información, que en su estado normal no se ven a simple vista. El origen de los datos es diverso pueden ser datos tabulares, imágenes, sonidos, texto y video.

La Ciencia de Datos recopila, procesa y analiza los datos para encontrar soluciones eficaces, uno de los propósitos de este documento es conocer algunas investigaciones que se han realizado en torno a la ciencia de datos e implementarlos en una aplicación.

1.1 Conceptos Generales

La Ciencia de Datos es el proceso de construir un modelo representativo que se ajuste a los datos de observación, según el problema se pueden clasificar en tareas tales como clasificación, análisis de asociación, agrupamiento y regresión. Cada tarea de ciencia de datos utiliza algoritmos de aprendizaje específicos como arboles de decisión, redes neuronales artificiales, K- vecinos más cercanos (K-NN) entre otros.

1.2 Planteamiento del Problema

Día a día estamos en contacto con sistemas que recolectan información con el objetivo de modelar nuestro comportamiento, el convertir esa información en datos requiere del uso de herramientas estadísticas y computacionales siendo la Ciencia de Datos la encargada de ello.

Hoy en día las empresas comienzan a apostar por la Ciencia de Datos y el uso de tecnologías como el Big Data, Data Analytics y Maching Learning entre otras herramientas para rentabilizar los datos y predecir una variedad de situaciones comerciales.

Desafortunadamente muchas empresas están desperdiciando el valor de los datos por desconocimiento, burocracia o incapacidad de adaptación a las nuevas tecnologías, falta de organización o de profesionales capacitados en la materia.

1.3 Objetivos Generales.

Implementar el uso de un algoritmo Naive Bayes que permita identificar patrones de repetición en canciones, esto mediante el uso de algoritmo de clasificación de textos.

1.4 Objetivos Específicos del Proyecto.

- ✓ Establecer los procedimientos que se usaran para formular y evaluar el clasificador.
- ✓ Desarrollar el clasificador bayesiano multinomial.
- ✓ Realizar pruebas.
- ✓ Analizar y comparar los resultados de acuerdo con la precisión, recall y puntaje F1.

1.5 Justificación de la Investigación

El análisis de datos mejorar la toma de decisiones, proporcionar nuevas perspectivas, ayuda a combatir problemas de desarrollo social y de salud.

Hoy en día nuestro país se enfrenta a gran cantidad de retos y la facilidad o dificultad de su resolución surge en gran necesidad de que herramientas tenemos a nuestra disposición. Las decisiones basadas en datos representan una ventaja competitiva para cualquier ámbito.

Lo anterior motiva este trabajo, en el que se pretende realizar un análisis de letras de canciones de diferentes géneros musicales para realizar un clasificador que pueda reconocer el género/categoría de una canción basándose únicamente en la letra.

Capítulo 2: Estado del Arte

Se analizaron cinco artículos científicos relacionados con la investigación de la ciencia de datos para conocer las áreas de oportunidad

2.1 Review on Data Science and Prediction

En el artículo Revisión sobre la Ciencia de Datos y Predicción nos menciona que la ciencia de datos ha sido muy eficaz para combatir los desafíos que enfrenta el Big Data ya que se ha producido una expansión de datos gracias a los dispositivos inteligentes, las redes sociales y la web.

Casi todos los aspectos de la vida humana se interconectaron con grandes cantidades de información creando una necesidad para las personas que pudieran administrar y rastrearla recopilación de información. Desde el punto de vista de la ingeniería la increíble escala del Big Data ha demostrado que muchos modelos de base de datos están desactualizados, estos modelos se crearon para resumir datos y acceso rápido y no para el descubrimiento de conocimientos. Están optimizados para la consulta de los usuarios y no para el descubrimiento de varios patrones complejos.

La toma de decisiones de hoy se basa en Big Data, donde las computadoras actúan como mejores tomadores de decisiones que las personas, refiriéndose como mejores en el aspecto escalabilidad, precisión y costo.

En conclusión, el Big Data realiza la extracción de grandes volúmenes de conjuntos de datos mientras que la ciencia de datos utiliza algoritmos para entrenarlos y obtener información relevante.

2.2 A survey study of success factors in data science projects

En el artículo Un estudio de encuestas sobre los factores de éxito en proyectos de ciencias de datos se realizó una encuesta a 237 profesionales de la ciencia de datos sobre el uso de metodologías de gestión de proyectos para la ciencia de datos,

Las preguntas más relevantes fueron:

Q10 ¿Sueles seguir alguna metodología de proyectos de ciencia de datos?

Q11 ¿Seleccione la metodología que conoce y/o ha utilizado para sus proyectos de ciencia de datos?

Las metodologías más mencionadas son: CRISP-DM, Microsoft TDSP, ciclos de vida Agile DS Lifecycle, Domino DS Lifecycle, IBM FMDS, RAMSY.

Con respecto a las metodologías de ciencia de datos, Agile DS Lifecycle la más utilizado seguido de CRISP-DM y Microsoft TDSP. Solo el 25% de los encuestados afirman usan algún tipo de metodología de proyectos de ciencia de datos este bajo porcentaje muestra la falta de una metodología de procesos definidos para la gestión

de proyectos de ciencia de datos siendo una de las principales causas de las fallas actuales y de los desafíos de gestión por el contrario de los profesionales que se adhieren a una metodología de proyecto ponen un mayor énfasis en los riesgos y peligros potenciales del proyecto, el control de versiones, la canalización de implementación a producción y la seguridad y privacidad de los datos.

2.3 Classification and Analysis of Techniques and Tools for Data Visualization Teaching

En articulo Clasificación y Análisis de Técnicas y Herramientas para la Enseñanza de la Visualización de Datos su principal objetivo es mejorar la enseñanza de la visualización de los datos.

La visualización de los datos es una forma de representar gráficamente la información y los datos, se pueden usar varios gráficos diferentes para fines operativos, como mejorar la eficiencia, monitorear procesos, estudiar la distribución geográfica de datos, buscar tendencia y relaciones, su principal función está en el análisis de datos.

Presentan una nueva clasificación de las técnicas graficas de datos en tres grupos diferentes con atributos diferentes de los gráficos:

- Uso: se definen según el uso o tratamiento que hagan de los datos crudos o procesados (Exploratorio o Explicativo).

- Forma: se pueden definir de acuerdo con los diferentes elementos gráficos utilizados para construirlos que se pueden resumir en su forma siendo el mayor atributo utilizado en la clasificación de gráficos de datos, existen diferentes subtipos de grafos: Hachas, círculos, mapas, rectángulos, redes y dibujos.
- Movimiento se definen según su movimiento es decir si la pantalla tiene o no elementos que varíen según el tiempo (Estático y Dinámico).

Se realizo un análisis comparativo de las principales herramientas de visualización de datos, dando como resultado que las ventajas de utilizar software libre es que tiene una mayor variedad de visualización de datos y una mayor capacidad de personalización de esta. El software con licencia, su principal ventaja es la velocidad al crear cualquier implementación.

2.4 Estudio de demanda y empleabilidad de la carrera de ingeniería en ciencia de datos en el Ecuador

Mediante una encuesta aplicada a las instituciones educativas del centro del país Ecuador, utilizando recursos virtuales, se obtuvieron como respuesta que 4,210 estudiantes y 81 instituciones públicas y privadas obteniendo como resultado una tendencia favorable para justificar la creación de la carrera de Ingeniería en Ciencia de Datos en el Ecuador.

El instrumento utilizado para la recolección fue una encuesta que consta de 18 preguntas y fue generada en línea lo que permitió obtener información sobre la

necesidad e intereses de estudiar la carrera de Ciencia de Datos, permitiendo conocer los criterios valorados entre el perfil profesional versus la oferta académica del programa,

Las tendencias de interés en el área de ciencia de datos se obtiene un interés del 57.93% en instituciones públicas, frente a un 43.03% de instituciones privadas. En cuanto a nivel estudiantil se determina que un 82.54% de los encuestados le interesaría una carrera profesional en área de ciencia de datos.

La importancia de las Tecnologías de la información y el análisis de datos en esta era hacen necesario la creación de profesionales en ciencia de datos, que sean capaces de seleccionar, preparar, analizar, evaluar y comunicar cantidades masivas de datos de cualquier tipo, de manera ética y responsable para la toma de decisiones inteligentes y la resolución de problemas complejos en los sectores científicos, tecnológicos, empresariales y sociales

2.5 Clasificación automática de papers de Ciencias de la Computación

La tesis consiste en el desarrollo de un sistema de clasificación automática de papers (artículos científicos), que permiten asignar la categoría adecuada de acuerdo con el contenido de estas características puede ahorrar a los publicadores el tiempo significativo que implica atravesar un proceso de selección de categorías jerárquicas para clasificarlos correctamente, y de esta manera facilitar su búsqueda una vez que se han publicado en los repositorios bibliográficos online.

Para la implementación de este sistema se obtuvo una colección de artículos de todo tipo del sitio web de la Librería Digital de la ACM (Association for Computing Machinery) a partir de la cual se generaron los data sets de entrenamiento que se utilizarán junto con una herramienta llamada MALLET para determinar la categoría de papers que aún no han sido clasificados, estudiando el caso particular de las últimas taxonomías que presenta dicho sitio (versiones de 1998 y 2012).

La clasificación automática de textos siempre ha sido un tema de investigación importante desde la existencia de documentos digitales, ya que ésta permite manejar la enorme cantidad de información disponible en la web. Está basada principalmente en técnicas de machine learning, las cuales construyen automáticamente un clasificador que aprende las características de las categorías a partir de un conjunto preclasificado de documentos. Estas técnicas juegan un rol muy importante en la extracción y resumen de información, recuperación de texto, entre otros.

En este trabajo se ha hecho una revisión general de los problemas de clasificación jerárquica, y un subconjunto de sus posibles soluciones. Se ha diseñado para tal fin una aplicación que permita automatizar el proceso de clasificación de artículos científicos y facilitar la configuración de los parámetros de esta, para poder comunicarse de manera sencilla con herramientas de clasificación externa, utilizando para este caso particular la implementación en Java de MALLET, la cual posee eficientes metodologías para realizar este tipo de tareas.

Capítulo 3 Marco Teórico

3.1 Ciencia de Datos

La Ciencia de Datos explora y analiza datos mediante tecnologías que provienen de las matemáticas, la estadística y la informática con el objetivo de obtener información para la toma de decisiones; entre las que encontramos el análisis exploratorio, el aprendizaje automático, el aprendizaje profundo el procesamiento del lenguaje natural, la visualización de datos y el diseño experimental.

Las dos tecnologías que más se utilizan dentro de la ciencia de datos son el Machine Learning y el Deep Learning ambas englobadas en el campo de la IA.

En ambos casos se busca la construcción de sistemas que sean capaces de aprender a resolver problemas a partir de conjuntos de datos que se entrenan de forma:

- ✓ supervisada: cuando los datos de entrenamiento están previamente etiquetados por humanos

- ✓ No supervisada cuando el conjunto de datos no está etiquetado.

Ejemplos de las tecnologías: sistemas de predicción ortográfica o traducción automática hasta los coches autónomos o los sistemas de visión artificial.

3.2 Machine Learning

El aprendizaje automático se dedica al estudio de agentes/algoritmos como: redes bayesianas, máquinas de vectores de soporte, análisis de clústeres, para analizar y procesar datos, que aprenden o evolucionan basados en su experiencia, consiste en dejar que los agentes descubran patrones recurrentes a partir de observaciones en los conjuntos de datos, estos datos pueden ser números, palabras, imágenes, estadísticas, etc.

Existe una amplia variedad de aplicaciones que utilizan agentes basados en aprendizaje entre ellas se encuentran:	
procesamiento de Lenguaje Natural	se utiliza para el análisis sintáctico y morfológico de los textos, extracción de Información, clasificación automática de documentos.
En el diagnóstico médico	según la historia clínica y los síntomas que presenta el paciente
En Biología	reconocimiento de tumores, patrones en cadenas de ADN.
En finanzas e industrias bancarias	para definir modelos de riesgo y fraude crediticio.
En el análisis de imágenes	detectar objetos dentro de una imagen como rostros, personas, etc.

Tabla 1 Ejemplos de aplicaciones de aprendizaje automático

3.3 Clasificación de Textos

La clasificación de textos consiste en determinar la categoría de un texto que establezca la clase a la que pertenece de acuerdo con ciertas características, facilitando la organización de la información

La decisión sobre que etiqueta debe asignar a las palabras se determina a partir de un modelo construido con el texto de entrenamiento.

3.4 Procesamiento del Lenguaje Natural

El procesamiento del lenguaje natura NLP permite que las computadoras interpreten el lenguaje humano.

El lenguaje puede expresarse por escrito (texto) u oralmente (voz). El NLP está más avanzado en el tratamiento de textos, ya que los datos son fáciles de encontrar en formato electrónico.

Las computadoras solo entienden de bytes y dígitos es por lo que el texto debe convertirse en números, específicamente en vectores de números.

3.4.1 Extracción de características o codificación de características

1. Tokenization

Divide el texto en pequeñas unidades como palabras o frases

2. Stop words

Las palabras generalmente se filtran antes de procesar un lenguaje natural se denominan stop words, son las palabras más comunes en cualquier idioma como artículos, preposiciones, pronombres y conjunciones, se filtran ya que no agregan mucha información al texto

3. Modelo BoW (Bag of Words)

Es un modelo que representa un texto como una bolsa de palabras, sin considerar su orden o relación entre ellas, extraer características del texto, es decir la frecuencia de palabras de un documento.

Representándose con el modelo vectorial en que cada documento es representado como un vector de dimensión igual al tamaño del vocabulario de la clase, y en que el valor de cada atributo corresponde a la frecuencia de la palabra en la clase correspondiente.

3.5 Algoritmo por Clasificación

Este algoritmo está basado en el teorema de Bayes y clasifica cada valor como independiente de cualquier otro. Permite predecir una categoría en función de un conjunto dado, utilizando la probabilidad.

3.3.1 Naive de Bayes

Es de los clasificadores más utilizados por su facilidad y rapidez. Constituye una técnica supervisada que necesita tener ejemplos clasificados para funcionar.

El funcionamiento del algoritmo destaca en dos partes:

1. **Crear el modelo**

- ✓ Calcula las probabilidades a priori de cada clase.
- ✓ Para cada clase, realiza un recuento de los valores de atributos que toma cada ejemplo. Se debe distribuir cada clase por separada para mayor eficiencia del algoritmo.
- ✓ Aplicar la Corrección de Laplace.
- ✓ Normalizar para obtener un rango de valores [0,1].

2. Clasificar

1. Para cada clase disponible se determinan los valores de probabilidad de cada valor de los atributos del nuevo ejemplo
2. Aplicar la fórmula de Naive Bayes.

3.3.2 Matemáticas del algoritmo Naive de Bayes

Dado un vector de características X= $(x_1, x_2, \ldots, x_n)$ y una variable de clase Y, el teorema de Bayes establece que:

$$P(y|X) = \frac{P(X|y) * P(y)}{P(X)} \qquad (1)$$

Donde:

$P(y)$ es la probabilidad a priori, y es la probabilidad de que la hipótesis sea cierta (independiente de los datos).

$P(X|y)$ es el likelihood de una clase Y dados los datos X, probabilidad de los datos dado que la hipótesis era cierta

$P(X)$ evidencia es la probabilidad de los datos (independientemente de la hipótesis.)

$P(y|X)$es la probabilidad posteriori de la hipótesis dado los datos, la distribución de probabilidad final para la clase.

Utilizando la regla de la cadena, la probabilidad $P(X|y)$ se puede descomponer como:

$$P(X|y) = P(x_1, x_2, \dots, x_n|y)$$

$$P(X|y) = P\left(x_1 \middle| x_2 \dots x_{n,} y\right) * P\left(x_2 \middle| x_3 \dots x_{n,} y\right) \dots P(x_n|y) \quad (2)$$

Debido a la suposición de independencia condicional de Naive, las probabilidades condicionales son independientes entre sí.

$$P(X|y) = P(x_1|y) * P(x_2|y) \dots P(x_n|y) \quad (3)$$

Entonces por independencia condicional:

$$P(y|X) = \frac{P(x_1|y) * P(x_2|y) \dots P(x_n|y) * P(y)}{P(x_1) * P(x_2) \dots P(X_n)} \quad (4)$$

Y como el denominador permanece constante para todos los valores, la probabilidad posteriori puede ser:

$$P(y|x_1, x_2, \dots, x_n) = P(y) \prod_{i=1}^{n} P\left(x_1|y\right) \quad (5)$$

El clasificador Naive Bayes combina este modelo con una regla de decisión. Una regla común es elegir la hipótesis que sea más probables: o la regla de decisión máxima a posteriori o MAP (máximum a posteriori hypothesis).

$$y = argmax_y P(y) \prod_{i}^{n} P\left(x_1|y\right) \quad (6)$$

3.3.3 Independencia Naive Bayes Multinomial

La independencia condicional asume que las probabilidades de $P(x_i \mid y_j)$ son independientes de la clase Y.

$$P(x_1, x_{2,} \dots x_n | y)$$

$$P(x_1, x_{2,} \dots x_n | y) = P(x_1|y) * P(x_2|y) * P(x_3|y) * \dots * P(x_n|y)$$

$$Y_{MAP} = argmax_{y \in Y} P(x_1, x_{2,} \dots x_n | y) P(y)$$

$$Y_{NB} = argmax_{y \in Y} P(y_j) \prod_{i}^{n} P(x_i | y_j) \qquad (7)$$

Los valores $P(x_i \mid y_j)$ se estiman con la frecuencia de los datos, no se hace busqueda de hipotesis, sino que se cuenta la frecuencia de ocurrencias.

Capítulo 4 Desarrollo del Clasificador

Con el objetivo de identificar el género de la canción basándose en parte de la letra (frase, coro) de forma automática, se desarrolla una aplicación en el lenguaje de programación Java, el método para realizar la clasificación es por Naive Bayes Multinomial, dando como resultado el género musical u interprete.

4.1 Método de clasificación jerárquica

Se clasificaron las canciones por medio de una organización jerárquica teniendo 5 categorías, cada categoría contiene de 5 a 10 canciones.

4.2 Recopilación de datos

El primer paso fue obtener el conjunto de datos; todas las letras de las canciones se obtuvieron de la página web música.com.

4.3 Clasificación de texto en el clasificador

Dada una clase Y y un conjunto de palabras X de la canción a clasificar, se calcula la probabilidad de que X_i se clasifique dentro de la clase Y_i , con ello se conoce la probabilidad de X_i dada una clase y la probabilidad de la clase Y_i pero se necesita encontrar el valor máximo de la expresión para encontrar la clase en la que mejor se clasifique X_i.

La definición de las clases son las categorías/genero de las canciones y la clasificación se realiza por palabras.

4.4 Implementación

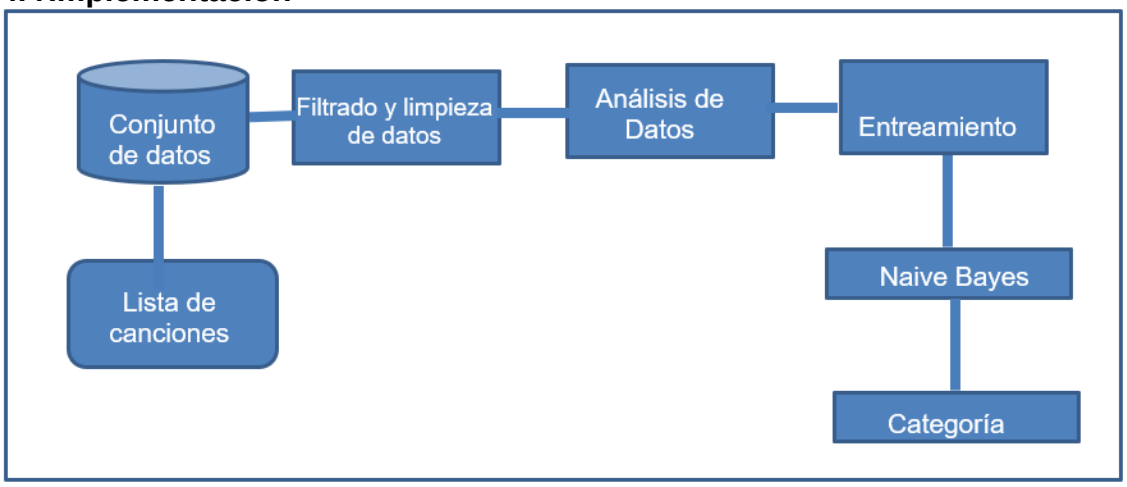


Fig. 1Arquitectura del Sistema

4.5 Procesamiento del algoritmo

a) Conjunto de datos

Al ser una clasificación supervisada, los datos ya están previamente clasificados manualmente en 5 categorías, contando con un total de 9238 palabras, en la tabla2 se muestra la cantidad de palabras por categorías.

