QFEbonn

Quantitative Finance & Economics

German | Spanish | Chinese

Warning

This version of the presentation is automatically machine-translated. Please use it with caution and always refer to the original English version for accuracy.

Spotify Musik Genre Analyse

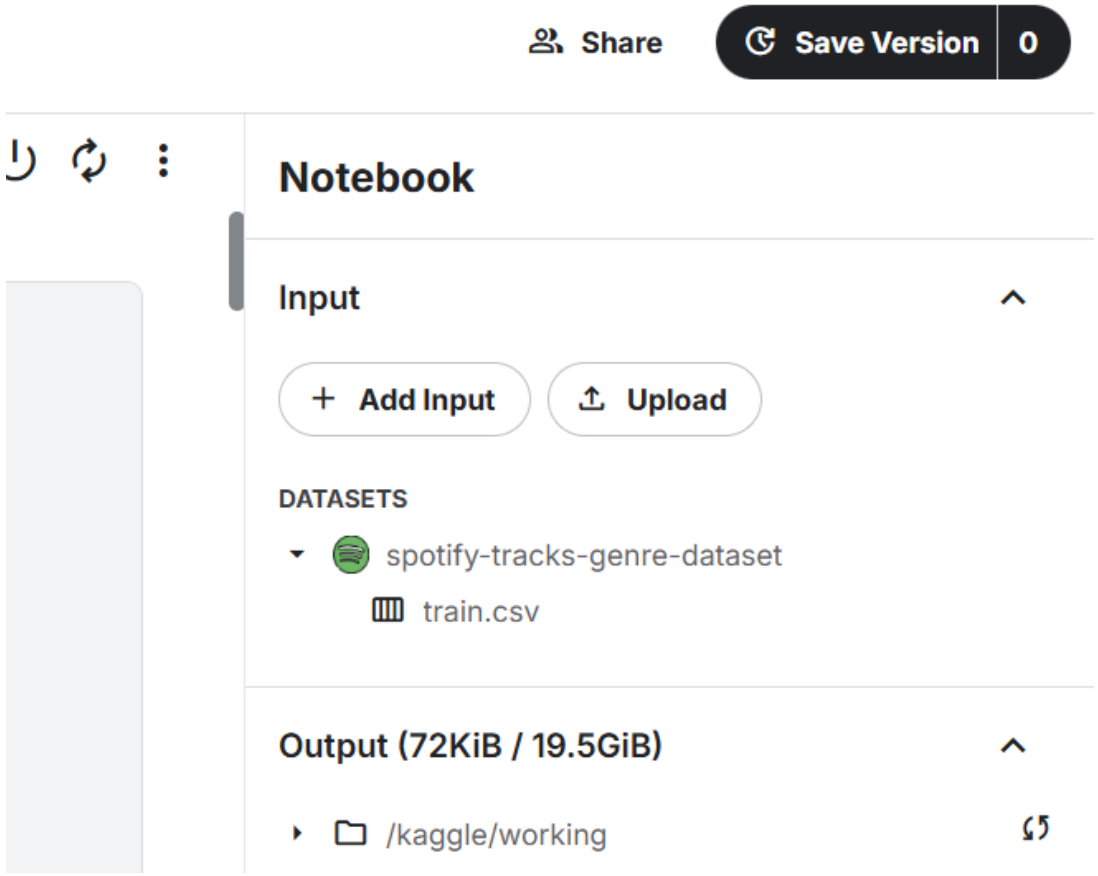
Quantitative Finanzierung und ÖkonomieBonn

Einführung

- Datensatz:Spotify Musik vonKaggleUndHugging Face
- Ziel:Reinigen Sie unordentliche Textdaten, um sich auf die Analyse vorzubereiten
- Warum die Reinigung von Bedeutung:
  schmutzige Daten ->**fehlerhaft**Erkenntnisse

EinstellungHoch

- Wir werden benutzenKaggleUm direkt auf Daten zuzugreifen und Code live auszuführen
- Sie müssen:
  - Melden Sie sich anKaggle.com
  - Erstellen Sie ein neues Notizbuch
  - Fügen Sie diesen Datensatz in Ihre Eingabe hinzu:kaggle.com/datasets/thedevastator/spotify-tracks-genre-dataset



DatenInspektion

- Verwendenhead(), shape, describe(), info(), nunique() einen kurzen Blick auf die Daten zu werfen
- Überprüfen Sie, ob es Duplikate gibt oder ob eine der Spalte fehlende Werte hat
- Versuchen Sie herauszufinden, welche Dateneinträge fehlende Werte haben
- Verwendenfillna() Fehlende Werte ersetzen(e.g. ‘’ or ‘Unknown’)
- Finden Sie heraus, welche Säulen nicht numerisch sind(‘object’)
- Finden Sie heraus, welcher Künstler die meisten Tracks hat

DatenVisualisierung

- Verwendenplt and sns.histplot() um die zu zeichnenVerteilung der Popularität der Spur
- Verwendengroupby() um die Top 10 Genres mit den höchsten zu findenbedeutenvonPopularität
- Verwendensns.boxplot() um die zu zeichnenVerteilung der Popularität durch Genre (Top 10 Genre)

DatenReinigung

- Erstellen wir eine neue Spalte‘clus_att’ kurz fürClustering -Attribute
- Wir konzentrieren uns jetzt auf die Reinigung dieser neuen Spalte:
  - Interpunktion entfernen
  - Entfernen Sie Non -ASCII -Zeichen
  - Stoppwörter entfernen
  - Duplikate entfernen
  - Wörter tokenisieren
  - Lemmatisieren Verben

```
df['clustering_attributes'] = (df['artists'] + ' ' +
                               df['track_name'] + ' ' +
                                df['track_genre'])
```

Nicht-ASCII-Zeichen und Stoppen Sie Wörter

❌ Examples of Non-ASCII Characters

These characters are **not** part of the basic ASCII set:

Character	Description
é	Latin small e with acute
ñ	Latin small n with tilde
Ω	Greek capital omega
£	British pound sign
™	Trademark symbol
😊	Smiling face emoji
—	Em dash (long dash)

🛑 What Are Stop Words?