.

Categoría	Conjunto de datos.
Queen/ Rock	1412
Eminen/ Hip Hop	4259
Daft Punk/ Electrónica	1915
Tiesto/Electropop	1033
Other	619
TOTAL	9238

Tabla 2 Total del conjunto de datos.

Para el entrenamiento del clasificador se ingresan las letras de las canciones que se muestran en la tabala3.

Queen/ Rock	Eminen/ Hip Hop	Daft Punk/ Electrónica	Tiesto/Electropop	Other
Bohemian rhapsody	8 mile	Instant crush	Century	BabaO'Riley
I want to break free	Darkness	Get lucky	Escapeme	I love rock ´n' roll
The show must go on	Lose yourself	Atom bomb	Feel it in my bones	Rockin' in the free world
Under pressure	Space bound	Harder, better, faster, stronger	Here on earth	Dream on
Don't stop me now	The real Slim shady	One more time	It's not the things you say	Born to be wild
Another one bites the dust	The way I am	Lose yourself to dance	I will be here	
Some body to love	Tilll collapse	Technologic	Red lights	
We are the champions	Without me	Face to face	Ritual	
Crazy Little thing called love	Rap god	Give life back to music	You are my diamond	
Who wants to live forever		Fragments of time		

Tabla 3 Títulos de las canciones (data set)

Los datos de entrenamiento están dentro de la clase Análisis, declarados como variables de tipo String, son un total de 43 canciones.

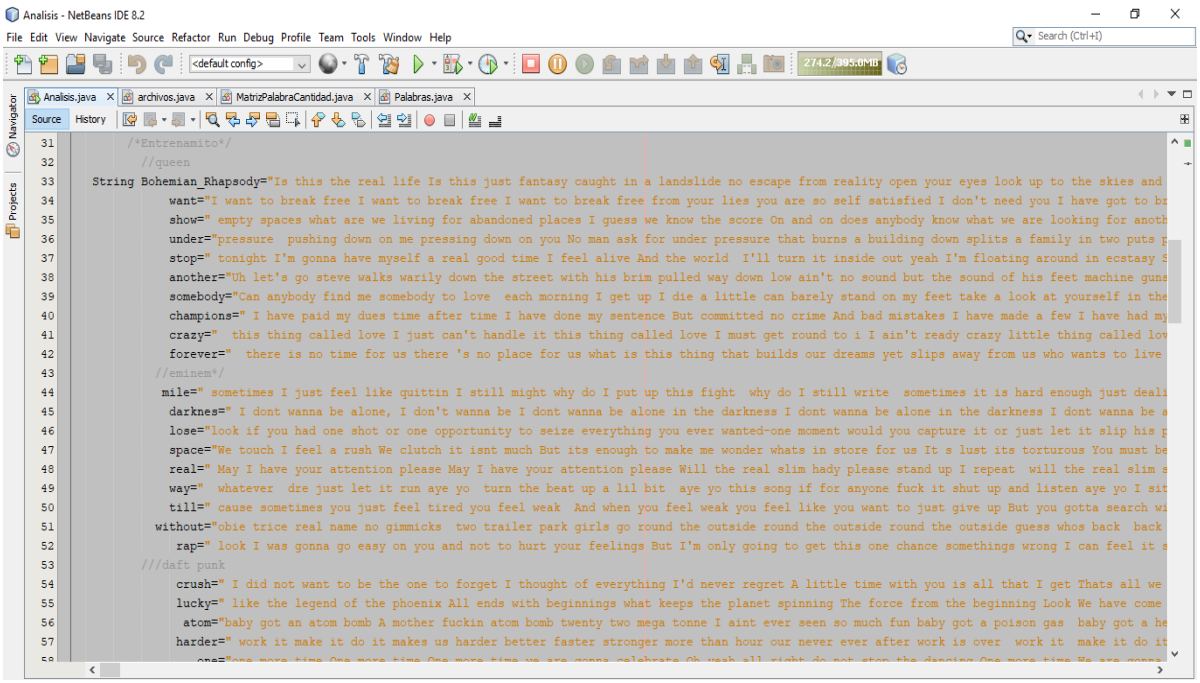


Fig. 2 Datos de entrenamiento (data set)

En la clase **Palabras** se crea un arrayList llamado palabrasL (lista de palabras) en el cual se va añadiendo cada una de las palabras en el método **IngresarPalabra.**

El método **verPalabras** imprime el arrayList palabrasL, que es el conjunto por considerar.

b) Filtrado y Limpieza de datos

✓ **Tokenization**

La tokenización se realiza dentro del método **separarpalabras** con ayuda de un Apalabras (Arreglo de palabras) y la función Split separa la palabra de la cadena, después de un espacio o de algunos de estos caracteres [, @ ‘!] + obteniendo las palabras/token´s.

```
************
0-this
1-real
2-life
3-just
4-fantasy
5-caught
6-landslide
7-escape
8-from
9-reality
10-open
11-your
12-eyes
13-look
14-skies
15-poor
16-need
17-sympathy
18-because
19-easy
20-come
21-little
22-high
23-anyway
24-wind
25-blows
26-does
27-really
28-matter
29-mama
30-killed
31-against
32-head
33-pulled
34-trigger
35-dead
36-begun
37-have
38-gone
39-thrown
40-away
41-mean
42-make
43-back
44-again
45-time
46-tomorrow
47-carry
48-nothing
49-matters
50-late
51-sends
52-shivers
53-down
54-spine
55-body
56-aching
57-goodbye
58-everybody
59-gotta
60-leave
61-behind
62-face
63-truth
64-wanna
65-sometimes
66-wish
67-never
68-been
69-born
70-silhouetto
71-scaramouche
72-will
73-fandango
74-thunderbolt
75-lightning
76-very
77-frightening
78-galileo
79-figaro
80-magnifico
81-nobody
82-loves
83-family
84-spare
85-monstrosity
86-bismillah
87-mamma
88-beelzebub
89-devil
90-aside
91-think
92-stone
93-spit
94-love
95-baby
96-right
97-outta
98-here
99-yeah
100-anyone
101-want
102-break
103-free
104-from
105-your
106-lies
107-self
108-satisfied
109-need
110-have
111-knows
112-fallen
113-love
114-first
115-time
116-this
117-know
118-real
119-yeah
120-strange
121-true
122-over
123-like
124-sure
125-when
126-walk
127-that
128-door
129-baby
130-life
131-still
132-goes
133-used
134-living
135-without
136-side
137-live
138-alone
139-make
140-heve
141-empty
142-spaces
143-what
144-living
145-abandoned
146-places
147-guess
148-know
149-score
150-does
151-anybody
152-looking
153-another
154-hero
155-mindless
156-crime
157-behind
158-curtain
159-pantomime
160-hold
161-line
162-want
163-take
164-anymoreThe
165-show
166-must
167-yeah
168-Inside
169-heart
170-breaking
171-make
172-flaking
173-smile
174-still
175-stays
176-whatever
177-happens
178-leave
179-chance
180-heartache
181-failed
182-romance
183-learning
184-warmer
185-soon
186-turning
187-round
188-corner
189-outside
190-dawn
191-inside
192-dark
193-aching
194-free
195-soul
196-painted
197-like
198-wings
199-butterflies
200-fairytales
201-yesterday
202-will
203-grow
204-never
205-friends
206-face
207-with
208-grin
209-giving
210-bill
211-overkill
212-have
213-find
214-carry
215-pressure
216-pushing
217-down
218-pressing
219-under
220-that
221-burns
222-building
223-splits
224-family
225-puts
226-people
227-streets
228-That
229-okay
230-terror
231-knowing
232-what
233-this
234-world
235-about
236-watching
237-some
238-good
239-friends
240-screaming
241-pray
242-tomorrow
243-gets
244-higher
245-chipping
246-around
247-kick
248-brains
249-floor
250-these
251-days
252-never
253-rains
254-pours
255-screaming:
256-high
257-turned
---
```

258-away
259-from
260-like
261-blind
262-fence
263-work
264-keep
265-coming
266-with
267-love
268-slashed
269-torn
270-Love
271-insanity
272-laughs
273-breaking
274-give
275-ourselves
276-more
277-chance
278-cause
279-such
280-fashioned
281-word
282-dares
283-care
284-edge
285-night
286-loves
287-change
288-caring
289-last
290-dance
291-tonight
292-gonna
293-have
294-myself
295-real
296-good
297-time
298-feel
299-alive
300-world
301-turn
302-inside
303-yeah
304-floating
305-around
306-ecstasy
307-stop
308-cause
309-having
310-shooting
311-star
312-leaping
313-through
314-like
315-tiger
316-defying
317-racing
320-passing
321-Lady
322-Godiva
323-there
324-stopping
325-burning
326-hundred
327-degrees
328-that
329-they
330-call
331-Mister
332-Fahrenheit
333-travelling
334-speed
335-light
336-wanna
337-make
338-supersonic
339-such
340-just
341-give
342-(cause
343-havin
344-rocket
345-ship
346-Mars
347-collision
348-course
349-satellite
350-control
351-machine
352-ready
353-reload
354-atom
355-bomb
356-about
357-explode
358-woman
359-ball
360-steve
361-walks
362-warily
363-down
364-street
365-with
366-brim
367-pulled
368-sound
369-feet
370-machine
371-guns
372-ready
373-this
374-hanging
375-edge
376-your
377-seat
378-doorway
379-bullets
380-beat
381-yeah
382-another
383-bites
384-dust
385-gone
386-gonna
387-think
388-going
389-along
390-without
391-when
392-took
393-everything
394-that
395-kicked
396-happy
397-satisfied
398-long
399-stand
400-heat
401-look
402-take
403-bite
404-shoot
405-there
406-plenty
407-ways
408-hurt
409-bring
410-ground
411-cheat
412-treat
413-leave
414-standing
415-repeating
416-anybody
417-find
418-somebody
419-love
420-each
421-morning
422-little
423-barely
424-stand
425-feet
426-take
427-look
428-yourself
429-mirror
430-Lord
431-what
432-doing
433-have
434-spent
435-years
436-believing
437-just
438-relief
439-work
440-hard
441-works
442-everyday
443-life
444-till
445-ache
446-bones
447-home
448-earned
449-goes
450-down
451-knees
452-start
453-pray
454-praise
455-tears
456-from
457-eyes
458-please
459-wants
460-help
461-every
462-everybody
463-they
464-goin
465-crazy
466-water
467-brain
468-common
469-sense
470-nobody
471-left
472-believe
473-yeah
474-someone
475-feel
476-rhythm
477-keep
478-losing
479-beat
480-alright
481-gonna
482-face
483-defeat
484-gotta
485-this
486-prison
487-cell
488-some
489-free
490-anywhere
491-have
492-paid
493-dues
494-time
495-after
496-done
497-sentence
498-committed
499-crime
500-mistakes
501-made
502-share
503-sand
504-kicked
505-face
506-come
507-through
508-mean
509-champions
510-friends
511-keep
512-fighting
513-till
514-losers
515-cause
516-world
517-taken
518-bows
519-curtain
520-calls
521-brought
522-fame
523-fortune
524-everything
525-that
526-goes
527-with
528-thank
529-been
530-roses
531-pleasure
532-cruise
533-consider
534-challenge
535-before
536-whole
537-human
538-race
539-gonna
540-lose
541-this
542-thing
543-called
544-love
545-just
546-handle
547-must
548-round
549-ready
550-crazy
551-little
552-(this
553-thing)
554-cries

555-like
556-baby
557-cradle
558-night
559-swings
560-jives
561-shakes
562-over
563-jelly
564-fish
565-kinda
566-there
567-goes
568-knows
569-rock
570-roll
571-drives
572-gives
573-cold
574-fever
575-leaves
576-cool
577-sweat
578-gotta
579-relax
580-track
581-take
582-back
583-seat
584-Hitchhike
585-long
586-ride
587-motor
588-bike
589-until
590-hitchhike
591-Freddie
592-yeah
593-there
594-time
595-place
596-what
597-this
598-thing
599-that
600-builds
601-dreams
602-slips
603-away
604-from
605-wants
606-live
607-forever
608-chance
609-decided
610-world
611-only
612-sweet
613-moment
614-aside
615-dares
616-love
617-when
618-must
619-touch
620-tears
621-with
622-your
623-lips
624-fingertips
625-have
626-today
627-sometimes
628-just
629-feel
630-like
631-quittin
632-still
633-might
634-this
635-fight
636-write
637-hard
638-enough
639-dealin
640-with
641-real
642-life
643-wanna
644-jump
645-stage
646-kill
647-mics
648-show
649-these
650-people
651-what
652-level
653-skills
654-white
655-hate
656-somethin
657-aint
658-right
659-brake
660-lights
661-case
662-fright
663-drawin
664-blank
665-fault
666-great
667-then
668-falls
669-insides
670-crawl
671-clam
672-wham
673-slam
674-shut
675-whole
676-manhoods
677-been
678-stripped
679-have
680-wicked
681-must
682-split
683-fuck
684-shit
685-goin
686-home
687-world
688-shoulder
689-back
690-mile
691-road
692-Chorus
693-make
694-plan
695-time
696-stand
697-travel
698-land
699-take
700-matters
701-into
702-hands
703-once
704-over
705-tracks
706-never
707-look
708-gone
709-know
710-where
711-sorry
712-momma
713-grown
714-alone
715-follow
716-footsteps
717-making
718-only
719-that
720-escape
721-from
722-eminem
723-walkin
724-train
725-tryin
726-regain
727-spirit
728-fore
729-same
730-crap
731-plant
732-pants
733-chase
734-gotta
735-move
736-asap
737-mommas
738-poor
739-little
740-baby
741-sister
742-dont
743-understand
744-sits
745-front
746-buries
747-nose
748-colors
749-until
750-crayon
751-gets
752-dull
753-hand
754-while
755-brother
756-mother
757-Aint
758-tellin
759-really
760-goes
761-head
762-wish
763-could
764-daddy
765-neither
766-keep
767-runnin
768-wanted
769-upset
770-cause
771-blew
772-grew
773-grow
774-nuts
775-Dont
776-step
777-pressures
778-much
779-whats
780-best
781-wont
782-tell
783-moment
784-pray
785-please
786-beggin
787-pigeon
788-holed
789-regular
790-hope
791-hear
792-homey
793-wherever
794-dawg
795-bailin
796-trailer
797-tomorrow
798-love
799-kiss
800-goodbye
801-whenever
802-need
803-there
804-second
805-blow
806-everything
807-work
808-live
809-didnt
810-wouldnt
811-deal
812-wasn
813-skillest
814-borderline
815-Detroit
816-city
817-limits
818-different
819-certain
820-significance
821-certificate
822-authenticity
823-even
824-credibility
825-seen
826-heard
827-smelled
828-incredible
829-upon
830-pedestal
831-unsigned
832-havin
833-rough
834-porch
835-friends
836-kick
837-dumb
838-rhymes
839-serve
840-lunchline
841-when
842-comes
843-crunch
844-punchlines
845-bust
846-flow
847-another
848-crab

907-suddenly
848-crab 908-surge
849-bucket 909-burst
850-luck 910-energy
851-Rabbit 911-occured
852-maybe 912-free
853-outlet 913-leaders
854-startin 914-three
855-doubt 915-third
856-feelin 916-scared
857-skeptic: 917-bird
858-hang 918-cross
859-clothes 919-median
860-about 920-curb
861-salvati(921-verbs
862-Army 922-blur
863-salvage 923-dont
864-outfit 924-wanna
865-cold 925-alone
866-plus 926-darkness
867-stuck 927-anymore
868-battlin 928-hello
869-mode 929-friend
870-defense: 930-here
871-thing 931-again
872-want 932-cant
873-pity 933-this
874-dark 934-hole
875-bein 935-like
876-pulled 936-walls
877-apart 937-closin
878-each 938-help
879-limbs 939-feel
880-skin 940-these
881-doin 941-curtains
882-stove 942-open
883-explode 943-something
884-kettle 944-pulls
885-mouth 945-closed
886-overloads 946-feels
887-learned 947-loathing
888-turn 948-vegas
889-takes 949-haven
890-burned 950-vagueat
891-fallin 951-lost
892-next 952-make
893-meet 953-small
894-girl 954-wager
895-longer 955-tomorrows
896-play 956-paper
897-stupid 957-would
898-immature 958-odds
899-every 959-favor
900-ingredien 960-much
901-courage 961-father
902-already 962-think
903-beat 963-that
904-words 964-knew
905-urge 965-keep
966-pacin
907-room
968-valium
969-then
970-chase
971-with
972-booze
973-little
974-taste
975-maybe
976-take
977-snooze
978-tear
979-stage
980-fuck
981-colt
982-need
983-somethin
984-stronger
985-caps
986-better
987-vodka
988-round
989-after
990-gettin
991-loaded
992-thats
993-shots
994-double
995-entendre
996-starin
997-service
998-menu
999-benzo
1000-hear
1001-music
1002-continue
1003-crescendo
1004-whole
1005-fuckin
1006-venue
1007-from
1008-window
1009-when
1010-know
1011-youre
1012-schizo
1013-cause
1014-peakin
1015-curtain
1016-hotel
1017-loud
1018-almost
1019-though
1020-sound
1021-should
1022-ready
1023-show
1024-wait
1024-wait
1025-crowd
1026-thought
1027-shit
1028-sold
1029-only
1030-opening
1031-early
1032-overreact
1033-told
1034-relax
1035-just
1036-hope
1037-packed
1038-before
1039-they
1040-fill
1041-each
1042-thatd
1043-totally
1044-wack
1045-murder
1046-nobodys
1047-what
1048-nobody
1049-shows
1050-panic
1051-mode
1052-bout
1053-snap
1054-motherfuckin
1055-wacko
1056-second
1057-cancel
1058-fans
1059-below
1060-rush
1061-entrance
1062-plan
1063-wreck
1064-cameras
1065-directions
1066-press
1067-apeshit
1068-bananas
1069-networks
1070-commando
1071-extra
1072-clips
1073-ammo
1074-hecklers
1075-armed
1076-teeth
1077-nother
1078-Valium
1079-fall
1080-ground
1081-crawl
1082-dresser
1083-alcohol
1084-breath
1085-reach
1086-scope
1087-blackin
1088-meds
1089-them
1090-benzodiazepines
1091-gone
1092-magazines
1093-sprawled
1094-floor
1095-media
1096-goin
1097-yeah
1098-people
1099-start
1100-time
1101-10:05
1102-starts
1103-already
1104-sweatin
1105-locked
1106-rapid
1107-fire
1108-spittin
1109-concert-goers
1110-scopes
1111-sniper
1112-vision
1113-surprise
1114-nowhere
1115-slide
1116-clip
1117-inside
1118-leanin
1119-going
1120-Kaiser
1121-soze
1122-finger
1123-trigger
1124-licensed
1125-owner
1126-prior
1127-convictions
1128-says
1129-skys
1130-limit
1131-supplies
1132-infinite
1133-strapped
1134-soldier
1135-hopping
1136-over
1137-climbing
1138-fences
1139-some
1140-john
1141-travolta

2092-constant
2093-lyrical
2094-content
2095-guilty
2096-conscience
2097-gotten
2098-such
2099-rotten
2100-responses
2101-controversy
2102-circles
2103-seems
2104-media
2105-immediately
2106-points
2107-finger
2108-point
2109-index
2110-pinky
2111-ring
2112-thumb
2113-give
2114-bullshit
2115-they
2116-pull
2117-full
2118-dudes
2119-bullied
2120-shoos
2121-school
2122-blame
2123-marylin
2124-heroine
2125-where
2126-were
2127-parents
2128-look
2129-middle
2130-america
2131-tragedy
2132-upper
2133-class
2134-city
2135-havin
2136-happenin
2137-attack
2138-eminem
2139-glad
2140-feed
2141-fuel
2142-That
2143-need
2144-fire
2145-burn
2146-burnin
2147-returned
2148-wasn
2149-sick
2150-bein
2151-admired
2152-wished
2153-fired
2154-drop
2155-from
2156-label
2157-stop
2158-fables
2159-gonna
2160-able
2161-name
2162-pigeon
2163-hold
2164-some
2165-poppy
2166-sensation
2167-caught
2168-rotating
2169-rock
2170-roll
2171-stations
2172-patience
2173-deal
2174-these
2175-cocky
2176-casions
2177-wigger
2178-tries
2179-black
2180-talk
2181-accent
2182-grab
2183-ball
2184-always
2185-keep
2186-askin
2187-same
2188-questions
2189-hood
2190-grew
2191-till
2192-grabbin
2193-hair
2194-tearing
2195-drivin
2196-crazy
2197-take
2198-racin
2199-pacing
2200-stand
2201-thankful
2202-every
2203-bathroom
2204-without
2205-someone
2206-standing
2207-sign
2208-autograph
2209-asshole
2210-cause
2211-sometimes
2212-just
2213-feel
2214-tired
2215-weak
2216-when
2217-like
2218-want
2219-give
2220-gotta
2221-search
2222-within
2223-find
2224-that
2225-inner
2226-strength
2227-pull
2228-shit
2229-motivation
2230-quitter
2231-matter
2232-fall
2233-flat
2234-your
2235-face
2236-collapse
2237-spilling
2238-these
2239-raps
2240-long
2241-drop
2242-youll
2243-never
2244-killing
2245-then
2246-stop
2247-pinning
2248-them
2249-hiphop
2250-eminem
2251-subliminal
2252-thoughts
2253-sending
2254-women
2255-caught
2256-webs
2257-spin
2258-hook
2259-venom
2260-adrenaline
2261-shots
2262-penicillin
2263-could
2264-illin
2265-amoxicillin
2266-real
2267-enough
2268-criminal
2269-filling
2270-minimal
2271-swap
2272-millions
2273-listeners
2274-coming
2275-with
2276-gonna
2277-fear
2278-showed
2279-spirit
2280-lives
2281-hear
2282-lyrics
2283-shock
2284-miracle
2285-product
2286-fizzing
2287-shizzle
2288-whizzle
2289-this
2290-plot
2291-listen
2292-bizzles
2293-forgot
2294-slizzle
2295-does
2296-fuck
2297-roof
2298-comes
2299-till
2300-lights
2301-legs
2302-cant
2303-shut
2304-mouth
2305-smoke
2306-clears
2307-high
2308-perhaps
2309-bone
2310-music
2311-magic
2312-there
2313-certain
2314-feeling
2315-spit
2316-people
2317-moment
2318-every
2319-single
2320-minute
2321-spittin
2322-trying
2323-hold
2324-onto
2325-again
2326-while
2327-much

3032-special
3033-cable
3034-channel
3035-went
3036-radio
3037-station
3038-very
3039-next
3040-kill
3041-lyrics
3042-supersonic
3043-speed
3044-sama
3045-lamaa
3046-duma
3047-assuming
3048-human
3049-superhuman
3050-innovative
3051-rubber
3052-anything
3053-ricocheting
3054-glue
3055-devastating
3056-more
3057-than
3058-demonstrating
3059-give
3060-audience
3061-levitating
3062-fading
3063-haters
3064-forever
3065-waiting
3066-fell
3067-celebrating
3068-motivated
3069-elevating
3070-elevator
3071-mainstream
3072-thats
3073-jealous
3074-confuse
3075-found
3076-hella
3077-fuse
3078-rock
3079-shock
3080-throw
3081-lose
3082-songs
3083-dont
3084-words
3085-occurs
3086-ripping
3087-verses
3088-diverse
3089-curtains
3090-inadvertently
3091-hurtin
3092-many
3093-murder
3094-prove
3095-nice
3096-sacrifice
3097-virgins
3098-flunkie
3099-pill
3100-junky
3101-accolades
3102-skills
3103-brung
3104-Full
3105-myself
3106-bully
3107-mind
3108-million
3109-leagues
3110-above
3111-speak
3112-tongues
3113-tongue
3114-cheek
3115-drunk
3116-satan
3117-wheel
3118-asleep
3119-seat
3120-bumping
3121-heavy
3123-funky
3124-something
3125-tugging
3126-struggling
3127-angels
3128-fight
3129-devils
3130-asking
3131-eliminate
3132-women
3133-hate
3134-consideration
3135-bitter
3136-hatred
3137-patient
3138-sympathetic
3139-situation
3140-understand
3141-discrimination
3142-life
3143-handing
3144-lemons
3145-lemonade
3146-cant
3147-batter
3148-supposed
3149-bake
3150-cake
3150-cake
3151-mistake
3152-fatal
3153-need
3154-overseas
3155-vacation
3156-trip
3157-broad
3158-fall
3159-retard
3160-want
3161-forget
3162-thought
3163-everything
3164-never
3165-regret
3166-little
3167-time
3168-with
3169-that
3170-Thats
3171-need
3172-because
3173-take
3174-thing
3175-same
3176-when
3177-your
3178-round
3179-believe
3180-lips
3181-ground
3182-wanna
3183-place
3184-roche
3185-gives
3186-anymore
3187-once
3188-look
3189-made
3190-offer
3191-then
3192-this
3193-picture
3194-kids
3195-head
3196-hear
3197-last
3198-said
3199-listened
3200-problems
3201-listen
3202-mine
3203-will
3204-alone
3205-again
3206-cause
3207-does
3208-happen
3209-every
3210-kinda
3211-counted
3212-being
3213-friend
3214-give
3215-away
3216-about
3217-what
3218-really
3219-know
3220-where
3221-chained
3222-myself
3223-unlocks
3224-like
3225-door
3226-some
3227-more
3228-matter
3229-they
3230-wanted
3231-someone
3232-looked
3233-just
3234-summer
3235-memory
3236-dies
3237-worked
3238-long
3239-hard
3240-sees
3241-right
3242-through
3243-easy
3244-lies
3245-cracks
3246-road
3247-would
3248-disguise
3249-runs
3250-scissor
3251-seem
3252-wall
3253-cannot
3254-break
3255-down
3256-else
3257-fall
3258-thousand
3259-lonely
3260-stars
3261-hiding
3262-cold
3263-sing
3264-doesn
3265-understand
3266-upset
3267-swimming

3964-find
3965-just
3966-yeah
3967-here
3968-fields
3969-fight
3970-meals
3971-back
3972-into
3973-living
3974-need
3975-prove
3976-right
3977-forgiven
3978-raise
3979-your
3980-only
3981-teenage
3982-wasteland
3983-sally
3984-take
3985-hand
3986-travel
3987-south
3988-cross
3989-land
3990-fire
3991-look
3992-past
3993-shoulder
3994-exodus
3995-happy
3996-ones
3997-near
3998-together
3999-before
4000-much
4001-older
4002-they
4003-wasted
4004-dancing
4005-there
4006-record
4007-machine
4008-knew
4009-must
4010-have
4011-been
4012-about
4013-seventeen
4014-beat
4015-going
4016-strong
4017-playing
4018-favorite
4019-song
4020-could
4021-tell
4022-wouldn
4023-long
4024-till
4025-with
4026-yeah
4027-would
4028-singing
4029-love
4030-rock
4031-roll
4032-another
4033-dime
4034-jukebox
4035-baby
4036-come
4037-take
4038-your
4039-time
4040-dance
4041-smiled
4042-asked
4043-name
4044-that
4045-matter
4046-said
4047-cause
4048-same
4049-home
4050-where
4051-alone
4052-next
4053-were
4054-moving
4055-there
4056-colors
4057-street
4058-white
4059-blue
4060-people
4061-shufflin
4062-their
4063-feet
4064-sleepin
4065-shoes
4066-warning
4067-sign
4068-road
4069-ahead
4070-sayin
4071-better
4072-dead
4073-feel
4074-like
4075-satan
4076-them
4077-forget
4078-keep
4079-rockin
4080-free
4081-world
4082-woman
4083-night
4084-with
4085-baby
4086-hand
4087-under
4088-light
4089-near
4090-garbage
4091-puts
4092-away
4093-gone
4094-hates
4095-life
4096-what
4097-done
4098-more
4099-that
4100-will
4101-never
4102-school
4103-fall
4104-love
4105-cool
4106-thousand
4107-points
4108-homeless
4109-kinder
4110-gentler
4111-machine
4112-department
4113-stores
4114-toilet
4115-paper
4116-styrofoam
4117-boxes
4118-ozone
4119-layer
4120-says
4121-hope
4122-alive
4123-fuel
4124-burn
4125-roads
4126-drive
4127-every
4128-time
4129-that
4130-look
4131-mirror
4132-these
4133-lines
4134-face
4135-gettin
4136-clearer
4137-past
4138-gone
4139-went
4140-like
4141-dusk
4142-dawn
4143-everybodys
4144-their
4145-dues
4146-life
4147-know
4148-nobody
4149-knows
4150-where
4151-comes
4152-goes
4153-lose
4154-half
4155-lifes
4156-books
4157-written
4158-pages
4159-lived
4160-learned
4161-from
4162-fools
4163-sages
4164-true
4165-things
4166-come
4167-back
4168-sing
4169-with
4170-year
4171-laughter
4172-tear
4173-just
4174-today
4175-maybe
4176-tomorrow
4177-good
4178-lord
4179-will
4180-take
4181-away
4182-dream
4183-yourself
4184-until
4185-your
4186-your
4187-motor
4188-runnin
4189-head
4190-highway
4191-looking
4192-adventure
4193-whatever
4194-comes
4195-yeah
4196-darlin
4197-gonna
4198-make
4199-happen

4494-round
4495-outside
4496-round
4497-outside
4498-guess
4499-whos
4500-back
4501-back
4502-again
4503-shadys
4504-back
4505-tell
4506-friend
4507-guess
4508-whoss
4509-back
4510-guess
4511-whos
4512-back
4513-guess
4514-whos
4515-back
4516-guess
4517-whos
4518-back
4519-guess
4520-whos
4521-back
4522-guess
4523-whos
4524-back
4525-guess
4526-whos
4527-back

4503-ovulating
4504-know
4505-cheney
4506-your
4507-husbands
4508-heart
4509-problem
4590-complicati
4591-wont
4592-they
4593-tried
4594-shut
4595-down
4596-feels
4597-empty
4598-without
4599-come
4600-your
4601-lips

5097-like
5098-have
5099-well
5100-truthful
5101-blueprints
5102-simply
5103-rage
5104-youthful
5105-exuberance
5106-everybody
5107-loves
5108-root
5109-nuisance
5110-earth
5111-like
5112-asteroid
5113-nothing
5114-shoot
5115-moon
5116-since
5117-taken
5118-school
5119-with
5120-this
5121-music
5122-cause
5123-vehicle
5124-rhyme
5125-lead
5126-school
5127-full
5128-students
5129-product
5130-rakim

6104-lives
6105-baby
6106-crystal
6107-ball
6108-baby
6109-does
6190-care
6191-baby
6192-having
6193-much
6194-shit
6195-kickin
6196-mother
6197-fuckin
6198-atom
6199-bomb
6200-baby
6201-fleet
6202-submarine

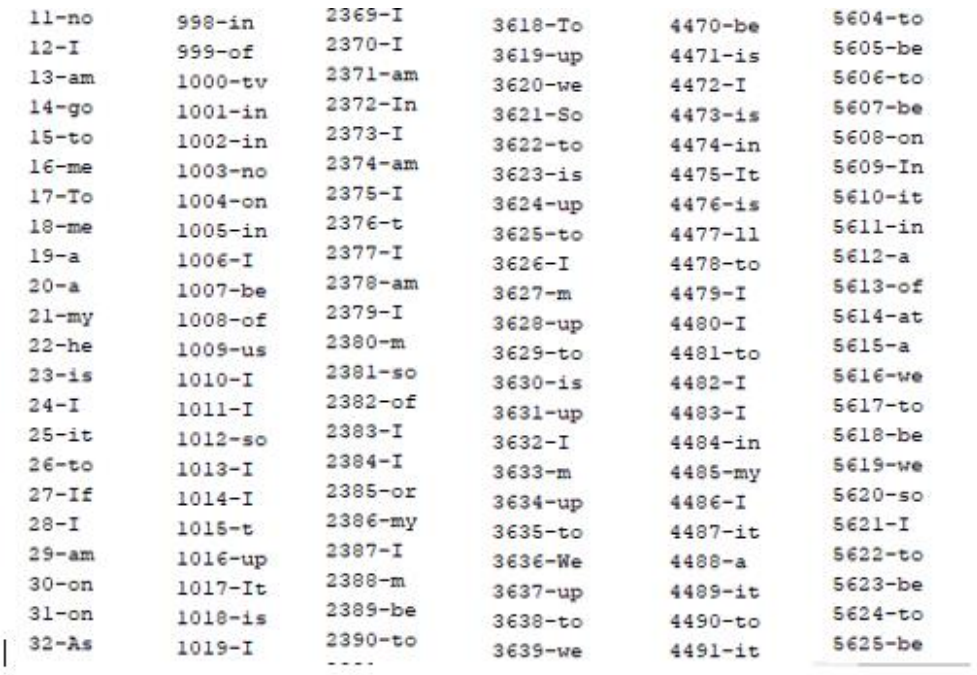
7067-read
7068-tune
7069-print
7070-scan
7071-send
7072-rename
7073-touch
7074-bring
7075-watch
7076-turn
7077-leave
7078-start
7079-format
7080-break
7081-trash
7082-change
7083-mail
7084-upgrade
7085-charge
7086-point
7087-zoom
7088-press
7089-snap
7090-work
7091-quick
7092-erase
7093-write
7094-paste
7095-save
7096-load
7097-check
7098-quit
7099-rewrite
7100-plug
7101-play
8472-always
8473-ritual
8474-always
8475-ritual
8476-come
8477-come
8478-come
8479-come
8480-come
8481-babe
8482-come
8483-come
8484-come
8485-come
8486-come
8487-babe
8488-love
8489-love
8490-love
8491-babe

9105-good
9106-lord
9107-will
9108-take
9109-away
9110-sing
9111-with
9112-sing
9113-year
9114-sing
9115-laughter
9116-sing
9117-tear
9118-sing
9119-with
9120-just
9121-today
9122-maybe
9123-tomorrow
9124-good
9125-lord
9126-will
9127-take
9128-away
9129-your
9130-motor
9131-runnin
9132-head
9133-highway
9134-locking
9135-adventure
9136-whatever
9137-comes
9138-yeah
9210-once
9211-explode
9220-into
9221-space
9222-like
9223-true
9224-natures
9225-child
9226-were
9227-born
9228-born
9229-wild
9230-climb
9231-high
9232-never
9233-wanna
9234-born
9235-wild
9236-born
9237-wild

Lista 1 9238 palabras/token´s

✓ Stopwords

Con el método **separarpalabras** también se realiza el stopwords, ignora todas las palabras del texto de entrenamiento que sean menores a 4 caracteres



11-no	998-in	2369-I	3618-To	4470-be	5604-to
12-I	999-of	2370-I	3619-up	4471-is	5605-be
13-am	1000-tv	2371-am	3620-we	4472-I	5606-to
14-go	1001-in	2372-In	3621-So	4473-is	5607-be
15-to	1002-in	2373-I	3622-to	4474-in	5608-on
16-me	1003-no	2374-am	3623-is	4475-It	5609-In
17-To	1004-on	2375-I	3624-up	4476-is	5610-it
18-me	1005-in	2376-t	3625-to	4477-ll	5611-in
19-a	1006-I	2377-I	3626-I	4478-to	5612-a
20-a	1007-be	2378-am	3627-m	4479-I	5613-of
21-my	1008-of	2379-I	3628-up	4480-I	5614-at
22-he	1009-us	2380-m	3629-to	4481-to	5615-a
23-is	1010-I	2381-so	3630-is	4482-I	5616-we
24-I	1011-I	2382-of	3631-up	4483-I	5617-to
25-it	1012-so	2383-I	3632-I	4484-in	5618-be
26-to	1013-I	2384-I	3633-m	4485-my	5619-we
27-If	1014-I	2385-or	3634-up	4486-I	5620-so
28-I	1015-t	2386-my	3635-to	4487-it	5621-I
29-am	1016-up	2387-I	3636-We	4488-a	5622-to
30-on	1017-It	2388-m	3637-up	4489-it	5623-be
31-on	1018-is	2389-be	3638-to	4490-to	5624-to
32-As	1019-I	2390-to	3639-we	4491-it	5625-be

Lista 2 Stop words

Con la función length () se obtiene la longitud del token, y un if condicional de que el token sea mayor a 3 caracteres, se agrega Apalabras a palabraL_temp con la función **add ().**

```
public ArrayList<String> separarPalabras(String texto,String catego ){
    ArrayList<String> palabrasL_temp;
      String[] Apalabras=texto.split("[, ?.@ ' !]+"); //separa la palabra de la cade
      palabrasL_temp = new ArrayList();
      for(int i=0; i<Apalabras.length;i++){
        //ingresa la palabras a la lista de palabras
        if (Apalabras[i].length() > 3 ) {// El contains()método comprueba si una cad
          ///verifica que no se repita la palabra
           palabrasL_temp.add(Apalabras[i]);
```

Fig. 3 Método separarpalabras

✓ Bag of Words

Utiliza un vector de palabras en el cual se lleva un recuento de cada una de ella y su posición en el vector; se clasifica con un switch la categoría del token´s/palabras dentro de la matriz cpcate (contador de palabras por categoría).

```
54      switch (catego.toUpperCase()) { //toUpperCase()método co
55
56          case "QUEEN":
57              cpcate.ingresarPalabra(Apalabras[i], 0 );
58
59              count[0]++;
60
61              break;
62          case "EMINEM":
63              cpcate.ingresarPalabra(Apalabras[i], 1);
64              count[1]++;
65              break;
66          case "DAFT PUNK":
67              cpcate.ingresarPalabra(Apalabras[i], 2);
68              count[2]++;
69              break;
70          case "TIESTO":
71              cpcate.ingresarPalabra(Apalabras[i], 3);
72              count[3]++;
73              break;
74          case "OTHER":
75              cpcate.ingresarPalabra(Apalabras[i], 4);
76              count[4]++;
77              break;
78          default:
```

Fig. 4 El switch convierte la categoría a mayúsculas

Esto se realiza en la clase **MatrizPalabraCantidad** con el ArreyList palabra dentro del método **ingresarPalabra**, con un if y la función **contains** verifica que la palabra/token no esté dentro de la cadena palabr, si es verdadera la agrega a la cadena palabr con la función **add** indicándole la posición o índice para agregarla, cuando la palabra ya se encuentra en la cadena palabr entonces ubica su posición con la variable dato y la función **indexOf** que devuelve la posición del token para contar la frecuencia de ocurrencia de cada palabra (token) por categoría, se utiliza para realizar la tabla de frecuencias.