Stop words are **common words** in a language that are often **ignored** in text analysis or search engines.

📌 **Examples (in English):**

csharp Copy Edit

```csharp
a  an  the  and  or  but  is  are  was  were  in  on  at  with
```

📘 Why are they ignored?

They **don't add much meaning** and are used frequently, so removing them helps:

- Speed up processing
- Focus on important words

Tokenize und Wörter lemmatisieren

🔤 What is Word Tokenization?

Word tokenization is the process of **splitting text into individual words**, called **tokens**.

📘 Example:

Text:

```sql
I love natural language processing.
```

Tokens:

```css
["I", "love", "natural", "language", "processing", "."]
```

🌱 What is Lemmatization?

Lemmatization is the process of reducing a word to its **base or dictionary form**, called a **lemma**.

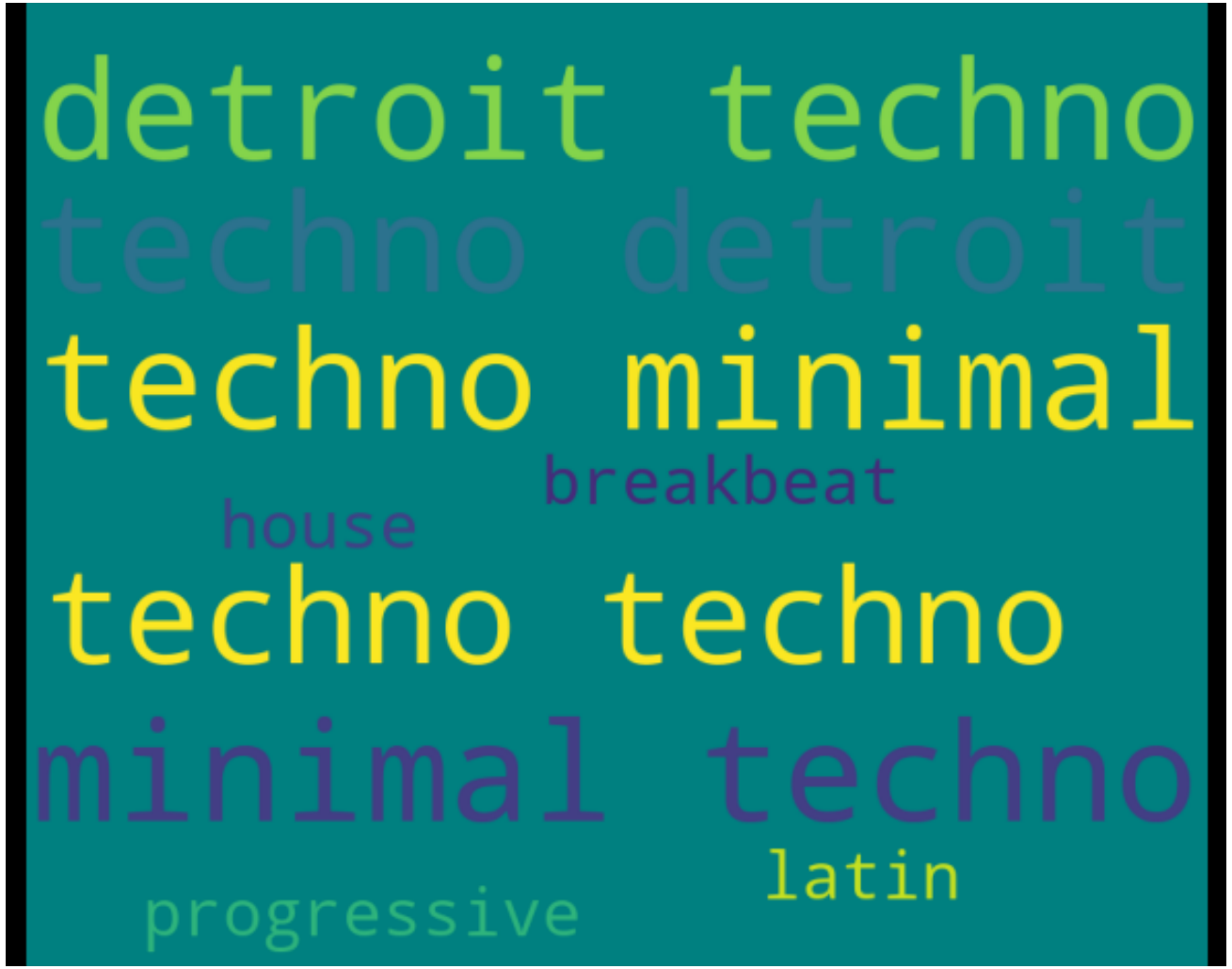
📘 Example:

Word	Lemma
running	run
better	good
studies	study
mice	mouse

WortWolke

Gemeinsam Musikgenres der 1. Cluster

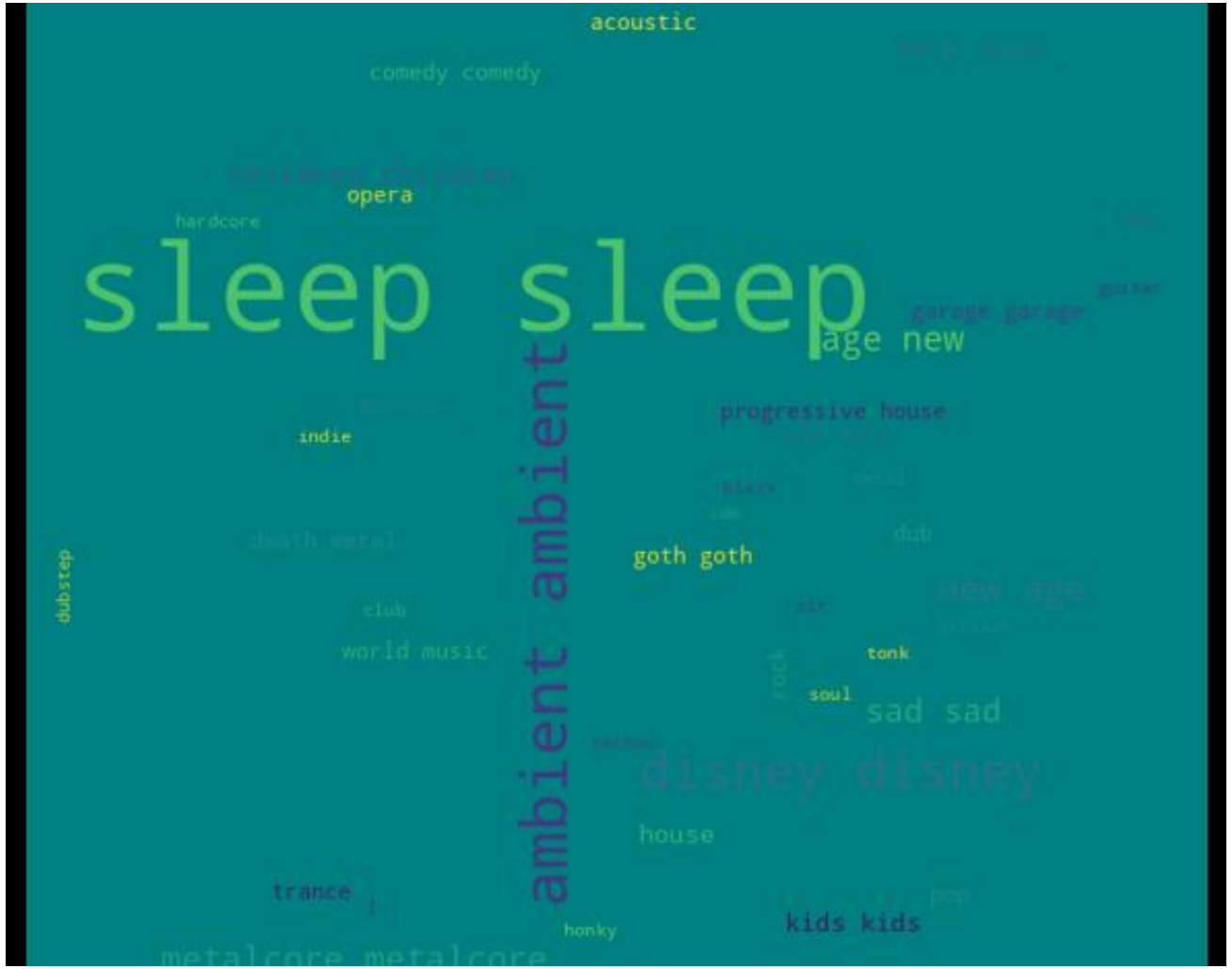
Berechnet von **K-Means Clustering** (K = 6)



WortWolke

Gemeinsam Musikgenres der 2. Cluster

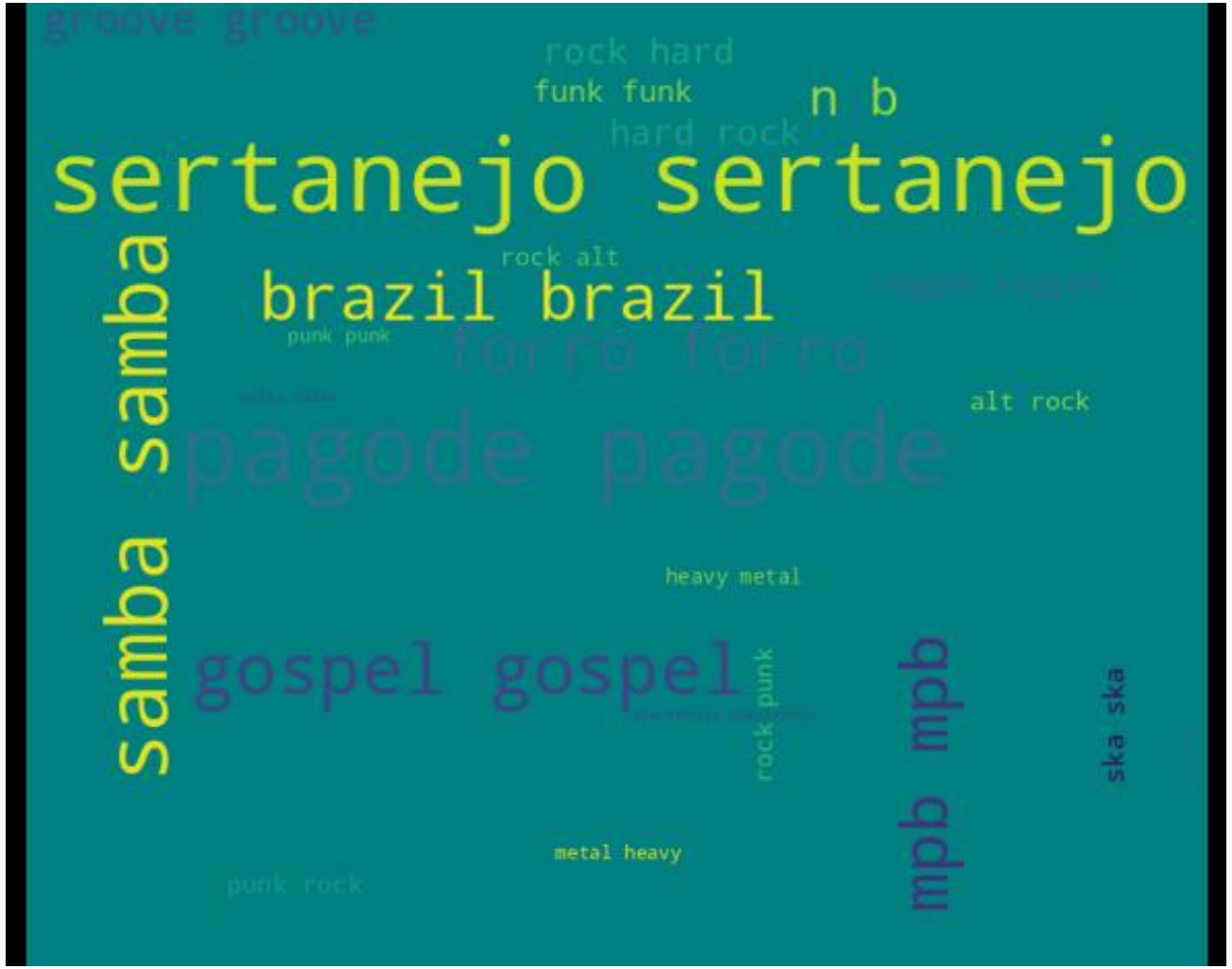
Berechnet von **K-Means Clustering** (K = 6)