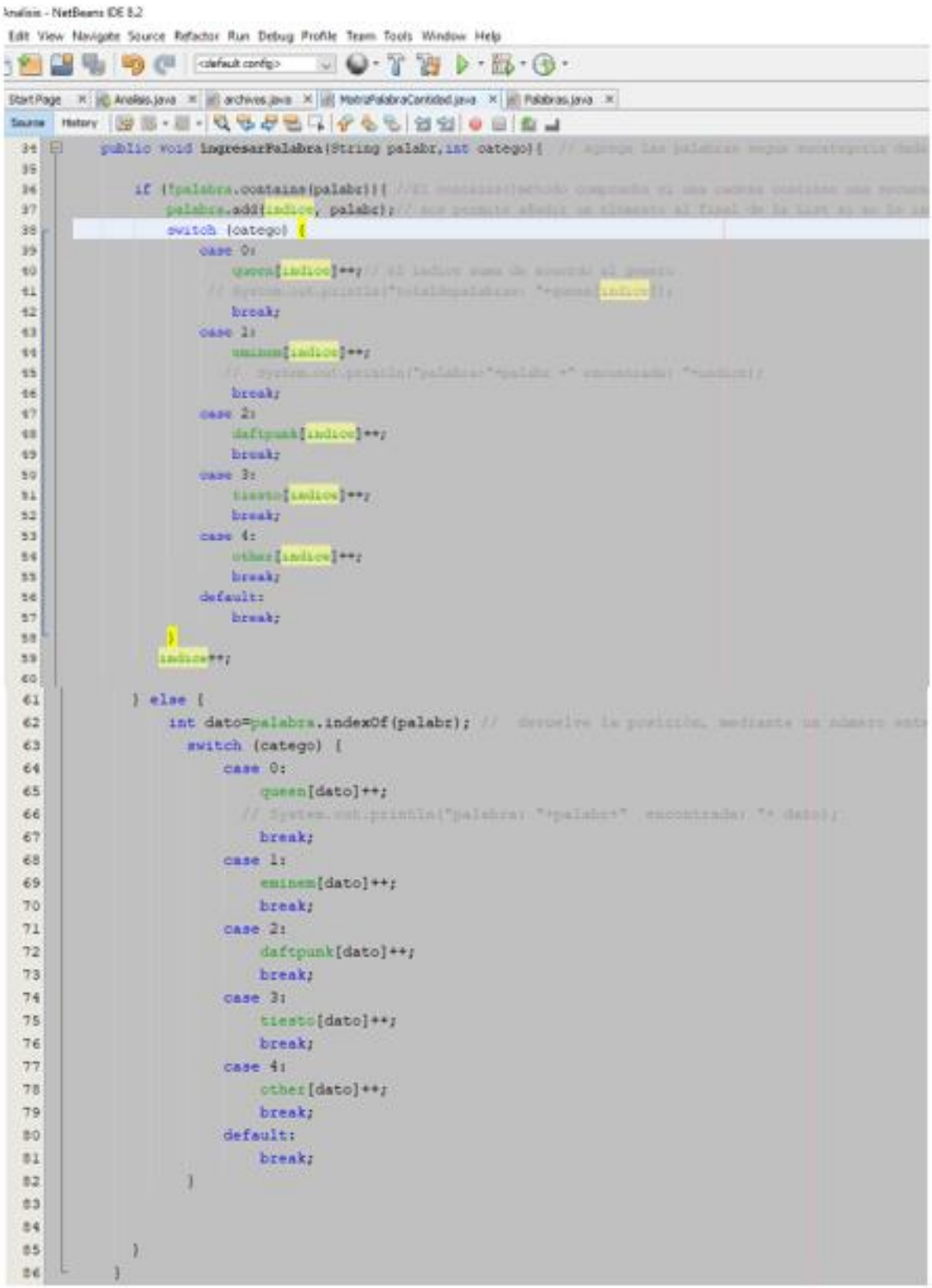


Fig. 5 Método ingresarPalabra separa la cadena en token´s y asigna un índice a cada una.

c) Análisis de los Datos

- ✓ **Tabla de frecuencias**

Con el método **ver_palabraCateCantidad** (ver palabra por categoría y cantidad) se imprimen los datos del método **ingresarPalabra** para observar la frecuencia de las palabras en cada categoría.

```
***************Tabla de frecuencias ******************************************
0-Palabra- this
 Queen 20.0  Eminem 79.0  Daft Punk 8.0   Tiesto 16.0   Other 0.0
1-Palabra- real
 Queen 3.0  Eminem 24.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
2-Palabra- life
 Queen 5.0  Eminem 6.0  Daft Punk 11.0   Tiesto 5.0   Other 2.0
3-Palabra- just
 Queen 16.0  Eminem 91.0  Daft Punk 14.0   Tiesto 17.0   Other 4.0
4-Palabra- fantasy
 Queen 1.0  Eminem 0.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
5-Palabra- caught
 Queen 1.0  Eminem 3.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
6-Palabra- landslide
 Queen 1.0  Eminem 0.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
7-Palabra- escape
 Queen 1.0  Eminem 1.0  Daft Punk 0.0   Tiesto 12.0   Other 0.0
8-Palabra- from
 Queen 7.0  Eminem 28.0  Daft Punk 1.0   Tiesto 5.0   Other 2.0
9-Palabra- reality
 Queen 1.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
10-Palabra- open
 Queen 1.0  Eminem 3.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
11-Palabra- your
 Queen 5.0  Eminem 37.0  Daft Punk 11.0   Tiesto 18.0   Other 14.0
12-Palabra- eyes
 Queen 2.0  Eminem 0.0  Daft Punk 0.0   Tiesto 2.0   Other 0.0
13-Palabra- look
 Queen 4.0  Eminem 8.0  Daft Punk 1.0   Tiesto 0.0   Other 2.0
14-Palabra- skies
 Queen 1.0  Eminem 0.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
15-Palabra- poor
 Queen 4.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
```

```
519-Palabra- asap
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
520-Palabra- mommas
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
521-Palabra- sister
 Queen 0.0  Eminem 2.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
522-Palabra- dont
 Queen 0.0  Eminem 38.0  Daft Punk 1.0   Tiesto 3.0   Other 0.0
523-Palabra- understand
 Queen 0.0  Eminem 2.0  Daft Punk 1.0   Tiesto 0.0   Other 0.0
524-Palabra- sits
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
525-Palabra- front
 Queen 0.0  Eminem 6.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
526-Palabra- buries
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
527-Palabra- nose
 Queen 0.0  Eminem 3.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
528-Palabra- colors
 Queen 0.0  Eminem 2.0  Daft Punk 0.0   Tiesto 0.0   Other 1.0
529-Palabra- crayon
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
530-Palabra- dull
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
531-Palabra- hand
 Queen 0.0  Eminem 2.0  Daft Punk 0.0   Tiesto 1.0   Other 3.0
532-Palabra- while
 Queen 0.0  Eminem 8.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
533-Palabra- brother
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
534-Palabra- mother
 Queen 0.0  Eminem 2.0  Daft Punk 2.0   Tiesto 0.0   Other 0.0
755-Palabra- wreck
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
756-Palabra- cameras
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
757-Palabra- directions
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
758-Palabra- press
 Queen 0.0  Eminem 2.0  Daft Punk 7.0   Tiesto 0.0   Other 0.0
759-Palabra- apeshit
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
760-Palabra- bananas
 Queen 0.0  Eminem 2.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
761-Palabra- networks
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
762-Palabra- commando
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
763-Palabra- extra
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
```

```
858-Palabra- reports
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
859-Palabra- number
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
860-Palabra- fatalities
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
861-Palabra- santa
 Queen 0.0  Eminem 2.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
862-Palabra- texas
 Queen 0.0  Eminem 2.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
863-Palabra- galveston
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
864-Palabra- mass
 Queen 0.0  Eminem 3.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
865-Palabra- southern
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
866-Palabra- california
 Queen 0.0  Eminem 2.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
867-Palabra- schoolwe
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
868-Palabra- following
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
869-Palabra- deadly
 Queen 0.0  Eminem 3.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
870-Palabra- houston
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
871-Palabra- twenty-six
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
872-Palabra- twenty
 Queen 0.0  Eminem 2.0  Daft Punk 1.0   Tiesto 0.0   Other 0.0
873-Palabra- other
 Queen 0.0  Eminem 11.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
992-Palabra- roof
 Queen 0.0  Eminem 11.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
993-Palabra- dogs
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
994-Palabra- caged
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
995-Palabra- playin
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
996-Palabra- beginnin
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
997-Palabra- mood
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
998-Palabra- changed
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
999-Palabra- chewed
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
1000-Palabra- booed
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
1001-Palabra- kept
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
1002-Palabra- rhymin
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
1003-Palabra- stepwritin
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
```

```
1128-Palabra- squeezing
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
1129-Palabra- neck
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
1130-Palabra- popsicle
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
1131-Palabra- stick
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
1132-Palabra- possible
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
1133-Palabra- house
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
1134-Palabra- streamed
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
1135-Palabra- cheeks
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
1136-Palabra- temple
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
1137-Palabra- wouldve
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
1138-Palabra- anything
 Queen 0.0  Eminem 2.0  Daft Punk 0.0   Tiesto 2.0   Other 0.0
1139-Palabra- adored
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
1140-Palabra- save
 Queen 0.0  Eminem 1.0  Daft Punk 7.0   Tiesto 0.0   Other 0.0
1141-Palabra- Just
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
1142-Palabra- youll
 Queen 0.0  Eminem 2.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
1143-Palabra- attention
 Queen 0.0  Eminem 4.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
1144-Palabra- Will
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
1416-Palabra- spilling
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
1417-Palabra- killing
 Queen 0.0  Eminem 3.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
1418-Palabra- pinning
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
1419-Palabra- hiphop
 Queen 0.0  Eminem 2.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
1420-Palabra- subliminal
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
1421-Palabra- thoughts
 Queen 0.0  Eminem 2.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
1422-Palabra- sending
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
1423-Palabra- webs
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
1424-Palabra- spin
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
1425-Palabra- hock
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
1426-Palabra- venom
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
```

```
1563-Palabra- disaster
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
1564-Palabra- catastrophe
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
1565-Palabra- nana
 Queen 0.0  Eminem 2.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
1566-Palabra- kshh
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
1567-Palabra- bent
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
1568-Palabra- antenna
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
1569-Palabra- tune
 Queen 0.0  Eminem 1.0  Daft Punk 6.0   Tiesto 0.0   Other 0.0
1570-Palabra- enter
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
1571-Palabra- endin
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
1572-Palabra- splinter
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
1573-Palabra- center
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
1574-Palabra- winter
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
1575-Palabra- interesting
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
1576-Palabra- wrestling
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
1577-Palabra- infesting
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
1578-Palabra- ears
 Queen 0.0  Eminem 1.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
1957-Palabra- celebrate
 Queen 0.0  Eminem 0.0  Daft Punk 72.0   Tiesto 0.0   Other 0
1958-Palabra- dancing
 Queen 0.0  Eminem 0.0  Daft Punk 16.0   Tiesto 0.0   Other 1
1959-Palabra- uhmmm
 Queen 0.0  Eminem 0.0  Daft Punk 4.0   Tiesto 0.0   Other 0.(
1960-Palabra- celebration
 Queen 0.0  Eminem 0.0  Daft Punk 4.0   Tiesto 0.0   Other 0.(
1961-Palabra- musics
 Queen 0.0  Eminem 0.0  Daft Punk 30.0   Tiesto 0.0   Other 0
1962-Palabra- often
 Queen 0.0  Eminem 0.0  Daft Punk 3.0   Tiesto 0.0   Other 0.(
1963-Palabra- speeding
 Queen 0.0  Eminem 0.0  Daft Punk 3.0   Tiesto 0.0   Other 0.(
1964-Palabra- shirt
 Queen 0.0  Eminem 0.0  Daft Punk 3.0   Tiesto 0.0   Other 0.(
1965-Palabra- ahead
 Queen 0.0  Eminem 0.0  Daft Punk 3.0   Tiesto 0.0   Other 1.(
1966-Palabra- wipe
 Queen 0.0  Eminem 0.0  Daft Punk 3.0   Tiesto 0.0   Other 0.(
1967-Palabra- dancee
 Queen 0.0  Eminem 0.0  Daft Punk 4.0   Tiesto 0.0   Other 0.(
1968-Palabra- everybodyon
 Queen 0.0  Eminem 0.0  Daft Punk 1.0   Tiesto 0.0   Other 0.(
```

```
2109-Palabra- fields
 Queen 0.0  Eminem 0.0  Daft Punk 0.0   Tiesto 0.0   Other 1.0
2110-Palabra- meals
 Queen 0.0  Eminem 0.0  Daft Punk 0.0   Tiesto 0.0   Other 1.0
2111-Palabra- forgiven
 Queen 0.0  Eminem 0.0  Daft Punk 0.0   Tiesto 0.0   Other 1.0
2112-Palabra- teenage
 Queen 0.0  Eminem 0.0  Daft Punk 0.0   Tiesto 0.0   Other 5.0
2113-Palabra- wasteland
 Queen 0.0  Eminem 0.0  Daft Punk 0.0   Tiesto 0.0   Other 5.0
2114-Palabra- sally
 Queen 0.0  Eminem 0.0  Daft Punk 0.0   Tiesto 0.0   Other 1.0
2115-Palabra- south
 Queen 0.0  Eminem 0.0  Daft Punk 0.0   Tiesto 0.0   Other 1.0
2116-Palabra- past
 Queen 0.0  Eminem 0.0  Daft Punk 0.0   Tiesto 0.0   Other 2.0
2117-Palabra- shoulder
 Queen 0.0  Eminem 0.0  Daft Punk 0.0   Tiesto 0.0   Other 1.0
2118-Palabra- exodus
 Queen 0.0  Eminem 0.0  Daft Punk 0.0   Tiesto 0.0   Other 1.0
2119-Palabra- ones
 Queen 0.0  Eminem 0.0  Daft Punk 0.0   Tiesto 0.0   Other 1.0
2120-Palabra- older
 Queen 0.0  Eminem 0.0  Daft Punk 0.0   Tiesto 0.0   Other 1.0
2121-Palabra- record
 Queen 0.0  Eminem 0.0  Daft Punk 0.0   Tiesto 0.0   Other 1.0
2122-Palabra- seventeen
 Queen 0.0  Eminem 0.0  Daft Punk 0.0   Tiesto 0.0   Other 1.0
2123-Palabra- wouldn
 Queen 0.0  Eminem 0.0  Daft Punk 0.0   Tiesto 0.0   Other 1.0
.... . . .    ..
2158-Palabra- lord
 Queen 0.0  Eminem 0.0  Daft Punk 0.0   Tiesto 0.0   Other 4.0
2159-Palabra- highway
 Queen 0.0  Eminem 0.0  Daft Punk 0.0   Tiesto 0.0   Other 2.0
2160-Palabra- adventure
 Queen 0.0  Eminem 0.0  Daft Punk 0.0   Tiesto 0.0   Other 2.0
2161-Palabra- darlin
 Queen 0.0  Eminem 0.0  Daft Punk 0.0   Tiesto 0.0   Other 3.0
2162-Palabra- embrace
 Queen 0.0  Eminem 0.0  Daft Punk 0.0   Tiesto 0.0   Other 3.0
2163-Palabra- lightnin
 Queen 0.0  Eminem 0.0  Daft Punk 0.0   Tiesto 0.0   Other 1.0
2164-Palabra- metal
 Queen 0.0  Eminem 0.0  Daft Punk 0.0   Tiesto 0.0   Other 1.0
2165-Palabra- thunder
 Queen 0.0  Eminem 0.0  Daft Punk 0.0   Tiesto 0.0   Other 1.0
2166-Palabra- natures
 Queen 0.0  Eminem 0.0  Daft Punk 0.0   Tiesto 0.0   Other 2.0
2167-Palabra- child
 Queen 0.0  Eminem 0.0  Daft Punk 0.0   Tiesto 0.0   Other 2.0
2168-Palabra- climb
 Queen 0.0  Eminem 0.0  Daft Punk 0.0   Tiesto 0.0   Other 2.0
**************************************************************************
total de palabras: 9238
```

Fig. 6 Tabla de ocurrencia de las 9238 palabras del texto de entrenamiento

Para el conteo de textos de entrenamiento se convierte la categoría a mayúsculas; con un if y la función **equals** compara que la categoría no este vacia para poder obtener la cantidad de textos incrementando la variable cantidadTextos, tambien se lleva un conteo por categoría del texto de entrenamiento y se guarda e incrementa en el vector categoria para poder calcular las probabilidades a priori de cada clase.

```
90              switch (catego.toUpperCase()) {//con
91                                              //c
92                      case "QUEEN":
93                          categoria[0]++;
94
95                          break;
96                      case "EMINEM":
97                          categoria[1]++;
98                          break;
99                      case "DAFT PUNK":
100                         categoria[2]++;
101                         break;
102                     case "TIESTO":
103                         categoria[3]++;
104                         break;
105                     case "OTHER":
106                         //revisar el toupp
107                         categoria[4]++;
108                         break;
```

Fig. 7 Conteo de textos por categoría.

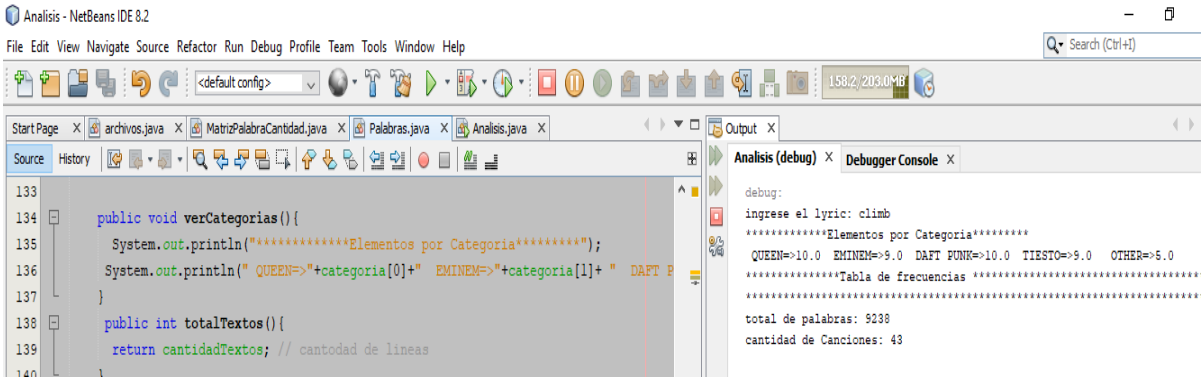


Fig. 8 En VerCategorias se imprime la cantidad de textos que tiene cada categoría, en totalTextos nos da el total de textos de entrenamiento.

d) Entrenamiento

✓ Crear Modelo

✓ Probabilidades a priori de cada clase/categoria

Una vez obtenido el total de textos de entrenamiento y categorizado se calcula la probabilidad a priori de cada clase.

$$P(y_i) = \frac{textosdecategoria}{totaldetextos} \qquad (8)$$

En el vector proCategoria(probabilidadCategoria) almacena el resultado de la división del valor categoria [i] entre cantidadtexto de acuerdo con la formula (8) y se realiza en el método **CalProbabilidadCategorias** teniendo un total de 1 entre el total de clases**,** los resultados se observan en la fig8.

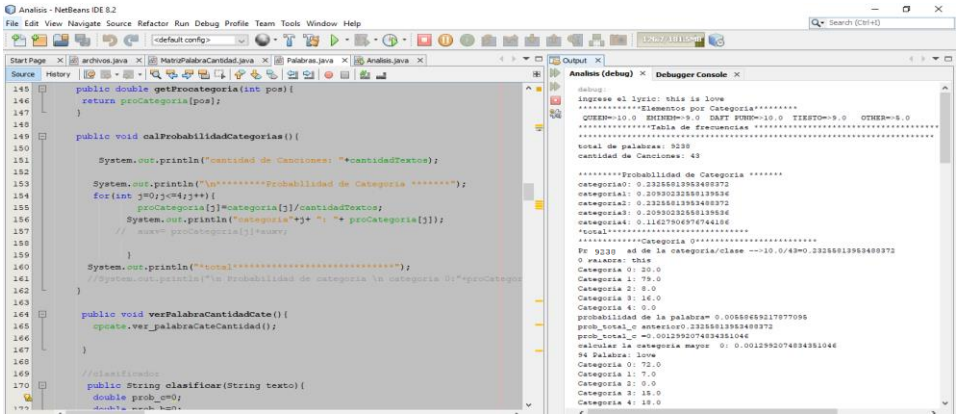


Fig. 9. Calcular probabilidad por categoría.

```
ingrese el lyric: this is love
*************Elementos por Categoria*********
 QUEEN=>10.0   EMINEM=>9.0   DAFT PUNK=>10.0   TIESTO=>9.0    OTHER=>5.0
****************Tabla de frecuencias ***************************************
*************************************************************************
total de palabras: 9238
cantidad de Canciones: 43

*********Probabllidad de Categoria *******
categoria0: 0.23255813953488372
categoria1: 0.20930232558139536
categoria2: 0.23255813953488372
categoria3: 0.20930232558139536
categoria4: 0.11627906976744186
```

Fig. 10 Probabilidad a priori de las clases

✓ Suavizado de Laplace

Para obtener la probabilidad de $P(X|y_j)$ que es la probabilidad de que la categoría y_j teniendo en cuenta la cantidad de ocurrencias de la palabra x_i entre todas las palabras de la categoría y_j, luego con esa frecuencia se divide entre el total de palabras del texto de entrenamiento.

$$P(x_i|y) = \frac{count(x_i|y_j)}{|V|} \qquad (9)$$

Un problema de la clasificación de textos es cuando una palabra no está en todas las categorías del texto de entrenamiento, la probabilidad seria cero y nos daría un error en la clasificación, para solucionarlo se utiliza el suavizado de Laplace que consiste en incrementar en una unidad la ocurrencia de las palabras en las categorías como se muestra en la formula (9).

$$count(x_i|y_j) + 1 \qquad | \qquad (10)$$

Por ejemplo, con la palabra climb que solo está en la categoría other solo estará presente en esa categoría como se muestra en la Fig11, cuando es solo una palabra no hay un mayor problema, pero cuando es una oración la probabilidad cambiaria a cero como se observa en la Fig12.

```
2168-Palabra- climb
 Queen 0.0  Eminem 0.0  Daft Punk 0.0   Tiesto 0.0   Other
**********************************************************************
```

2.

```
*************Categoria 3*************************
Probabilidad de la categoria/clase -->9.0/43=0.20930232558139536
2168 Palabra: climb
Categoria 0: 0.0
Categoria 1: 0.0
Categoria 2: 0.0
Categoria 3: 0.0
Categoria 4: 2.0
probabilidad de la palabra= 0.0
prob_total_c anterior0.20930232558139536
prob_total_c =0.0
Rango mayor : 0.0****0.0 valor de i: 3 categoria final: 0
*************Categoria 4*************************
Probabilidad de la categoria/clase -->5.0/43=0.11627906976744186
2168 Palabra: climb
Categoria 0: 0.0
Categoria 1: 0.0
Categoria 2: 0.0
Categoria 3: 0.0
Categoria 4: 2.0
probabilidad de la palabra= 7.176175098672408E-4
prob_total_c anterior0.11627906976744186
prob_total_c =8.34438964961908E-5
calcular la categoria mayor  4: 8.34438964961908E-5
Rango mayor : 8.34438964961908E-5****8.34438964961908E-5 valor de i: 4
*************Resutado*************************
palabraslocal [climb]
rango de probabilidad: 8.34438964961908E-5
 categoria final: 4
```

Fig. 11. Probabilidad de la palabra climb sin suavizado

```
*************Categoria 3*************************
Probabilidad de la categoria/clase -->9.0/43=0.20930232558139536
67 Palabra: never
Categoria 0: 9.0
Categoria 1: 17.0
Categoria 2: 30.0
Categoria 3: 3.0
Categoria 4: 5.0
probabilidad de la palabra= 9.372071227741331E-4
prob_total_c anterior0.20930232558139536
prob_total_c =1.961596303480744E-4
2168 Palabra: climb
Categoria 0: 0.0
Categoria 1: 0.0
Categoria 2: 0.0
Categoria 3: 0.0
Categoria 4: 2.0
probabilidad de la palabra= 0.0
prob_total_c anterior1.961596303480744E-4
prob_total_c =0.0
Rango mayor : 0.0****0.0 valor de i: 3 categoria final: 2
```

Fig. 12 Probabilidad de las palabras never climb sin suavizado

Para solucionar el error se incrementa una unidad en cada palabra en la Fig11 se observa que se incrementa en uno la ocurrencia del token en cada categoría y así no se pierde el valor de la probabilidad anterior.

```
*************Categoria 4*************************
Probabilidad de la categoria/clase -->5.0/43=0.11627906976744186
2129 Palabra: blue
Categoria 0: 1.0
Categoria 1: 1.0
Categoria 2: 1.0
Categoria 3: 1.0
Categoria 4: 2.0
probabilidad de la palabra= 7.176175098672408E-4
prob_total_c anterior0.11627906976744186
prob_total_c =8.34438964961908E-5
calcular la categoria mayor  4: 8.34438964961908E-5
2168 Palabra: climb
Categoria 0: 1.0
Categoria 1: 1.0
Categoria 2: 1.0
Categoria 3: 1.0
Categoria 4: 3.0
probabilidad de la palabra= 0.001076426264800861
prob_total_c anterior8.34438964961908E-5
prob_total_c =8.982120182582431E-8
Rango mayor : 8.982120182582431E-8****8.982120182582431E-8 valor de i: 4 categ
*************Resutado*************************
palabraslocal [blue, climb]
rango de probabilidad: 8.982120182582431E-8
 categoria final: 4
```

Fig. 13. Incrementa en uno las palabras blue y climb con suavizado.