WortWolke

Gemeinsam Musikgenres der 3. Cluster

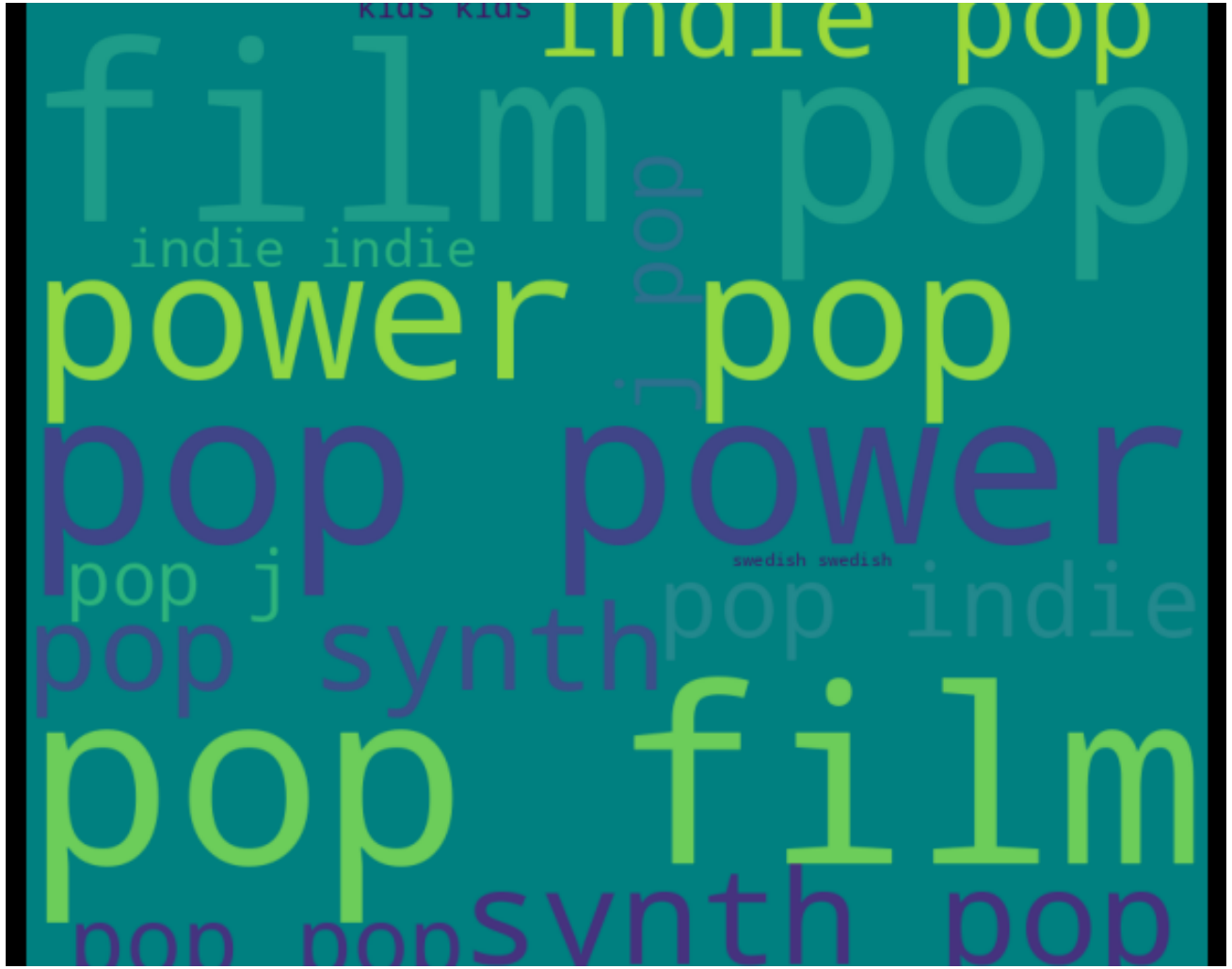
Berechnet von **K-Means Clustering** (K = 6)