✓ Normalizar los datos

Entonces para conocer el valor $P(x_i|y_j)$, se normalizan los valores de acuerdo con a la formula (11), donde se realiza el suavizado de Laplace entre la suma del total de palabras de la categoría más el vocabulario total.

$$P(x_i|y) = \frac{count(x_i|y_j) + 1}{count(y_j) + |\mathrm{V}|} \qquad (11)$$

✓ Palabras desconocidas

Para las palabras que se ingresan en los datos, pero no está en el texto de entrenamiento la solución es ignorarlas para no incluir ninguna probabilidad.

Por ejemplo, al ingresar la oración:" this rain is real" la palabra que no está en el entrenamiento es "rain" y "is" es un stop words. En la Fig14. se observa el total de elementos por categoría y la tabla de frecuencias de las palabras this y real que ocurren en cada categoría.

```
*************Elementos por Categoria*********
 QUEEN=>10.0  EMINEM=>9.0  DAFT PUNK=>10.0  TIESTO=>9.0   OTHER=>5.0
***************Tabla de frecuencias ************************************
0-Palabra- this
 Queen 20.0  Eminem 79.0  Daft Punk 8.0   Tiesto 16.0   Other 0.0
1-Palabra- real
 Queen 3.0  Eminem 24.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
```

Fig. 14. Ocurrencia de this y real por categoría.

En la Fig15. muestra la probabilidad a priori de las clases/categoría y se muestra el suavizado de las palabras this y rain se ignora con un valor de -1.

Al no estar en el entrenamiento la palabra “rain” no realiza operación y continua.

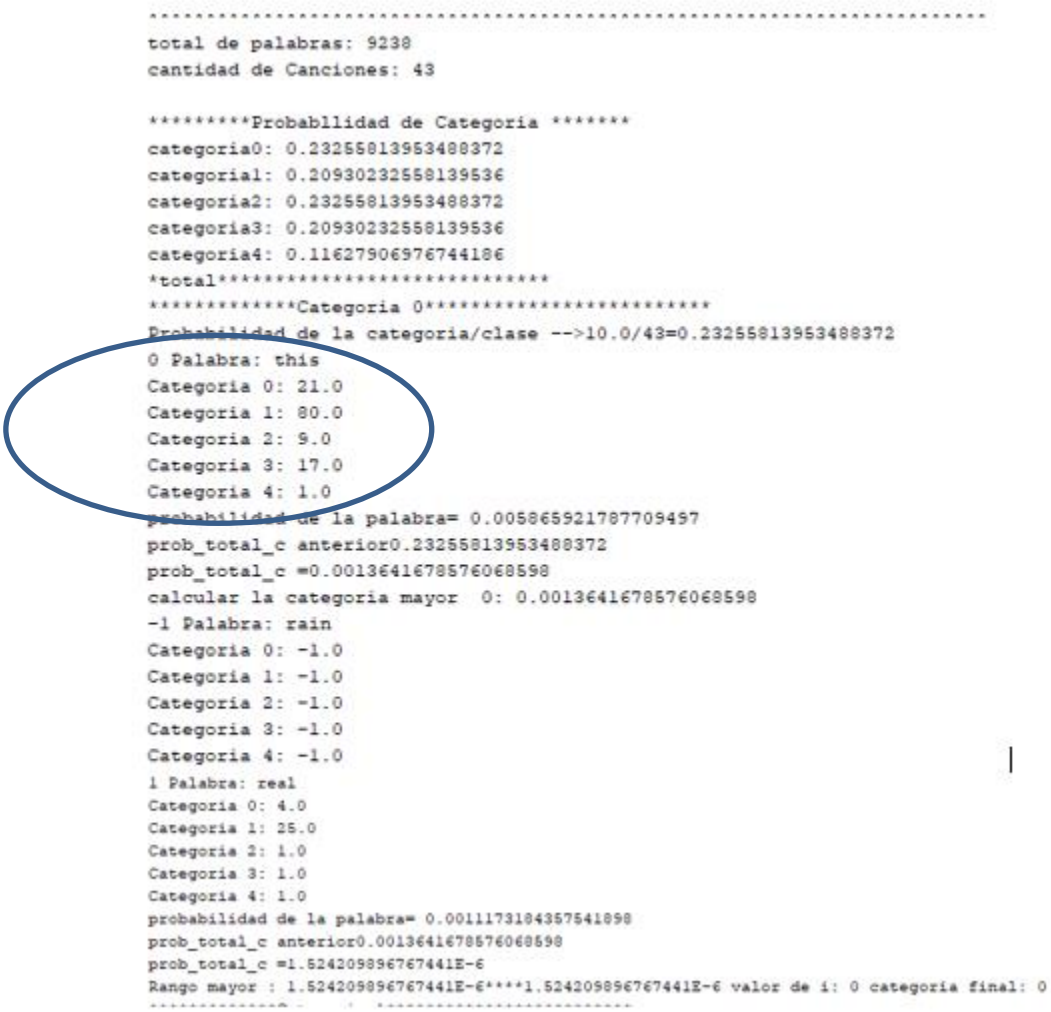


Fig. 15 Realiza el suavizado de Laplace en los token´s this y real.

Al no estar la palabra rain en el texto de entrenamiento le agrega un valor de -1 en cada categoría y no realiza operaciones y pasa a la siguiente palabra.

- ✓ **Clasificar**

- ✓ **Clasificar una oración nueva.**

Para cada categoría disponible se determinan los valores de probabilidad de cada palabra/token del nuevo ejemplo.

Siguiendo con el ejemplo anterior de la Fig15; en la Fig16. se muestra la probabilidad de las palabras por categoría de acuerdo con la formula (11).

```
*************Categoria 0************************
Probabilidad de la categoria/clase -->10.0/43=0.23255813953488372
0 Palabra: this
Categoria 0: 21.0
Categoria 1: 80.0
Categoria 2: 9.0
Categoria 3: 17.0
Categoria 4: 1.0
probabilidad de la palabra= 0.005865921787709497
*************Categoria 1************************
Probabilidad de la categoria/clase -->9.0/43=0.20930232558139536
0 Palabra: this
Categoria 0: 21.0
Categoria 1: 80.0
Categoria 2: 9.0
Categoria 3: 17.0
Categoria 4: 1.0
probabilidad de la palabra= 0.012447487163528862
*************Categoria 2************************
Probabilidad de la categoria/clase -->10.0/43=0.23255813953488372
0 Palabra: this
Categoria 0: 21.0
Categoria 1: 80.0
Categoria 2: 9.0
Categoria 3: 17.0
Categoria 4: 1.0
probabilidad de la palabra= 0.002204261572373255
*************Categoria 3************************
Probabilidad de la categoria/clase -->9.0/43=0.20930232558139536
0 Palabra: this
Categoria 0: 21.0
Categoria 1: 80.0
Categoria 2: 9.0
Categoria 3: 17.0
Categoria 4: 1.0
probabilidad de la palabra= 0.005310840362386754
*************Categoria 4************************
Probabilidad de la categoria/clase -->5.0/43=0.11627906976744186
0 Palabra: this
Categoria 0: 21.0
Categoria 1: 80.0
Categoria 2: 9.0
Categoria 3: 17.0
Categoria 4: 1.0
probabilidad de la palabra= 3.588087549336204E-4
```

Fig. 16 Probabilidades del token this en las cinco categorías.

✓ Naive Bayes

Aplicar la formula (7) del capítulo 3 Independencia Naive Bayes Multinomial para obtener la probabilidad $P(Y_j|X)$, que es la probabilidad a posteriori del token para determinar la clasificación de la categoría Y_j a la que pertenece.

$$P(x_1, x_2, \dots x_n|y)$$

$$P(x_1, x_2, \dots x_n|y) = P(x_1|y) * P(x_2|y) * P(x_3|y) * \dots * P(x_n|y)$$

$$Y_{MAP} = argmax_{y \in Y} P(x_1, x_2, \dots x_n|y) P(y)$$

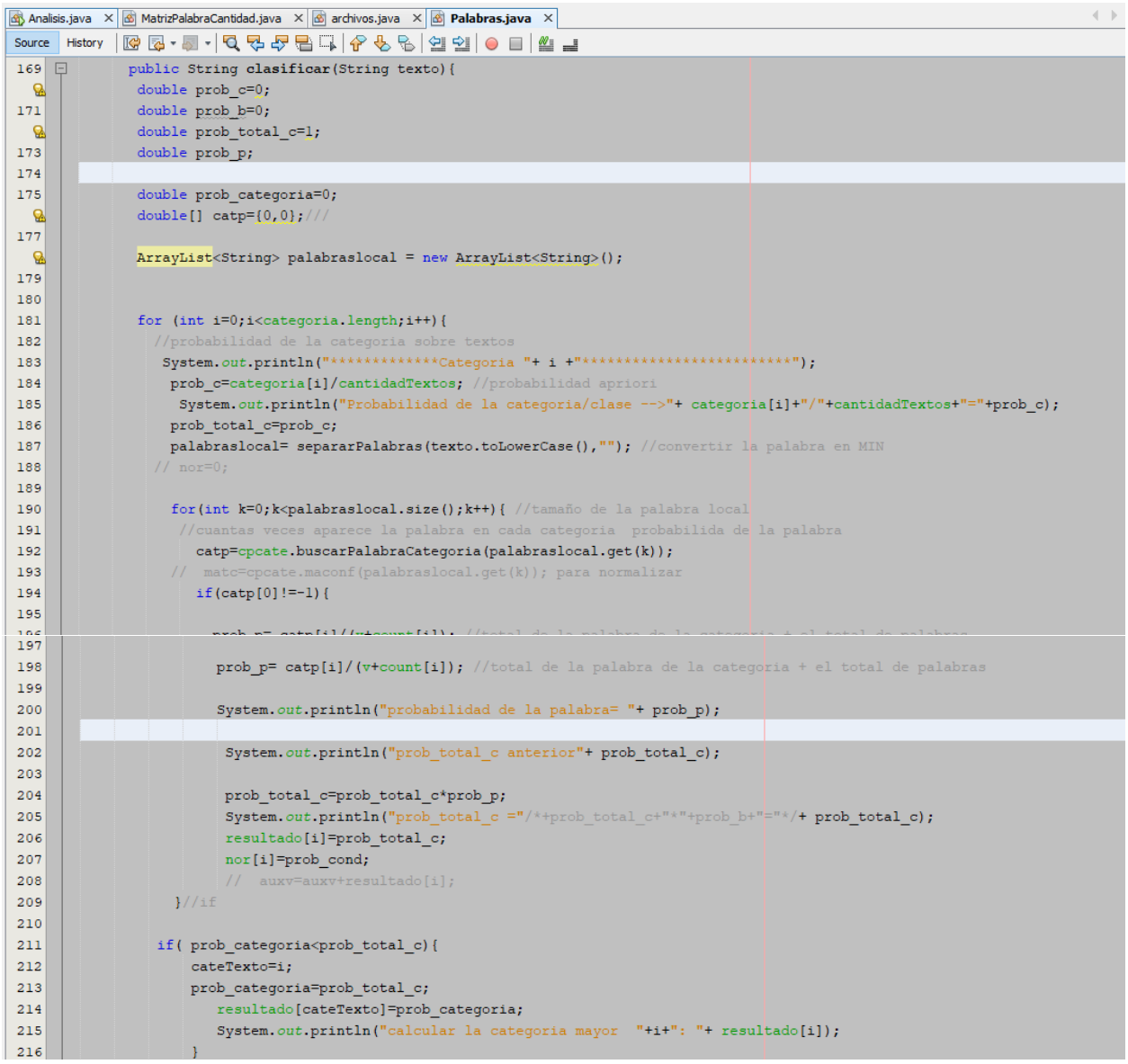
$$Y_{NB} = argmax_{y \in Y} P(y_j) \prod_{i}^{n} P(x_i|y_j) \qquad (7)$$

Siguiendo con el ejemplo, ahora después de conocer la probabilidad del primer token se aplicará la formula (7) que nos dice: multiplica la probabilidad a priori de la clase por la probabilidad del token i por categoría, hasta n-token para estimar la hipótesis más probable o MAP (máximum a posteriori hypothesis).

Esto se realiza dentro de la clase **Palabras** en el método **clasificar,** realiza las operaciones dentro de dos estructuras for; dentro del primer for se obtiene las palabras locales que es la frase ingresada al inicio, en cada categoría, el ArrayList palabraslocal, llama al método separar palabras y convierte en minúsculas las palabras de la frase ingresada.

Dentro del segundo for se realizan las operaciones de la formula (7)

Primero se debe de conocer la cantidad de ocurrencia de cada token de palabraslocal, y se realiza dentro el método buscarPalabraCategoria de la clase MatrizPalabraCantidad en base a la matriz cpcate que es donde se guarda la frecuencia del token y lo almacena dentro de arreglo catp (cantidad de palabras).




```
169     public String clasificar(String texto){
         double prob_c=0;
171      double prob_b=0;
         double prob_total_c=1;
173      double prob_p;
174
175      double prob_categoria=0;
         double[] catp={0,0};///
177
         ArrayList<String> palabraslocal = new ArrayList<String>();
179
180
181      for (int i=0;i<categoria.length;i++){
182        //probabilidad de la categoria sobre textos
183         System.out.println("*************Categoria "+ i +"*************************");
184          prob_c=categoria[i]/cantidadTextos; //probabilidad apriori
185           System.out.println("Probabilidad de la categoria/clase -->"+ categoria[i]+"/"+cantidadTextos+"="+prob_c);
186          prob_total_c=prob_c;
187          palabraslocal= separarPalabras(texto.toLowerCase(),""); //convertir la palabra en MIN
188       // nor=0;
189
190        for(int k=0;k<palabraslocal.size();k++){ //tamaño de la palabra local
191         //cuantas veces aparece la palabra en cada categoria  probabilida de la palabra
192           catp=cpcate.buscarPalabraCategoria(palabraslocal.get(k));
193       //  matc=cpcate.maconf(palabraslocal.get(k)); para normalizar
194          if(catp[0]!=-1){
195
196
197
198            prob_p= catp[i]/(v+count[i]); //total de la palabra de la categoria + el total de palabras
199
200            System.out.println("probabilidad de la palabra= "+ prob_p);
201
202             System.out.println("prob_total_c anterior"+ prob_total_c);
203
204             prob_total_c=prob_total_c*prob_p;
205             System.out.println("prob_total_c ="/*+prob_total_c+"*"+prob_b+"="*/+ prob_total_c);
206             resultado[i]=prob_total_c;
207             nor[i]=prob_cond;
208             //  auxv=auxv+resultado[i];
209        }//if
210
211      if( prob_categoria<prob_total_c){
212          cateTexto=i;
213          prob_categoria=prob_total_c;
214             resultado[cateTexto]=prob_categoria;
215             System.out.println("calcular la categoria mayor  "+i+": "+ resultado[i]);
216          }
```

Fig. 17 Método clasificar

Siguiendo con los pasos de la formula se obtiene la probabilidad a priori de la categoría prob_c, en seguida la probabilidad de la palabra prob_p para cada palabra de la frase, se multiplican ambas guardando el resultado en prob_total_c; para poder comparar los resultados y obtener la categoría que maximiza la probabilidad (argmax) se guarda el resultado de cada categoría en el arreglo resultado para finalizar dentro de un if compara la probabilidad total anterior con la actual para obtener la mayor. Como resultados se imprime las palabras que se consideraron separadas por comas, el resultado de la probabilidad mayor y la categoría.

```
242         System.out.println("************Resutado************************");
243         System.out.println("palabraslocal "+palabraslocal);
244          System.out.println("rango de probabilidad: "+resultado[cateTexto]+"\n categoria final: "+cateTexto);
245         // System.out.println("normalizar: "+nor[cateTexto]+"\n categoria final: "+cateTexto);
246          System.out.println("************************************");
```

Fig. 18: Código para la impresión de los resultados

- ✓ **Continuando con la clasificar de la oración anterior aplicando NBM.**

Probabilidad de las palabras this y real por categoría de acuerdo con la formula (11) para el token rain al no estar en el entrenamiento no se toma en cuenta.

$$P(x_i|y) = \frac{count(x_i|y_j) + 1}{count(y_j) + |V|} \qquad (11)$$

$$P(\text{this}\,|\text{categoria0}) = \frac{20+1}{1412+2168} = 0.00586592178$$

$$P(\text{real}\,|\text{categoria0}) = \frac{3+1}{1412+2168} = 0.00111731843$$

$$P(\text{this real}|\text{categoria0}) = P(\text{this}|\text{categoria0}) * P(\text{real}\,|\text{categoria0})$$

$$P(\text{categoria0}|\text{this real}) = P(\text{this}|\text{categoria0}) * P(\text{real}\,|\text{categoria0}) * P(\text{categoria0})$$

$$P(\text{categoria0}|\text{this real}) = \frac{21}{3580} * \frac{4}{3580} * \frac{10}{43} = 0.0000015242$$

Ahora hay que calcular la probabilidad de las demás categorías usando el mismo procedimiento:

$$P(\text{categoria1}|\text{this real}) = \frac{80}{6427} * \frac{25}{6427} * \frac{9}{43} = 0.00001013415$$

$$P(\text{categoria2}|\text{this real}) = \frac{9}{4083} * \frac{1}{4083} * \frac{10}{43} = 0.00000012554$$

$$P(\text{categoria3}|\text{this real}) = \frac{17}{3201} * \frac{1}{3201} * \frac{9}{43} = 0.00000034725$$

$$P(\text{categoria4}|\text{this real}) = \frac{1}{2787} * \frac{1}{2787} * \frac{5}{43} = 0.00000001497$$