WortWolke

Gemeinsam Musikgenres der 4. Cluster

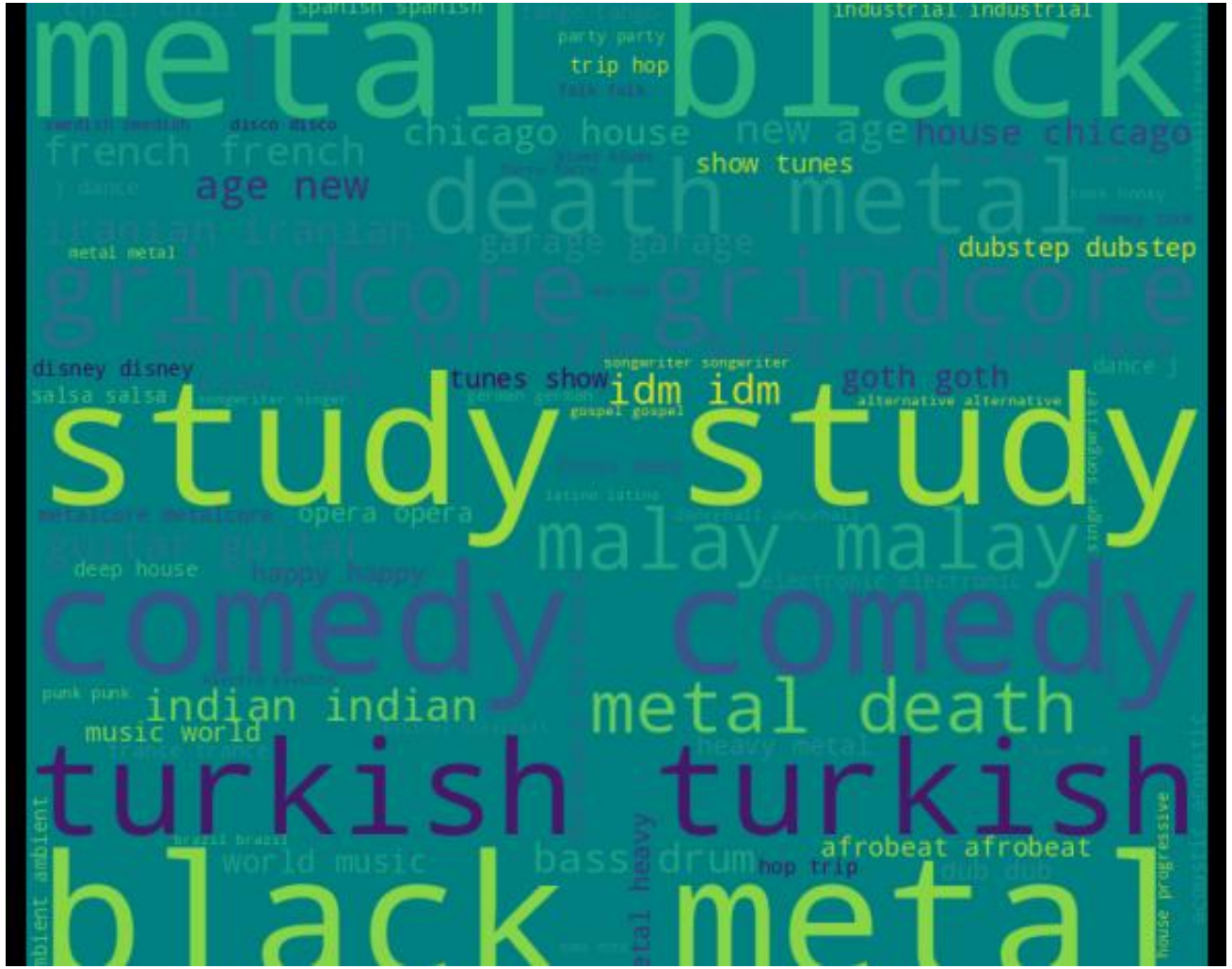
Berechnet von **K-Means Clustering** (K = 6)