Conociendo la probabilidad a posteriori de cada categoría, la probabilidad más alta es el resultado de la clasificación. En este ejemplo la categoría mayor es categoria1, lo verificaremos en las Fig18 a Fig22.

```
*************Categoria 0************************
Probabilidad de la categoria/clase -->10.0/43=0.23255813953488372
0 Palabra: this
Categoria 0: 21.0
Categoria 1: 80.0
Categoria 2: 9.0
Categoria 3: 17.0
Categoria 4: 1.0
probabilidad de la palabra= 0.005865921787709497
prob_total_c anterior0.23255813953488372
prob_total_c =0.0013641678576068598
calcular la categoria mayor  0: 0.0013641678576068598
-1 Palabra: rain
Categoria 0: -1.0
Categoria 1: -1.0
Categoria 2: -1.0
Categoria 3: -1.0
Categoria 4: -1.0
1 Palabra: real
Categoria 0: 4.0
Categoria 1: 25.0
Categoria 2: 1.0
Categoria 3: 1.0
Categoria 4: 1.0
probabilidad de la palabra= 0.0011173184357541898
prob_total_c anterior0.0013641678576068598
prob_total_c =1.524209896767441E-6
```

Fig. 19. Probabilidad de la oración ingresada para la categoría 0.

```
*************Categoria 1************************
Probabilidad de la categoria/clase -->9.0/43=0.20930232558139536
0 Palabra: this
Categoria 0: 21.0
Categoria 1: 80.0
Categoria 2: 9.0
Categoria 3: 17.0
Categoria 4: 1.0
probabilidad de la palabra= 0.012447487163528862
prob_total_c anterior0.20930232558139536
prob_total_c =0.0026052880109711575
calcular la categoria mayor  1: 0.0026052880109711575
-1 Palabra: rain
Categoria 0: -1.0
Categoria 1: -1.0
Categoria 2: -1.0
Categoria 3: -1.0
Categoria 4: -1.0
1 Palabra: real
Categoria 0: 4.0
Categoria 1: 25.0
Categoria 2: 1.0
Categoria 3: 1.0
Categoria 4: 1.0
probabilidad de la palabra= 0.003889839738602769 6
prob_total_c anterior0.0026052880109711575
prob_total_c =1.0134152835580976E-5
```

Fig. 20 Probabilidad de la oración ingresada para la categoría 1.

```
*************Categoria 2*************************
Probabilidad de la categoria/clase -->10.0/43=0.23255813953488372
0 Palabra: this
Categoria 0: 21.0
Categoria 1: 80.0
Categoria 2: 9.0
Categoria 3: 17.0
Categoria 4: 1.0
probabilidad de la palabra= 0.002204261572373255
prob_total_c anterior0.23255813953488372
prob_total_c =5.126189703193616E-4
-1 Palabra: rain
Categoria 0: -1.0
Categoria 1: -1.0
Categoria 2: -1.0
Categoria 3: -1.0
Categoria 4: -1.0
1 Palabra: real
Categoria 0: 4.0
Categoria 1: 25.0
Categoria 2: 1.0
Categoria 3: 1.0
Categoria 4: 1.0
probabilidad de la palabra= 2.449179524859172E-4
prob_total_c anterior5.126189703193616E-4
prob_total_c =1.2554958861605722E-7
```

Fig. 21 Probabilidad de la oración ingresada para la categoría 2.

```
*************Categoria 3*************************
Probabilidad de la categoria/clase -->9.0/43=0.20930232558139536
0 Palabra: this
Categoria 0: 21.0
Categoria 1: 80.0
Categoria 2: 9.0
Categoria 3: 17.0
Categoria 4: 1.0
probabilidad de la palabra= 0.005310840362386754
prob_total_c anterior0.20930232558139536
prob_total_c =0.001111571238639088
-1 Palabra: rain
Categoria 0: -1.0
Categoria 1: -1.0
Categoria 2: -1.0
Categoria 3: -1.0
Categoria 4: -1.0
1 Palabra: real
Categoria 0: 4.0
Categoria 1: 25.0
Categoria 2: 1.0
Categoria 3: 1.0
Categoria 4: 1.0
probabilidad de la palabra= 3.1240237425804435E-4
prob_total_c anterior0.001111571238639088
prob_total_c =3.4725749410780635E-7
```

Fig. 22 Probabilidad de la oración ingresada para la categoría 3.

```
*************Categoria 4*************************
Probabilidad de la categoria/clase -->5.0/43=0.11627906976744186
0 Palabra: this
Categoria 0: 21.0
Categoria 1: 80.0
Categoria 2: 9.0
Categoria 3: 17.0
Categoria 4: 1.0
probabilidad de la palabra= 3.588087549336204E-4
prob_total_c anterior0.11627906976744186
prob_total_c =4.17219482480954E-5
-1 Palabra: rain
Categoria 0: -1.0
Categoria 1: -1.0
Categoria 2: -1.0
Categoria 3: -1.0
Categoria 4: -1.0
1 Palabra: real
Categoria 0: 4.0
Categoria 1: 25.0
Categoria 2: 1.0
Categoria 3: 1.0
Categoria 4: 1.0
probabilidad de la palabra= 3.588087549336204E-4
prob_total_c anterior4.17219482480954E-5
prob_total_c =1.4970200304304056E-8
```

Fig. 23 Probabilidad de la oración ingresada para la categoría 4

En la Fig24. muestra las probabilidades de las cinco categorías y se elige la categoría con la probabilidad a posteriori mayor como resultado final.

```
*************Resutado*************************
palabraslocal [this, rain, real]
rango de probabilidad: 1.0134152835580976E-5
 categoria final: 1
**************************************
Probabilidad de la categoria  0: 1.5242098967674414E-6
total de palabras de categoria: 1412
Probabilidad de la categoria  1: 1.0134152835580976E-5
total de palabras de categoria: 4259
Probabilidad de la categoria  2: 1.2554958861605722E-7
total de palabras de categoria: 1915
Probabilidad de la categoria  3: 3.4725749410780635E-7
total de palabras de categoria: 1033
Probabilidad de la categoria  4: 1.4970200304304056E-8
total de palabras de categoria: 619
**************************************
Frase ingresada: this rain is real
 Categoria::1
 Genero:  Hip Hop
Interprete: EMINEM
```

Fig. 24. Resultado de la clasificación.

Se asigno la etiqueta de la clase que maximiza la probabilidad de la categoría y se compara con la etiqueta de la clase para evaluar el desempeño.

Ahora que se tienen los criterios para la clasificación, se aplicaran para una muestra de 144 datos (coros / frase de las canciones).

Dentro de tabla:4: Pruebas de clasificador, se observan los resultados para ambos métodos Naive Bayes Multinomial NMB y Naive Bayes NB, con ello se obtendrán las matrices de confusión.

En la siguiente tabla, **C** es igual a correcto, **I** es igual a incorrecto, **C0** categoría 0, C**1** categoría 1, **C2** categoría 2, **C3** categoría 3, **C** categoría 4, **X** es igual a incorrecto más la categoría que arrojo el clasificador y ✓ corresponde a la categoría a la que pertenece.

Prubas para el texto de entrenamiento con los metodos Naive Bayes Multinomial y Naive Bayes Simple						
Texto/frase	Categoria a la que pertence	Categoria que arroja el clasificador	NMB		NB	
			C	I	C	I
Is this the real life?	Categoria 0	**********Resutado********** palabraslocal [this, real, life] rango de probabilidad: 1.1037664516736711E-8 categoria final: 1		X C1	✓	
Open your eyes	Categoria 0	**********Resutado********** palabraslocal [open, your, eyes] rango de probabilidad: 1.1983750046742712E-10 categoria final: 1		X C1		X C1
Mama! Ooh!/Galileo	Categoria 0	rango mayor : 3.2480187008577615E-4**** **********Resutado********** palabraslocal [mama, /galileo] rango de probabilidad: 3.2480187008577615E-4 categoria final: 0	✓		✓	
Mama, Just killed a man	Categoria 0	**********Resutado********** palabraslocal [mama, just, killed] rango de probabilidad: 7.253323396712694E-10 categoria final: 1		X C1	✓	
Nothing really matters	Categoria 0	**********Resutado********** palabraslocal [nothing, really, matters] rango de probabilidad: 7.6028027872190782E-10 categoria final: 0	✓			X C1
Any way the wind blows	Categoria 0	**********Resutado********** palabraslocal [wind, blows] rango de probabilidad: 1.6330820322508298E-7 categoria final: 0	✓		✓	

I want to break free	Categoria 0	**********Resutado********** palabraslocal [want, break, free] rango de probabilidad: 9.347001046505361E-9 categoria final: 2		X C2		X C3
I've fallen in love for the first time	Categoria 0	**********Resutado********** palabraslocal [fallen, love, first, time] rango de probabilidad: 2.5838203038966885E-11 categoria final: 0	✓		✓	
Living without you by my side	Categoria 0	**********Resutado********** palabraslocal [living, without, side] rango de probabilidad: 3.0411211028879516-10 categoria final: 0	✓		✓	
The show must go on	Categoria 0	**********Resutado********** palabraslocal [show, must] rango de probabilidad: 2.6129312516013277E-6 categoria final: 0	✓			X C1
Does anybody want to take it anymore?	Categoria 0	**********Resutado********** palabraslocal [does, anybody, want, take, anymore] rango de probabilidad: 3.555253459501356-15 categoria final: 0	✓			X C2
Inside my heart is breaking	Categoria 0	**********Resutado********** palabraslocal [inside, heart, breaking] rango de probabilidad: 1.0521710600120076-10 categoria final: 1		X C1	✓	
What are we living for?	Categoria 0	**********Resutado********** palabraslocal [what, living] rango de probabilidad: 8.7059770838671052E-7 categoria final: 0	✓			X C3

I have to find the will to carry on	Categoria 0	**********Resutado********** palabraslocal [have, find, will, carry] rango de probabilidad: 4.923550076072225E-11 categoria final: 0	✓		✓	
Pressure on people, people on streets	Categoria 0	**********Resutado********** palabraslocal [pressure, people, people, streets] rango de probabilidad: 6.192673977660485E-12 categoria final: 0	✓		✓	
These are the days it never rains but it pours	Categoria 0	**********Resutado********** palabraslocal [these, days, never, rains, pours] rango de probabilidad: 5.040553154736432E-10 categoria final: 1		X C1		X C1
Why can't we give love that one more chance	Categoria 0	**********Resutado********** palabraslocal [give, love, that, more, chance] rango de probabilidad: 1.3706636935473 3E-14 categoria final: 2		X C2		X C2
Turned away from it all like a blind man	Categoria 0	**********Resutado********** palabraslocal [turned, away, from, like, blind] rango de probabilidad: 2.3247691960852047E-16 categoria final: 1		X C1		X C2
'Cause love's such an old fashioned word	Categoria 0	**********Resutado********** palabraslocal [cause, love, such, fashioned, word] rango de probabilidad: 3.2333404674463017E-15 categoria final: 0		X C1		X C1
This is our last dance	Categoria 0	Rango mayor : 3.0749348179293955E-9****4.2971611 **********Resutado********** palabraslocal [this, last, dance] rango de probabilidad: 3.0749348179293955E-9		X C2	✓	

I don't wanna stop at all	Categoria 0	**********Resutado********** palabraslocal [wanna, stop] rango de probabilidad: 3.592780470305183247E-6 categoria final: 0	✓			X C2
'Cause I'm having a good time, having a good time	Categoria 0	**********Resutado********** palabraslocal [cause, having, good, time, having, good, time] rango de probabilidad: 1.693415124369565E-18 categoria final: 0	✓			X C1
I'm a shooting star leaping through the sky	Categoria 0	**********Resutado********** palabraslocal [shooting, star, leaping, through] rango de probabilidad: 1.104036327220325E-14 categoria final: 1		X C1	✓	
That's why they call me Mister Fahrenheit	Categoria 0	**********Resutado********** palabraslocal [that, they, call, mister, fahrenhe] rango de probabilidad: 1.0265028000966514E-8 categoria final: 1		✓	✓	
I wanna make a supersonic man out of you	Categoria 0	**********Resutado********** palabraslocal [wanna, make, supersonic] rango de probabilidad: 5.550360442701869E-10 categoria final: 1		X C1		X C2
Are you ready, hey, are you ready for this?	Categoria 0	**********Resutado********** palabraslocal [ready, ready, this] rango de probabilidad: 1.5327250350555273E-8 categoria final: 0	✓		✓	

How do you think I'm going to get along Without you when you're gone	Categoria 0	**********Resutado********** palabraslocal [think, going, along] rango de probabilidad: 1.1826069125075042E-10 categoria final: 1		X C1		X C1
And another one gone, and another one gone	Categoria 0	**********Resutado********** palabraslocal [another, gone, another, gone] rango de probabilidad: 8.360109869852458-11 categoria final: 0	✓		✓	
Can anybody find me somebody to love?	Categoria 0	**********Resutado********** palabraslocal [anybody, find, somebody, love] rango de probabilidad: 2.3440417796952553E-9 categoria final: 0	✓		✓	
I have spent all my years in believing you	Categoria 0	**********Resutado********** palabraslocal [have, spent, years, believing] rango de probabilidad: 2.438845391793411E-13 categoria final: 0	✓		✓	
I work hard everyday of my life	Categoria 0	rango mayor : 1.62629632566011318-12****5.76195 **********Resutado********** palabraslocal [work, hard, everyday, life] rango de probabilidad: 1.5262963256601131E-12 categoria final: 2		X C2		X C2
They say I'm goin' crazy	Categoria 0	**********Resutado********** palabraslocal [they, goin, crazy] rango de probabilidad: 2.4440542050400427E-10 categoria final: 1		X C1		X C1
Can anybody find me somebody to love?	Categoria 0	**********Resutado********** palabraslocal [anybody, find, somebody, love] rango de probabilidad: 2.3440417796952553E-9 categoria final: 0	✓		✓	

We are the champions , my friends	Categoria 0	*************Resutado************************ palabraslocal [champions, friends] rango de probabilidad: 1.6512273881647280E-6 categoria final: 0	✓		✓	
nd we'll keep on fighting till the end	Categoria 0	*************Resutado************************ palabraslocal [keep, fighting] rango de probabilidad: 5.080699655891147E-7 categoria final: 0	✓		✓	
No time for losers	Categoria 0	*************Resutado************************ palabraslocal [time, losers] rango de probabilidad: 1.0145355913890105E-6 categoria final: 0	✓		✓	
Crazy little thing called love, yeah, yeah	Categoria 0	*************Resutado************************ palabraslocal [crazy, little, thing, called, love, yeah, yeah] rango de probabilidad: 3.552459500359704E-16 categoria final: 0	✓		✓	
This thing called love	Categoria 0	*************Resutado************************ palabraslocal [this, thing, called, love] rango de probabilidad: 8.481636221093537E-10 categoria final: 0	✓		✓	
I gotta be cool, relax, get hip!	Categoria 0	*************Resutado************************ palabraslocal [gotta, cool, relax] rango de probabilidad: 5.3219619300539156E-10 categoria final: 0	✓		✓	
This world has only one sweet moment	Categoria 0	*************Resutado************************ palabraslocal [this, world, only, sweet, moment] rango de probabilidad: 9.894584168024071E-16 categoria final: 1		x C1	✓	
Oh, when love must die?	Categoria 0	*************Resutado************************ palabraslocal [when, love, must] rango de probabilidad: 2.2200184051082042E-8 categoria final: 0	✓			X C3

But touch my tears with your lips	Categoria 0	************Resutado********* palabraslocal [touch, tears, with, your, lips] rango de probabilidad: 6.237558078036035E-16 categoria final: 1		X C1		X C3
Who wants to live forever?	Categoria 0	************Resutado********* palabraslocal [wants, live, forever] rango de probabilidad: 8.363083032941866E-9 categoria final: 0	✓		✓	
Time for me to just stand up, and travel new land	Categoria 1	************Resutado********* palabraslocal [time, just, stand, travel, land] rango de probabilidad: 9.736171423094794E-15 categoria final: 1	✓			X C0
Only way that I know how to escape from this 8 Mile Road	Categoria 1	************Resutado********* palabraslocal [only, that, know, escape, from, this, mile, road] rango de probabilidad: 6.0[illegible]E-12 categoria final: 1	✓		✓	
Sorry momma I'm grown, I must travel alone	Categoria 1	************Resutado********* palabraslocal [sorry, momma, grown, must, travel, alone] rango de probabilidad: 2.6[illegible]E-10 categoria final: 1	✓			X C2
I don't wanna be alone in the darkness	Categoria 1	************Resutado********* palabraslocal [wanna, alone, darkness] rango de probabilidad: 8.3255526640523333E-9 categoria final: 1	✓			X C2

Hello darkness, my old friend	Categoria 1	**********Resutado********** palabraslocal [hello, darkness, friend] rango de probabilidad: 3.4437513292218546E-9 categoria final: 1	✓			X C3
Yeah, we just wanted to share that with you	Categoria 1	**********Resutado********** palabraslocal [yeah, just, wanted, share, that, with] rango de probabilidad: 8.947670076484693E-17 categoria final: 1 normalizer: 0.5166356732565029	✓			X C0
you better lose yourself in the music, the moment	Categoria 1	**********Resutado********** palabraslocal [better, lose, yourself, music, moment] rango de probabilidad: 3.532820008123336E-14 categoria final: 2		X C2		X C2
the soul's escaping, through this hole that it's gaping	Categoria 1	**********Resutado********** palabraslocal [soul's, escaping, through, this, hole, that, it's, gaping] rango de probabilidad: 1.3304613071306618E-10 categoria final: 1	✓			X C0
you only get one shot, do not miss your chance to blow	Categoria 1	**********Resutado********** palabraslocal [only, shot, miss, your, chance, blow] rango de probabilidad: 4.8751505812057826E-15 categoria final: 1	✓		✓	
I'm a space bound rocket ship and your heart's the moon	Categoria 1	**********Resutado********** palabraslocal [space, bound, rocket, ship, your, heart, moon] rango de probabilidad: 1.7501436571645123E-22 categoria final: 1	✓			X C4

And I'm aiming right at you	Categoria 1	**********Resutado********** palabraslocal [aiming, right] rango de probabilidad: 5.1684179461462908E-7 categoria final: 1	✓			X C2
Two hundred fifty thousand miles on a clear night in june	Categoria 1	**********Resutado********** palabraslocal [hundred, fifty, thousand, miles, clear, night, june] rango de probabilidad: 7.870497803852261E-14 categoria final: 1	✓			X C0
I'm Slim Shady, yes I'm the real Shady	Categoria 1	**********Resutado********** palabraslocal [slim, shady, real, shady] rango de probabilidad: 1.0363220991500114E-10 categoria final: 1	✓		✓	
Please stand up, please stand up?	Categoria 1	**********Resutado********** palabraslocal [please, stand, please, stand] rango de probabilidad: 1.4547776280149493E-10 categoria final: 1	✓		✓	
And if I'm lucky, you might just give it a little kiss	Categoria 1	**********Resutado********** palabraslocal [lucky, might, just, give, little, kiss] rango de probabilidad: 3.0294605330931347E-10 categoria final: 1	✓			X C2
I'm so sick and tired of bein' admired	Categoria 1	**********Resutado********** palabraslocal [sick, tired, bein, admired] rango de probabilidad: 1.1776387490350131E-14 categoria final: 1	✓		✓	
That I	Categoria 1	**********Resutado********** palabraslocal [that, wished, that] rango de probabilidad: 7.726365161715695E-9 categoria final: 1	✓			

wished that I would just die or get fired						X C0
In the paper, the news, everyday	Categoria 1	**********Resutado********** palabraslocal [paper, news, everyday] rango de probabilidad: 6.0549473920304222E-10 categoria final: 1	✓		✓	
'Til I collapse I'm spilling these raps long as you feel 'em '	Categoria 1	**********Resutado********** palabraslocal [collapse, spilling, these, raps, long, feel] rango de probabilidad: 4.5437631513431094E-19 categoria final: 1	✓			X C2
Music is like magic there's a certain feeling you get	Categoria 1	**********Resutado********** palabraslocal [music, like, magic, there, certain, feeling] rango de probabilidad: 2.2341095525767564E-19 categoria final: 1	✓		✓	
Adrenaline shots of penicillin could not get the illin' to stop	Categoria 1	**********Resutado********** palabraslocal [adrenaline, shots, penicillin, could, illin, stop] rango de probabilidad: 5.1317795551433004E-21 categoria final: 1	✓			X C2
Guess who's back, guess	Categoria 1	**********Resutado********** palabraslocal [guess, back, guess, back] rango de probabilidad: 2.0320279285302186E-11 categoria final: 1	✓		✓	

who's back?						
Some vodka that will jump start my heart quicker	Categoria 1	**********Resutado********** palabraslocal [some, vodka, that, will, jump, start, heart, quicker] rango de probabilidad: 0.11625409942000E-15 categoria final: 1	✓			X C3
Cause we need a little controversy	Categoria 1	**********Resutado********** palabraslocal [cause, need, little, controversy] rango de probabilidad: 1.05014407715580208E-11 categoria final: 1	✓		✓	
I'm beginning to feel like a Rap God, Rap God	Categoria 1	**********Resutado********** palabraslocal [beginning, feel, like] rango de probabilidad: 5.5180322580360354E-9 categoria final: 1	✓			X C2
But for me to rap like a computer must be in my genes	Categoria 1	**********Resutado********** palabraslocal [like, computer, must, genes] rango de probabilidad: 3.0913017162169056E-13 categoria final: 1	✓			X C0
Rappers are hungry looking at me like it's lunchtime	Categoria 1	**********Resutado********** palabraslocal [rappers, hungry, looking, like, lunchtime] rango de probabilidad: 6.1620274960707603E-16 categoria final: 1	✓		✓	
I listened to your	Categoria 2	**********Resutado********** palabraslocal [listened, your, problems, listen, mine] rango de probabilidad: 1.9910630962870072E-16 categoria final: 2	✓		✓	

problems, now listen to mine						
Kinda counted on you being a friend	Categoria 2	**********Resutado********** palabraslocal [kinda, counted, being, friend] rango de probabilidad: 4.2642577146666193E-12 categoria final: 2	✓		✓	
The summer memory that just never dies	Categoria 2	**********Resutado********** palabraslocal [summer, memory, that, just, never, dies] rango de probabilidad: 3.4626837976068374E-19 categoria final: 1		X C1	✓	
We're up all night to get lucky	Categoria 2	**********Resutado********** palabraslocal [night, lucky] rango de probabilidad: 2.4649565231619232E-5 categoria final: 2	✓			X C0
What keeps the planet spinning	Categoria 2	**********Resutado********** palabraslocal [what, keeps, planet, spinning] rango de probabilidad: 1.2049707834158705E-13 categoria final: 2	✓			X C3
She's up all night for good fun, I'm up all night to get lucky	Categoria 2	**********Resutado********** palabraslocal [night, good, night, lucky] rango de probabilidad: 5.059612266278105E-10 categoria final: 2	✓			X C0
She got a kung fu star as a body guard,	Categoria 2	**********Resutado********** palabraslocal [kung, star, body, guard] rango de probabilidad: 1.3380564260176431E-14 categoria final: 2	✓		✓	

Going coat to coast on a campaign trail,	Categoria 2	**********Resutado********** palabraslocal [going, coat, coast, campaign, trail] rango de probabilidad: 9.037299235905E-10 categoria final: 2	✓		✓	
Baby' got an atom bomb,	Categoria 2	**********Resutado********** palabraslocal [baby, atom, bomb] rango de probabilidad: 1.0933101574860074E-9 categoria final: 2	✓			X C0
Do it faster, makes us stronger	Categoria 2	**********Resutado********** palabraslocal [faster, makes, stronger] rango de probabilidad: 1.6785727511652354E-8 categoria final: 2	✓		✓	
Harder, better, faster, stronger	Categoria 2	**********Resutado********** palabraslocal [harder, better, faster, stronger] rango de probabilidad: 6.988914222330722E-11 categoria final: 2	✓		✓	
Work it harder, make it better	Categoria 2	**********Resutado********** palabraslocal [work, harder, make, better] rango de probabilidad: 1.037915158815726-10 categoria final: 2	✓		✓	
One more time	Categoria 2	**********Resutado********** palabraslocal [more, time] rango de probabilidad: 6.467198809156014E-5 categoria final: 2	✓		✓	
Music's got me feeling so free	Categoria 2	**********Resutado********** palabraslocal [music, feeling, free] rango de probabilidad: 1.0516277077318533E-7 categoria final: 2	✓			X C1

You know we're gonna do it right	Categoria 2	**********Resutado********** palabraslocal [know, gonna, right] rango de probabilidad: 5.301135931376441E-8 categoria final: 2	✓		✓	
Celebrate and dance so free	Categoria 2	**********Resutado********** palabraslocal [celebrate, dance, free] rango de probabilidad: 7.108224320780118E-7 categoria final: 2	✓		✓	
Lose yourself to dance	Categoria 2	**********Resutado********** palabraslocal [lose, yourself, dance] rango de probabilidad: 1.1548088538445949E-7 categoria final: 2	✓		✓	
I know you don't get chance to take a break this often	Categoria 2	**********Resutado********** palabraslocal [know, chance, take, break, this, often] rango de probabilidad: 9.626727799583189E-10 categoria final: 1		X C1	✓	
You take my shirt and just go ahead and wipe up all the sweat	Categoria 2	**********Resutado********** palabraslocal [take, shirt, just, ahead, wipe, sweat] rango de probabilidad: 4.810662373737753E-10 categoria final: 2	✓			X C0
technologic. technologic. technologic. technologic.	Categoria 2	**********Resutado********** palabraslocal [technologic, technologic, technologic, technologic] rango de probabilidad: 1.0546737057429048E-12 categoria final: 3	✓		✓	

View it, code it, jam - unlock it,	Categoria 2	**********Resutado********** palabraslocal [view, code, unlock] rango de probabilidad: 5.1248913632156 59E-10 categoria final: 2	✓		✓	
Touch it, bring it, pay it, watch it	Categoria 2	**********Resutado********** palabraslocal [touch, bring, watch] rango de probabilidad: 1.1710910250553141E-9 categoria final: 2	✓			X C3
Just to leave this unresolved	Categoria 2	**********Resutado********** palabraslocal [just, leave, this, unresolved] rango de probabilidad: 3.611425497040707E-12 categoria final: 1		X C1	✓	
When you finally get involved face to face	Categoria 2	**********Resutado********** palabraslocal [when, finally, involved, face, face] rango de probabilidad: 7.0[illegible]E-16 categoria final: 2	✓			X C3
All because I hoped that you'd be someone different	Categoria 2	**********Resutado********** palabraslocal [because, hoped, that, someone, different] rango de probabilidad: 2.1377188017336958E-17 categoria final: 1		X C1	✓	
It's not hard to go the distance	Categoria 2	**********Resutado********** palabraslocal [hard, distance] rango de probabilidad: 1.6739945148807626E-7 categoria final: 2	✓		✓	
Let the music in	Categoria 2	**********Resutado********** palabraslocal [music, tonight, just, turn, music] rango de probabilidad: 5.533480020241912E-14 categoria final: 2	✓			X

tonight, just turn on the music						C1
Music	Categoria 2	**********Resutado********** palabraslocal [music] rango de probabilidad: 8.643649505322694E-4 categoria final: 2	✓			X C1
These fragments of time	Categoria 2	**********Resutado********** palabraslocal [these, fragments, time] rango de probabilidad: 1.0089462011512043E-6 categoria final: 1		X C1		X C1
Turning our days into melodies	Categoria 2	**********Resutado********** palabraslocal [turning, days, into, melodies] rango de probabilidad: 7.657690750930179E-11 categoria final: 3		X C3		X C0
And it's crystal clear	Categoria 2	**********Resutado********** palabraslocal [crystal, clear] rango de probabilidad: 5.679901716269209 4E-8 categoria final: 2	✓		✓	
Now put your hands in the air for the century!	Categoria 3	**********Resutado********** palabraslocal [your, hands, century] rango de probabilidad: 5.941097947596666E-9 categoria final: 3	✓		✓	
Hello..if you hear me	Categoria 3	**********Resutado********** palabraslocal [hello, hear] rango de probabilidad: 3.757633041444344E-6 categoria final: 3	✓		✓	

can you hear me?	Categoria 3	**********Resutado********** palabraslocal [hear] rango de probabilidad: 5.884708910442231E-4 categoria final: 3	✓		✓	
Escape me	Categoria 3	**********Resutado********** palabraslocal [escape] rango de probabilidad: 8.500250648416557E-4 categoria final: 3	✓			X C0
Forget shout friends	Categoria 3	**********Resutado********** palabraslocal [forget, shout, friends] rango de probabilidad: 4.085302203621251E-7 categoria final: 3	✓		✓	
don't shred me down to strips	Categoria 3	**********Resutado********** palabraslocal [shred, down, strips] rango de probabilidad: 1.5315397501860359E-10 categoria final: 3	✓		✓	
you're way to good at it	Categoria 3	**********Resutado********** palabraslocal [good] rango de probabilidad: 0.0010393659067400036 categoria final: 0		X C0		X C0
I feel you in my bones	Categoria 3	**********Resutado********** palabraslocal [feel, bones] rango de probabilidad: 1.0724128494505704E-5 categoria final: 3	✓			X C1
Take my hand up again	Categoria 3	**********Resutado********** palabraslocal [take, hand, again] rango de probabilidad: 1.9067796130126077E-10 categoria final: 1		X C1		X C0
You're	Categoria 3	**********Resutado********** palabraslocal [knocking, windows] rango de probabilidad: 6.1280734254310708E-7 categoria final: 3	✓		✓	

knocking at my windows						
For as long as you are, here on this earth	Categoria 3	**********Resutado********** palabraslocal [long, here, this, earth] rango de probabilidad: 3.6601934040849477E-11 categoria final: 3	✓			X C0
Feel alive, feel alive, feel alive, feel alive	Categoria 3	**********Resutado********** palabraslocal [feel, alive, feel, alive, feel, alive, feel, alive] rango de probabilidad: 1.0546462115522605E-15 categoria final: 3	✓		✓	
I feel alive	Categoria 3	**********Resutado********** palabraslocal [feel, alive] rango de probabilidad: 1.0009106594872066E-5 categoria final: 3	✓			X C1
Make it so, it shakes your heart	Categoria 3	**********Resutado********** palabraslocal [make, shakes, your, heart] rango de probabilidad: 1.193340599022146SE-12 categoria final: 1		X C1		X C0
It's a not the things you say	Categoria 3	**********Resutado********** palabraslocal [things] rango de probabilidad: 0.0013731174124365206 categoria final: 3	✓			X C0
In my mind I can feel you coming	Categoria 3	**********Resutado********** palabraslocal [mind, feel, coming] rango de probabilidad: 1.2060075532715034E-8 categoria final: 3	✓		✓	
I'm still waiting	Categoria 3	**********Resutado********** palabraslocal [still, waiting] rango de probabilidad: 1.3174339068625527E-7 categoria final: 1		X C1		X C0

I will be here	Categoria 3	**********Resutado********** palabraslocal [will, here] rango de probabilidad: 2.022264230352515E-6 categoria final: 3	✓		✓	
When the big world falls apart	Categoria 3	**********Resutado********** palabraslocal [when, world, falls, apart] rango de probabilidad: 6.977492874168396E-12 categoria final: 3	✓		✓	
You don't mind if life's not that pretty	Categoria 3	**********Resutado********** palabraslocal [mind, life, that, pretty] rango de probabilidad: 6.4591648320873138E-12 categoria final: 3	✓		✓	
We could just run them, red lights	Categoria 3	**********Resutado********** palabraslocal [could, just, them, lights] rango de probabilidad: 9.143306662310266E-11 categoria final: 3	✓			X C1
Don't ever turn around	Categoria 3	**********Resutado********** palabraslocal [donDt, ever, turn, around] rango de probabilidad: 1.2299739271717602E-9 categoria final: 2	✓			X C0
There ain't no reason to stay	Categoria 3	**********Resutado********** palabraslocal [there, ainDt, reason, stay] rango de probabilidad: 1.8378477002232431E-9 categoria final: 3	✓			X C0
Oh, you'll always be my ritual	Categoria 3	**********Resutado********** palabraslocal [always, ritual] rango de probabilidad: 6.577465476630215E-6 categoria final: 3	✓		✓	
You know what I	Categoria 3	**********Resutado********** palabraslocal [know, what, want, meet, dare] rango de probabilidad: 8.657761147021062E-16 categoria final: 1		X		X

want, now meet me if you dare				C1		C2
Be there when the sun is rising	Categoria 3	**********Resutado********** palabraslocal [there, when, rising] rango de probabilidad: 1.2762831251550302E-8 categoria final: 3	✓			X C0
You are my diamond when I'm with you I will shine	Categoria 3	**********Resutado********** palabraslocal [diamond, when, with, will, shine] rango de probabilidad: 5.5491106804073S2E-13 categoria final: 3	✓		✓	
All my life I been looking for love	Categoria 3	**********Resutado********** palabraslocal [life, been, looking, love] rango de probabilidad: 9.0234699980200403E-13 categoria final: 1		X C1		X C0
Teenage wasteland	Categoria 4	**********Resutado********** palabraslocal [teenage, wasteland] rango de probabilidad: 5.3892721095494S8E-7 categoria final: 4	✓		✓	
The happy ones are near	Categoria 4	**********Resutado********** palabraslocal [happy, ones, near] rango de probabilidad: 1.9144246877325448E-11 categoria final: 3		X C3		X C0
I get my back into my living	Categoria 4	**********Resutado********** palabraslocal [back, into, living] rango de probabilidad: 4.0839358711925S3E-10 categoria final: 1		X C1		X C1
I love rock and roll	Categoria 4	**********Resutado********** palabraslocal [love, rock, roll] rango de probabilidad: 1.4800122700721359E-9 categoria final: 0		X C0		X C0

Sing for the year	Categoria 4	**********Resutado********** palabraslocal [sing, year] rango de probabilidad: 1.5710710319519255E-6 categoria final: 4	✓			X C2
Born to be wild	Categoria 4	**********Resutado********** palabraslocal [born, wild] rango de probabilidad: 9.4812261917115546E-7 categoria final: 4	✓			X C0
Yeah, darlin', gonna make it happen	Categoria 4	**********Resutado********** palabraslocal [yeah, darlin, gonna, make, happen] rango de probabilidad: 2.1949473920292917E-14 categoria final: 2	✓			X C0
We were born, born to be wild	Categoria 4	**********Resutado********** palabraslocal [were, born, born, wild] rango de probabilidad: 5.463950259438169E-12 categoria final: 4	✓			X C2

Tabla 4 Pruebas del clasificador

Se puede observar que las categorías en las cuales tuvo mejor desempeño son las categorías que mayor cantidad de palabras aportan al vocabulario, asignando la mayoría de las oraciones de prueba a la categoria1(Eminen) que cuenta con un total de 4259 palabras, como el clasificador selecciona la probabilidad de pertenencia mayor, asigna las oraciones a esta categoría.

5.2 Aprendizaje Supervisado

Al realizar la tabla 4, se obtuvieron datos para guiar el proceso de aprendizaje con algunas oraciones, la mayoría son coros de las canciones, con base a las pruebas realizadas, clasifica algunas oraciones/frases ingresadas con las canciones y categoría que pertenecen.

Por lo tanto, la Arquitectura del sistema que se había mencionado en el capítulo 4.4 cambia, quedando como se nuestra en la Fig.25.

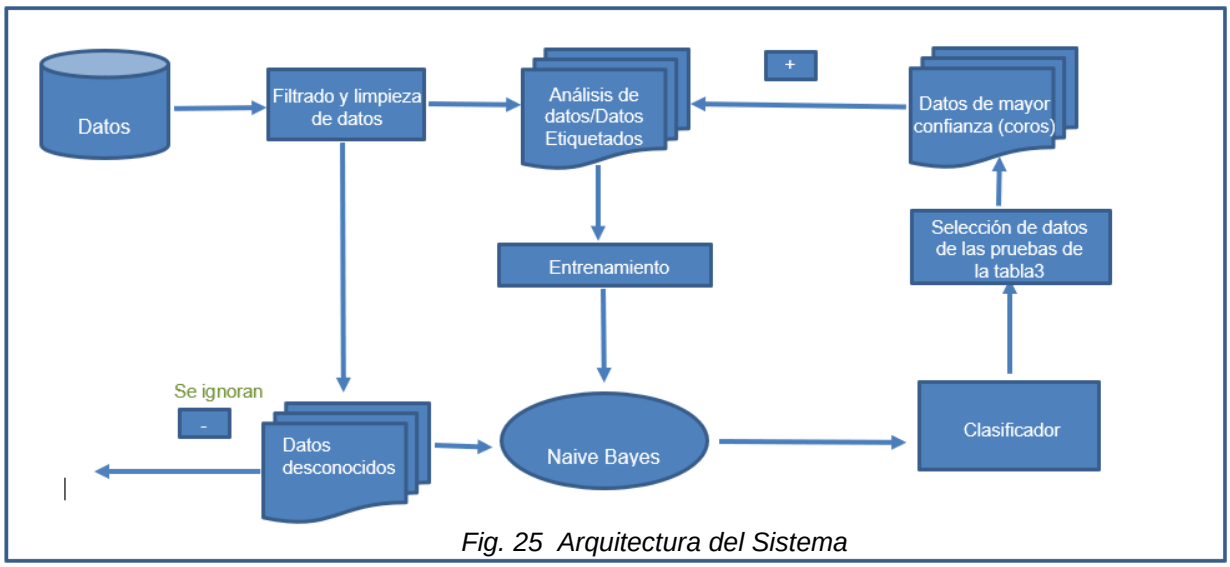


Fig. 25 Arquitectura del Sistema

De acuerdo con la tabla4 se agregan más datos para aprendizaje supervisado, pero ahora clasifica algunas canciones ya no solo el género al que pertenecen. Esto se realiza dentro de la clase Palabras en el método clasificar a partir de la línea 262 con los datos obtenidos en la tabla4, como se muestra en la Fig26.

Dentro de la clasificación se llama a el método leer, de la clase Archivo donde se leen las líneas del archivo. De acuerdo con la clasificación, se asigna el archivo a imprimir. Se puede ver un ejemplo en la línea 274 de la Fig26

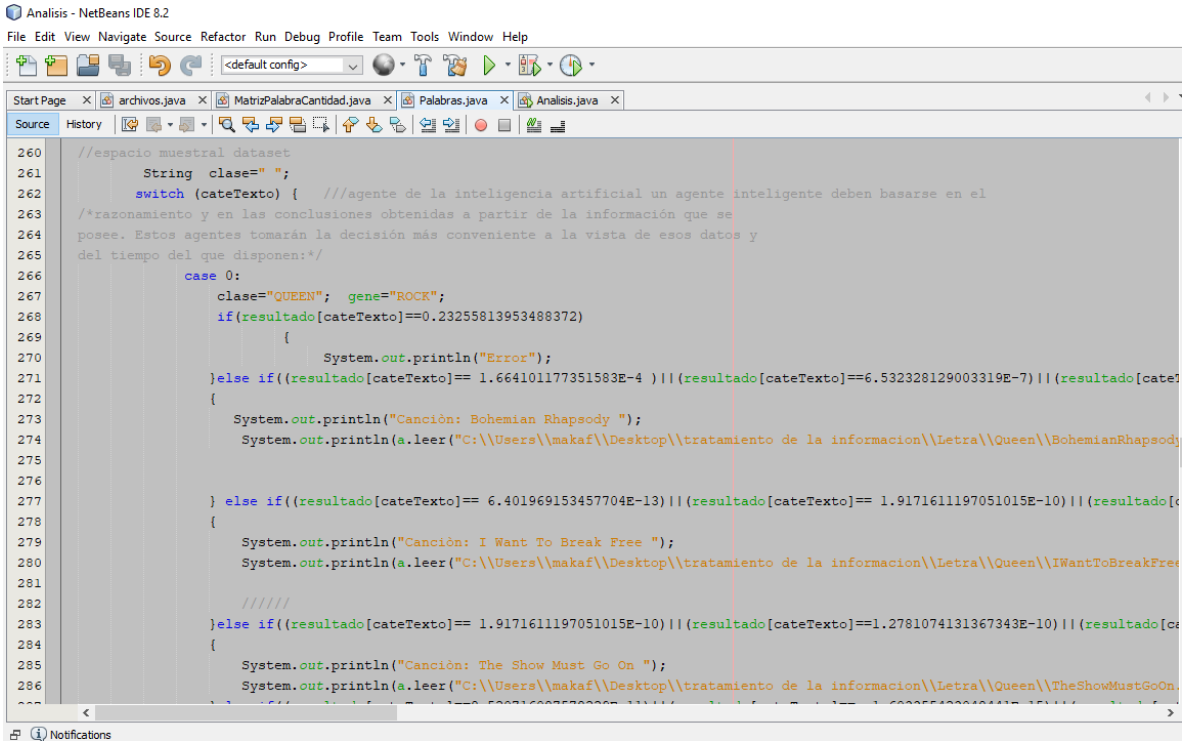


Fig. 26. Dentro de los if anidados se ingresaron los valores verdaderos positivos de la tabla 3.