WortWolke

Gemeinsam Musikgenres der 5. Cluster

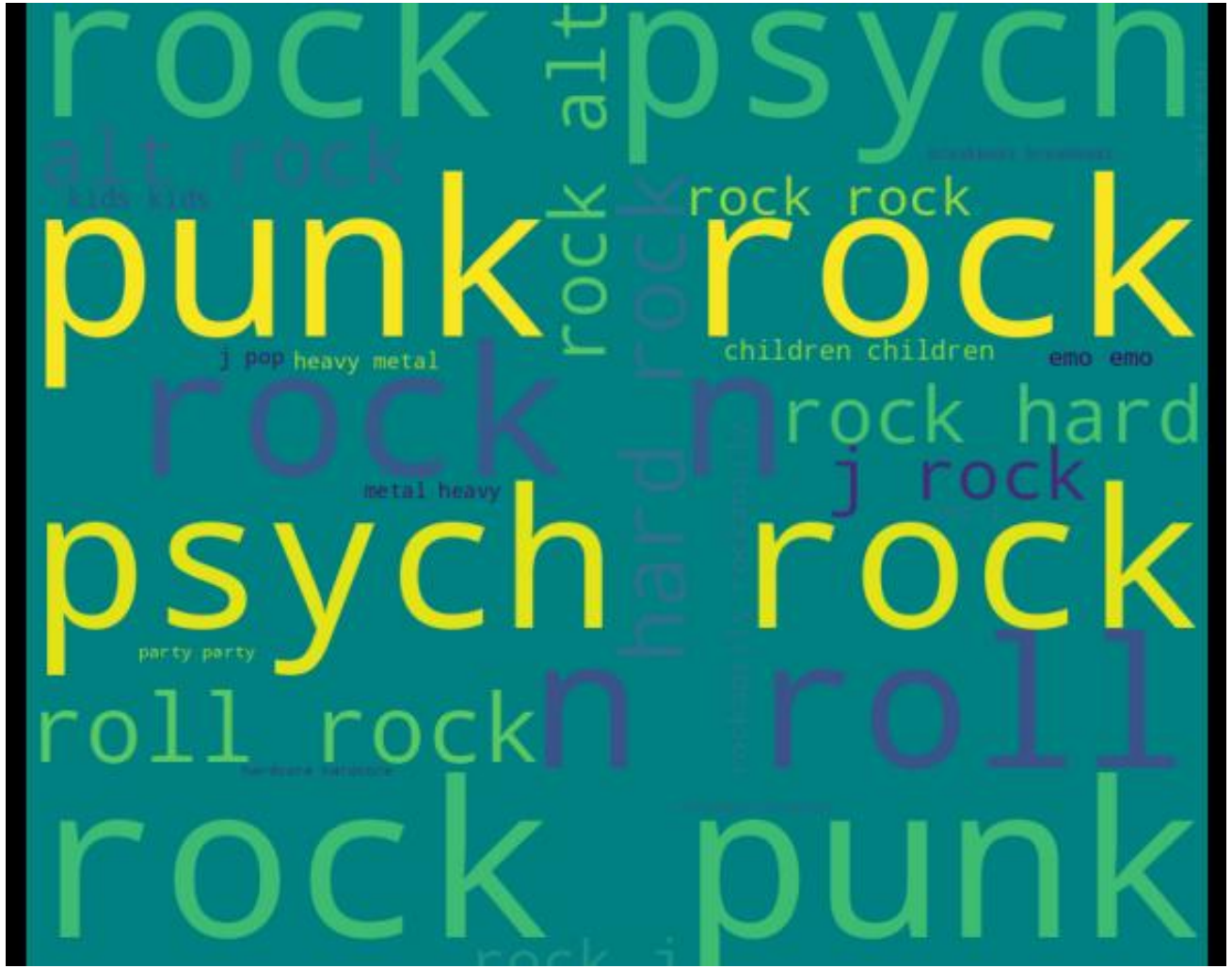
Berechnet von **K-Means Clustering** (K = 6)



WortWolke

GemeinsamMusikgenresder6. Cluster

Berechnet von**K-Means Clustering** (K = 6)



Spotify Música Género Análisis

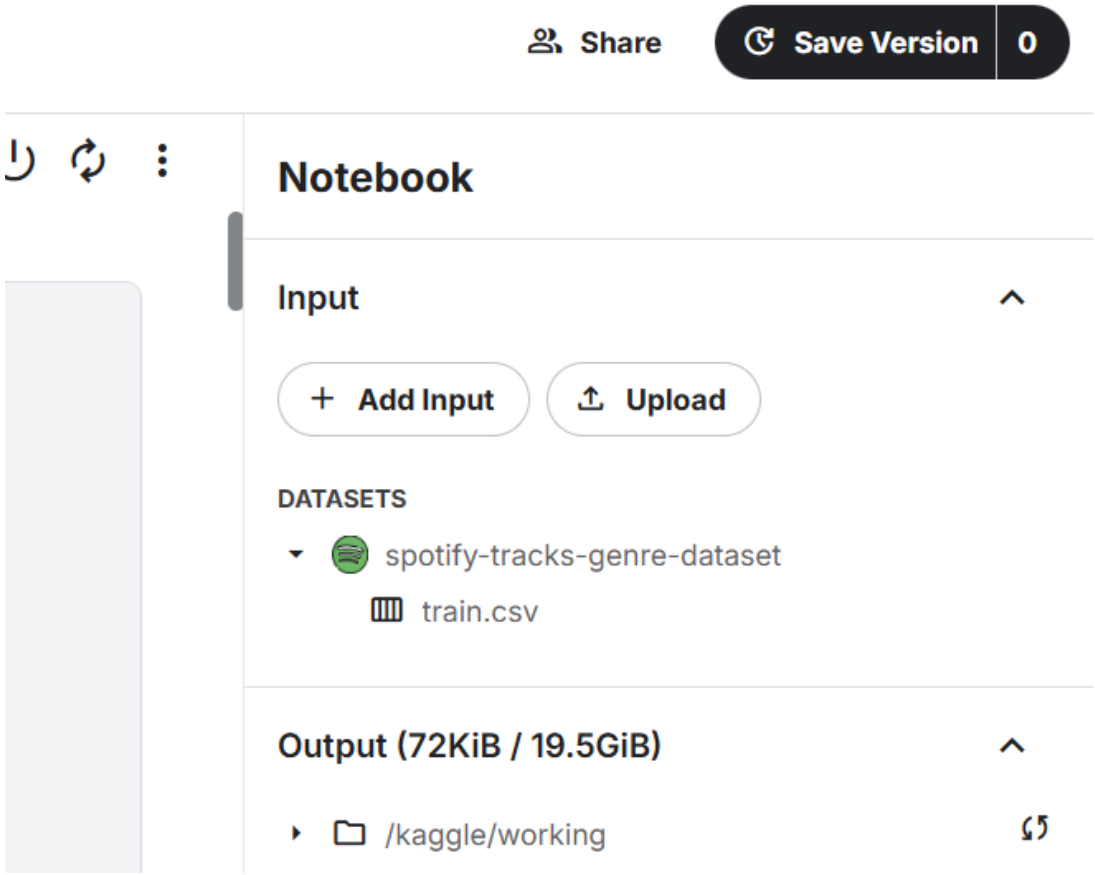
Finanzas cuantitativas y economíaBonn

Introducción

- Conjunto de datos: Música de Spotify de Kaggle y Hugging Face
- Meta: Limpiar datos de texto desordenados para prepararse para el análisis
- Por qué la limpieza es importante:
  datos sucios -
  >**flawed** perspectivas