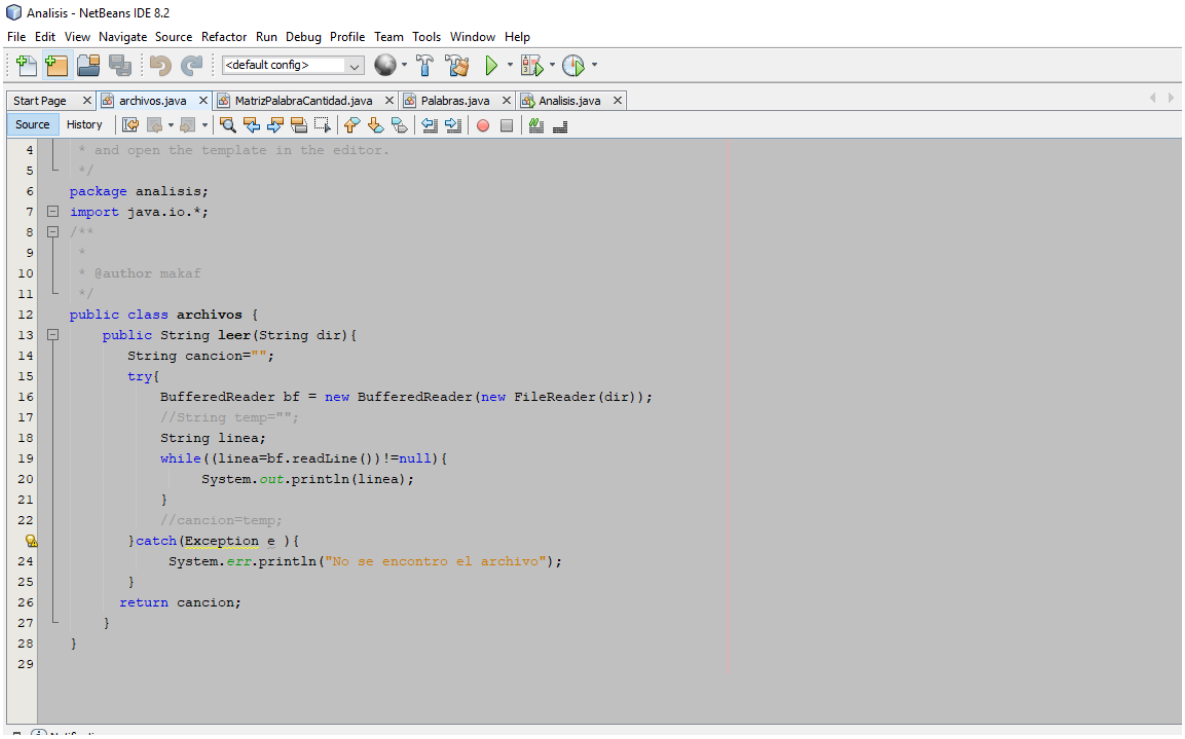


Fig. 27 Clase Archivo, método leer imprime las líneas del dataset.

Capítulo 6: Evaluación del desempeño del clasificador

Para observar el funcionamiento de la arquitectura de sistema Fig25, que se obtiene de la tabla4: Pruebas del clasificador, se ingresara el coro:" You don't mind if life's not that pretty" de la categoría:3 que corresponde a la canción I will be here, by Tiesto, se observa el resultado de la clasificación dentro del clasificador.

1. Filtrado y limpieza de datos: Las palabras menores de 4 caracteres son ignoradas y solo se consideran: "mind life that pretty" de la oración ingresada.
2. Análisis de Datos: aquí se encuentra BoW/ Tabla de frecuencias

```
debug:
ingrese el lyrics: You don't mind if life's not that pretty
**************Elementos por Categoria*********
 QUEEN=>10.0  EMINEM=>9.0  DAFT PUNK=>10.0  TIESTO=>9.0   OTHER=>5.0
****************Tabla de frecuencias **********************************
0-Palabra- this
 Queen 20.0  Eminem 79.0  Daft Punk 8.0   Tiesto 16.0   Other 0.0
1-Palabra- real
 Queen 3.0  Eminem 24.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
2-Palabra- life
 Queen 5.0  Eminem 6.0  Daft Punk 11.0   Tiesto 5.0   Other 2.0

117-Palabra- walk
 Queen 1.0  Eminem 6.0  Daft Punk 0.0   Tiesto 0.0   Other 0.0
118-Palabra- that
 Queen 10.0  Eminem 69.0  Daft Punk 9.0   Tiesto 11.0   Other 6.0

 1288-Palabra- mind
  Queen 0.0  Eminem 2.0  Daft Punk 0.0   Tiesto 8.0   Other 0.0
 1289-Palabra- song
 2042-Palabra- pretty
  Queen 0.0  Eminem 0.0  Daft Punk 0.0   Tiesto 4.0   Other 0.0
```

Fig. 28 Frecuencia del token´s de la oración

3. Entrenamiento: se calcula la probabilidad para los token´s para las cinco categorías

```
*************Categoria 3*************************
Probabilidad de la categoria/clase -->9.0/43=0.20930232558139536
1288 Palabra: mind
Categoria 0: 1.0
Categoria 1: 3.0
Categoria 2: 1.0
Categoria 3: 9.0
Categoria 4: 1.0
probabilidad de la palabra= 0.0028116213683223993
prob_total_c anterior0.20930232558139536
prob_total_c =5.884788910442231E-4
calcular la categoria mayor  3: 5.884788910442231E-4
2 Palabra: life
Categoria 0: 6.0
Categoria 1: 7.0
Categoria 2: 12.0
Categoria 3: 6.0
Categoria 4: 3.0
probabilidad de la palabra= 0.0018744142455482662
prob_total_c anterior5.884788910442231E-4
prob_total_c =1.1030532165777379E-6
118 Palabra: that
Categoria 0: 11.0
Categoria 1: 70.0
Categoria 2: 10.0
Categoria 3: 12.0
Categoria 4: 7.0
probabilidad de la palabra= 0.0037488284910965324
prob_total_c anterior1.1030532165777379E-6
prob_total_c =4.1351573255022976E-9
2042 Palabra: pretty
Categoria 0: 1.0
Categoria 1: 1.0
Categoria 2: 1.0
```

Fig. 29. Probabilidad de los token´s dentro de la categoría 3.

4. Naive Bayes se calcula la probabilidad de la categoría y de los token´s y se van multiplicando como lo dice la fórmula para cada categoría.

```
Categoria 3: 5.0
Categoria 4: 1.0
probabilidad de la palabra= 0.001562011871290222
prob_total_c anterior4.1351573255022976E-9
prob_total_c =6.4591648320873l3E-12
Rango mayor : 6.4591648320873l3E-12****6.4591648320873l3E-12 valor de i: 3 categoria final: 3
```

Fig. 30 Probabilidad de la categoría 3

5. Clasificación: se comparan los resultados obtenidos y se obtener la categoría que maximiza la probabilidad.

```
************Resutado**************************
palabraslocal [mind, life, that, pretty]
rango de probabilidad: 6.459164832087313E-12
 categoria final: 3
**************************************
Probabilidad de la categoria  0: 9.344226852449011E-14
total de palabras de categoria: 1412
Probabilidad de la categoria  1: 1.8032593344598638E-13
total de palabras de categoria: 4259
Probabilidad de la categoria  2: 1.0041423195132321E-13
total de palabras de categoria: 1915
Probabilidad de la categoria  3: 6.459164832087313E-12
total de palabras de categoria: 1033
Probabilidad de la categoria  4: 4.047370562546792E-14
total de palabras de categoria: 619
**************************************
```

Fig. 31 Resultado categoría final 3

6. Selección de datos: las probabilidades se incluyen en un switch donde se clasifican las canciones para imprimir la letra.

```
}else if((resultado[cateTexto]== 1)||(resultado[cateTexto]==2.022264230392519E-5 )||(resultado[cateTexto
{
    System.out.println("Canción:I Will Be Here");
     System.out.println(a.leer("C:\\Users\\makaf\\Desktop\\tratamiento de la informacion\\Letra\\Tiesto'
```

Fig. 32 Probabilidades para la canción

7. Datos de mayor confianza: con estas probabilidades se imprimen: la letra de la canción, nombre, categoría e interprete.

   En este ejemplo se muestra la clasificación de la oración: You don´t mind if life´s not that pretty de la tabla4: Pruebas del clasificador.

```
Frase ingresada: You don't mind if life's not that pretty
 Categoria::3
 Genero: Electropop
Interprete: TIESTO
```

Canción:I Will Be Here
I don't know, what went wrong
If I did, would it matter cause
It just wasn't enough
You know when the moment comes
To be strong, to resistance
And that is what
We're lead to believe

When the big world falls apart
And you think that the feeling will linger
You need somewhere to start
I will be here

And when it all seems to fall apart
You can't breathe
You don't know what you're thinking
You need somewhere to start
I will be here

You don't mind if life's not that pretty
It will soon disappear
It will be miles away
Away from here

Guess that things didn't work out
It will soon disappear and will be miles away
Away from here
You don't mind if life's not that pretty
It will soon disappear and will be miles away
Away from here

When the big world falls apart
And you think that the feeling will linger
You need somewhere to start
I will be here

And when it all seems to fall apart
You can't breathe
You don't know what you're thinking

Fig. 33 .Letra de la canción: I Will Be Hero by Tiesto

6.1 Matriz de confusión

En la matriz de confusión se analizan los resultados, mostrando los aciertos y errores para cada categoría con base a cuatro conceptos básicos: verdadero y falso positivo, verdaderos y falsos negativos.

VP	**FP**
FN	**VN**

Tabla 5 Matriz de confusión

Donde:

1. Verdadero Positivo [VP]: Los datos son positivos y en la prueba dio positivo
2. Verdadero Negativo [VN]: Los datos son negativos y en la prueba dio negativo
3. Falso Negativo [FN]: Los datos son positivos y en la prueba dio negativo
4. Falso Positivo [FP]: los datos son negativos y la prueba dio positiva

6.2 Métrica para la matriz de confusión

Accuracy: La exactitud mide el rendimiento de un sistema de IA

$$\boldsymbol{accuracy} = \frac{numerodepredicciones correctas}{totaldepredicciones} \quad (12)$$

Para la exactitud se considera como excelente los rangos mayores o igual a 0.9, como buena los rangos mayores o igual a 0.7 y menor a 0.7 tienen una mala exactitud.

Dos métricas mejores para evaluar clases desequilibradas son: precisión y recall.

Precision (Precisión): es la probabilidad de que la canción etiquetada en la categoría i corresponda realmente a esa categoría

$$\textbf{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}} \tag{13}$$

Recall (Sensibilidad): es la probabilidad de que la canción que pertenece a la clase i es etiquetada dentro de esa clase

$$\boldsymbol{Recall} = \frac{true\ positive}{true\ positive + Falso\ negativo} \tag{14}$$

Puntaje F1: Resume la precisión y la sensibilidad en una sola métrica.

$$PuntajeF1 = \frac{2 * Precision * Sensibilidad}{Precision + Sensibilidad} \tag{15}$$

Teniendo en cuenta que el valor bajo es menor o igual a 0.5 y los valores altos es mayor o igual a 0.7 para precisión y recall se realiza la evaluación como lo marca la tabla 5.

Valores para Precision y Recall	Evaluación
Alta precisión y alto recall	El clasificador detecta correctamente la categoría.
Alta precisión y bajo recall	El clasificador no detecta la categoría muy bien, pero cuando lo hace es realmente confiable.
Bajo precisión y alto recall	El clasificador detecta bien la categoría, pero también incluye muestras de otras categorías.
Bajo precisión y bajo recall	El clasificador no detecta las categorías correctamente.

Tabla 6 Casos posibles

6.3 Matriz de confusión del clasificador

Para evaluar al clasificador se realizan las matrices de confusión de ambos métodos, donde se almacenan el conjunto de posibilidades entre la clase/categoría correcta, evaluando mediante oraciones de las canciones.

Los valores de las columnas son los aciertos de las categorías mientras que los valores de las filas son los errores de la clasificación.

En las siguientes tablas 6 y 7 se muestran las matrices de confusión para ambos métodos de clasificación NB Y NBM, la evaluación de la matriz de acuerdo con la métrica establecida anteriormente y el nivel de confianza de ambos métodos.

CATEGORIA ASIGNADA POR NB									
CATEGORIA REAL		C0	C1	C2	C3	C4	PRECISION	RECALL	PUNTAJE F1	EVALUACIÓN
	C0	25	8	6	4	0	0.462962963	0.5813953488	0.5154639175	El clasificador no detecta las categorías correctamente.
	C1	6	10	8	2	1	0.3846153846	0.3703703704	0.3773584906	El clasificador no detecta las categorías correctamente.
	C2	5	4	19	3	0	0.5	0.6129032258	0.5501246377	El clasificador no detecta las categorías correctamente.
	C3	11	3	1	13	0	0.5909090909	0.4642857143	0.52	El clasificador no detecta las categorías correctamente.
	C4	7	1	4	0	3	0.75	0.2	0.3157894737	El clasificador no detecta la categoría muy bien, pero cuando lo hace es realmente confiable.
	MACROPOROMEDIO:						0.5376974877	0.4457909319	0.4874498964	Teniendo en cuenta la métrica el nivel de confianza es del 48% entra en el rango de media
	ACCURACY						70/144	0.48%		Tiene una baja exactitud

Tabla 7 Matriz de confusión NB.

El método NB presenta un desbalance debido a que solo contempla que el token se presente dentro de la categoría, con la finalidad de mejorar el nivel de confianza se realiza la clasificación de Naive Bayes Multinomial que toma en cuenta la frecuencia de las palabras en cada categoría.

Para medir el nivel de confianza del clasificador la métrica es la siguiente:

- ✓ Porcentaje del 70% su nivel de confianza es optima.
- ✓ Porcentaje del 80% su nivel de confianza es excelente.
- ✓ Porcentaje del 100% su nivel de confianza es perfecto

		CATEGORIA ASIGNADA POR NBM								
CATEGORIA REAL		C0	C1	C2	C3	C4	PRECISION	RECALL	PUNTAJE F1	EVALUACIÓN
	C0	26	13	4	0	0	0.9285714286	0.6046511628	0.73239436	El clasificador no detecta la categoría muy bien, pero cuando lo hace es realmente confiable.
	C1	0	26	1	0	0	0.52	0.962962963	0.6753246753	El clasificador detecta bien la categoría, pero también incluye muestras de otras categorías.
	C2	0	4	26	1	0	0.7878787879	0.8387096774	0.8125	El clasificador detecta correctamente la categoría.
	C3	1	5	0	22	0	0.9166666667	0.7857142857	0.84615384	El clasificador detecta correctamente la categoría
	C4	1	2	2	1	9	1	0.6	0.75	El clasificador no detecta la categoría muy bien, pero cuando lo hace es realmente confiable.
	MACROPOROMEDIO:						0.8306233767	0.7584076158	0.7952026415	Teniendo en cuenta la métrica su nivel de confianza es del 79% entra en el rango de excelente
	ACCURACY						109/144	0.75%		Tiene una buena exactitud

Tabla 8 Matriz de confusión NBM y la evaluación.

La clasificación con el método NBM mejoro el desempeño del clasificador, ya que considera la frecuencia del token´s, más el suavizado de Laplace y el normalizar los datos permite que se ajuste mejor la categoría a la que pertenece.

6.4 Macro promedio

Como el clasificador tiene más de dos categorías entonces se utiliza el macro promedio, que calcula el rendimiento de cada categoría/clase y luego promedio sobre el total categorías.

categoría	0	1	2	3	4
Precisión	0.93	0.52	0.78	0.91	1
Recall	0.60	0.96	0.84	0.78	0.6

Tabla 9 Valores de precisión y recall de las cinco categorías.

$$Precision\ macro\ promedio = \frac{0.93 + 0.52 + 0.78 + 0.91 + 1}{5} = \frac{4.153}{5} = 0.83$$

$$Recall\ macro\ promedio = \frac{0.60 + 0.96 + 0.84 + 0.78 + .6}{5} = \frac{3.791}{5} = 0.75$$

Tiene alta precisión, alto recall y el valor del puntaje F1 nos da un resultado de 78% de confiabilidad ´para el clasificador.

$$PuntajeF1 macro\ promedio = \frac{2 * (precision * recall)}{precision + recall} = \frac{2 * (0.6225)}{1.58} = 0.78$$

Teniendo en cuenta la métrica su nivel de confianza es excelente.

Capítulo 7: Conclusión

Con la realización de este trabajo se presenta una solución de ciencias de datos para el análisis de datos cumpliendo con los objetivos que se plantearon al inicio. Dado que el algoritmo Naive Bayes Multinomial se utiliza para el análisis de datos de texto de múltiples clases. El clasificador identifica las categorías/intérprete de las canciones, se clasifica por género y en algunos casos el nombre de la canción e imprime la letra.

La clasificación está basada en tres aspectos principales:

1. Preprocesamiento: Al tener el conjunto de datos se realiza la limpieza de los datos, la cual se realizó mediante la tokenización, remoción de stop words y el método Bag of words.
2. Análisis de datos: se lleva a cabo mediante la tabla de frecuencias y normalizando los datos
3. Clasificación: se realiza mediante el método Naive Bayes Multinomial obteniendo como resultados que el clasificador tiene mayor efectividad en las categorías con mayor frecuencia de palabras.

Una de las desventajas del método es que, al no estar equilibradas las categorías respecto a la cantidad de palabras, se asigna la probabilidad a posteriori a la que tiene mayor vocabulario.

7.1 Trabajo a futuro realizar la clasificación de sentimientos.

En el presente todo lo que vivimos es digital, los datos fluyen a través de las aplicaciones que utilizamos esto es una oportunidad para comprender lo que nos rodea dentro de las ciencias sociales y de comunicación.

La ciencia de datos nos permite que la comunicación se adapte logrando aumentar la efectividad de los mensajes considerando una gran cantidad de perspectivas sobre diferentes temas fomentando que la toma de decisiones sea basada en datos.

Teniendo en cuenta la desventaja que presento método, las mejoras que se tomarían en cuenta para el trabajo a futuro.

Evaluar el desempeño del clasificador en:

- ✓ Aumentar el nivel de categorías
- ✓ Agregar más idiomas
- ✓ Equilibrar el vocabulario de las categorías

Los datos no son solo un recurso, son una ventaja para la toma de decisiones.

Bibliografía

Abdulwahab Alazeb, M. A. (2021). Review on Data Science and Prediction on Computing . *International Conference on Computing and Data Science*, 8.

Araujo, B. S. (2006). *Aprendizaje automatico:Conceptos basicos y avanzados.* Madrid : Pearson Prentice Hall.

Arce, J. I. (20 de 11 de 2022). Obtenido de La matriz de confusión y sus metricas - Inteligencia artificial La matriz de confusión y sus métricas Inteligencia Artificial : https://www.juanbarrios.com/la-matriz-de-confusion-y-sus-metricas/

datos.gob.es. (05 de 04 de 2018). *Ciencia de datos, machine learning y deep learning*. Obtenido de https://datos.gob.es/es/blog/ciencia-de-datos-machine-learning-y-deep-learning

Deshpande, V. k. (2019). *Data Science Concepts and Practice.* United States: Morgan Kaufman.

Fernando Molina Graja, L. B. (2022). Demand and employability study of the data science engineering career in Ecuador. *Iberian Conference on Informattion Systems and Technologies* , 5.

Gonzalo Pajares Martinsanz, M. S. (2006). *Inteligencia Artificial e ingenieria del conocimiento.* Mexico: Alfaomega.

Hernández, L. d. (s.f.). *Programarfacil*. Obtenido de https://programarfacil.com/tutoriales/fragmentos/servomotor-con-arduino/

Iñigo Martinez, E. V. (2021). A survey study of success factors in data science projects. *International Conference on Big Data*, 6.

Juan J. Cuadrado-Gallego, Y. D. (2021). Classifitacion and Analysis of Techniques and Tools for Data Visualizations Teaching. *IEEE Global Engineering Education Conference*, 7.

Jurafsky, D. &. (1999). *Speech and Language Processing.* New Jersy: Prentice-Hall.

Kotu, V. (219). Introduction. En B. Deshpande, *Data Science Concepts and Practice* (pág. 549). Cambridge, MA: Morgan Kaufmann Publishers , and imprint of Elservier.

Martín, J. A. (Febrero de 2023). *Blog de Tecnología - IMF Smart Education*. Obtenido de Glosario de Procesamiento de Lenguaje Natural (NLP): https://blogs.imf-formacion.com/blog/tecnologia/glosario-de-procesamiento-de-lenguaje-natural-nlp-202302/

Scherz, A. (2018). Clasificación automática de papers de Ciencias de la Computación. 64.