ConfiguraciónArriba

- Estaremos usadosKagglePara acceder directamente a los datos y ejecutar el código en vivo
- Necesitará:
  - Registrarse enKaggle.com
  - Crea un nuevo cuaderno
  - Agregue este conjunto de datos en su entrada:kaggle.com/datasets/thedevastator/spotify-tracks-genre-dataset



DatosInspección

- Usarhead(), shape, describe(), info(), nunique() Dar un vistazo rápido a los datos
- Compruebe si hay duplicados o si alguno de la columna tiene valores faltantes
- Intente averiguar qué entradas de datos tiene valores faltantes
- Usarfillna() Para reemplazar los valores faltantes(e.g. ‘’ or ‘Unknown’)
- Descubra qué columnas no son numéricas(‘object’)
- Descubra qué artista tiene la mayoría de las pistas

DatosVisualización

- Usarplt and sns.histplot() Para trazar elDistribución de la popularidad de la pista
- Usargroupby() para encontrar los 10 principales géneros con los más altossignificardePopularidad
- Usarsns.boxplot() Para trazar elDistribución de popularidad por género (género 10 top)

DatosLimpieza

- Creemos una nueva columna‘clus_att’ inquietudAtributos de agrupación
- Nos centramos ahora en limpiar esta nueva columna:
  - Eliminar la puntuación
  - Eliminar caracteres no ascii
  - Eliminar las palabras de parar
  - Eliminar los duplicados
  - Tokenize Words
  - Lemmatizar verbos

```
df['clustering_attributes'] = (df['artists'] + ' ' +
                               df['track_name'] + ' ' +
                                df['track_genre'])
```

Personajes no ascii yDetener las palabras

❌ Examples of Non-ASCII Characters

These characters are **not** part of the basic ASCII set:

Character	Description
é	Latin small e with acute
ñ	Latin small n with tilde
Ω	Greek capital omega
£	British pound sign
™	Trademark symbol
😊	Smiling face emoji
—	Em dash (long dash)

🛑 What Are Stop Words?

Stop words are **common words** in a language that are often **ignored** in text analysis or search engines.

📌 **Examples (in English):**

csharp Copy Edit

```csharp
a  an  the  and  or  but  is  are  was  were  in  on  at  with
```

📘 Why are they ignored?

They **don't add much meaning** and are used frequently, so removing them helps:

- Speed up processing
- Focus on important words

Tokenizar yLemmatizar palabras

🔤 What is Word Tokenization?

Word tokenization is the process of **splitting text into individual words**, called **tokens**.

📘 Example:

Text:

```sql
I love natural language processing.
```

Tokens:

```css
["I", "love", "natural", "language", "processing", "."]
```

🌱 What is Lemmatization?

Lemmatization is the process of reducing a word to its **base or dictionary form**, called a **lemma**.

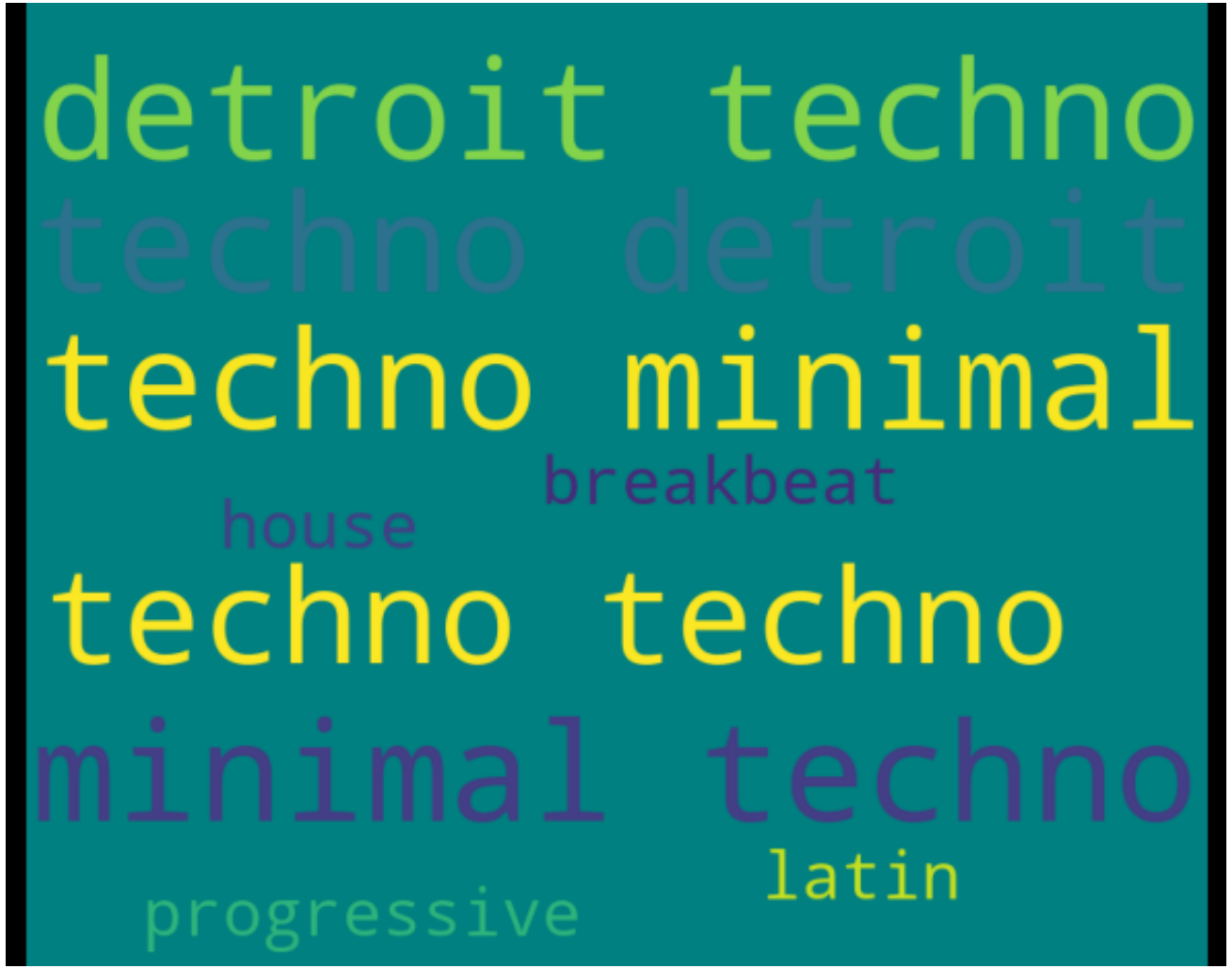
📘 Example:

Word	Lemma
running	run
better	good
studies	study
mice	mouse

PalabraNube

Comúngéneros musicalesdel1st clúster

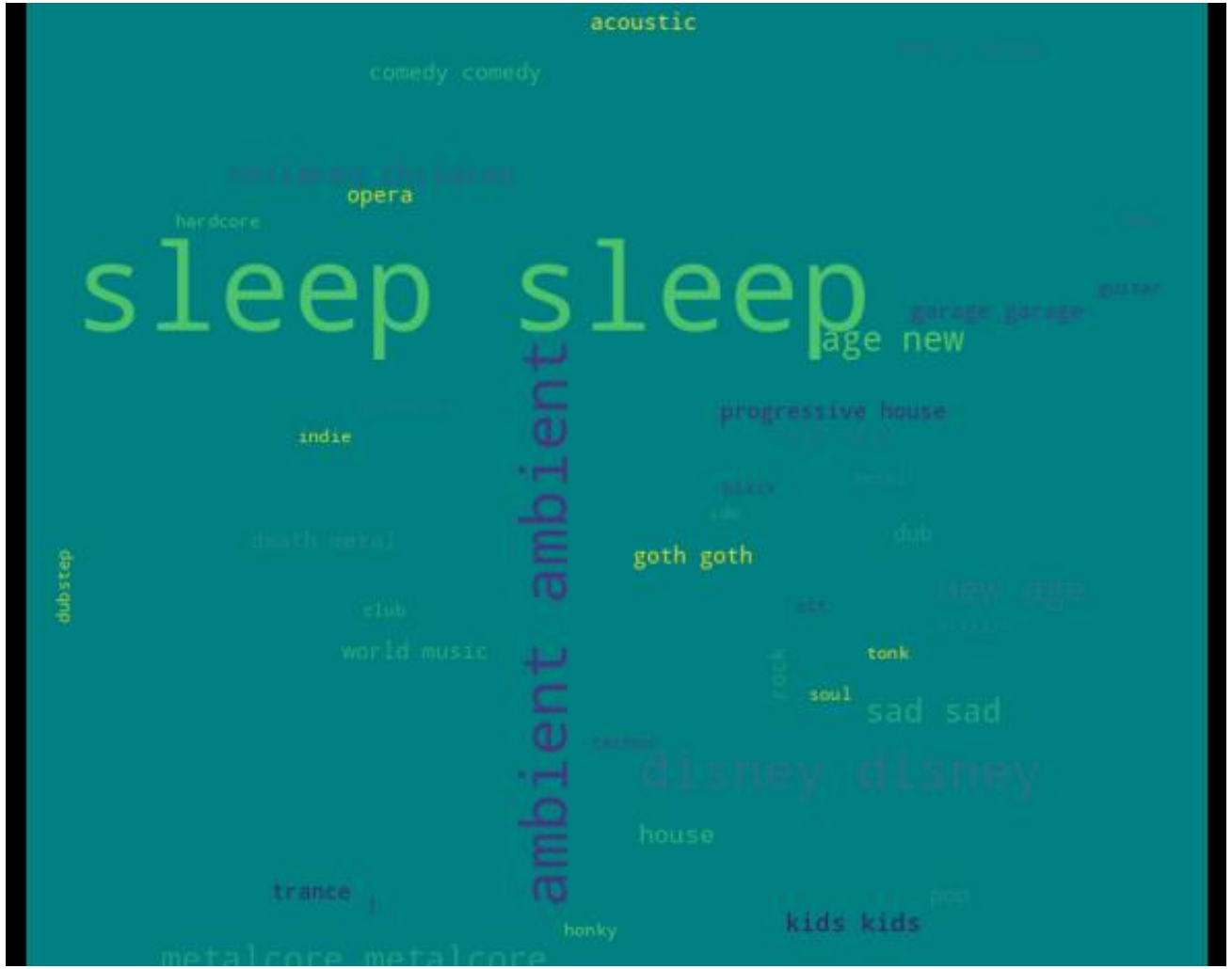
Calculado por**Clúster K-means** (K = 6)



PalabraNube

Comúngéneros musicalesdel2do grupo

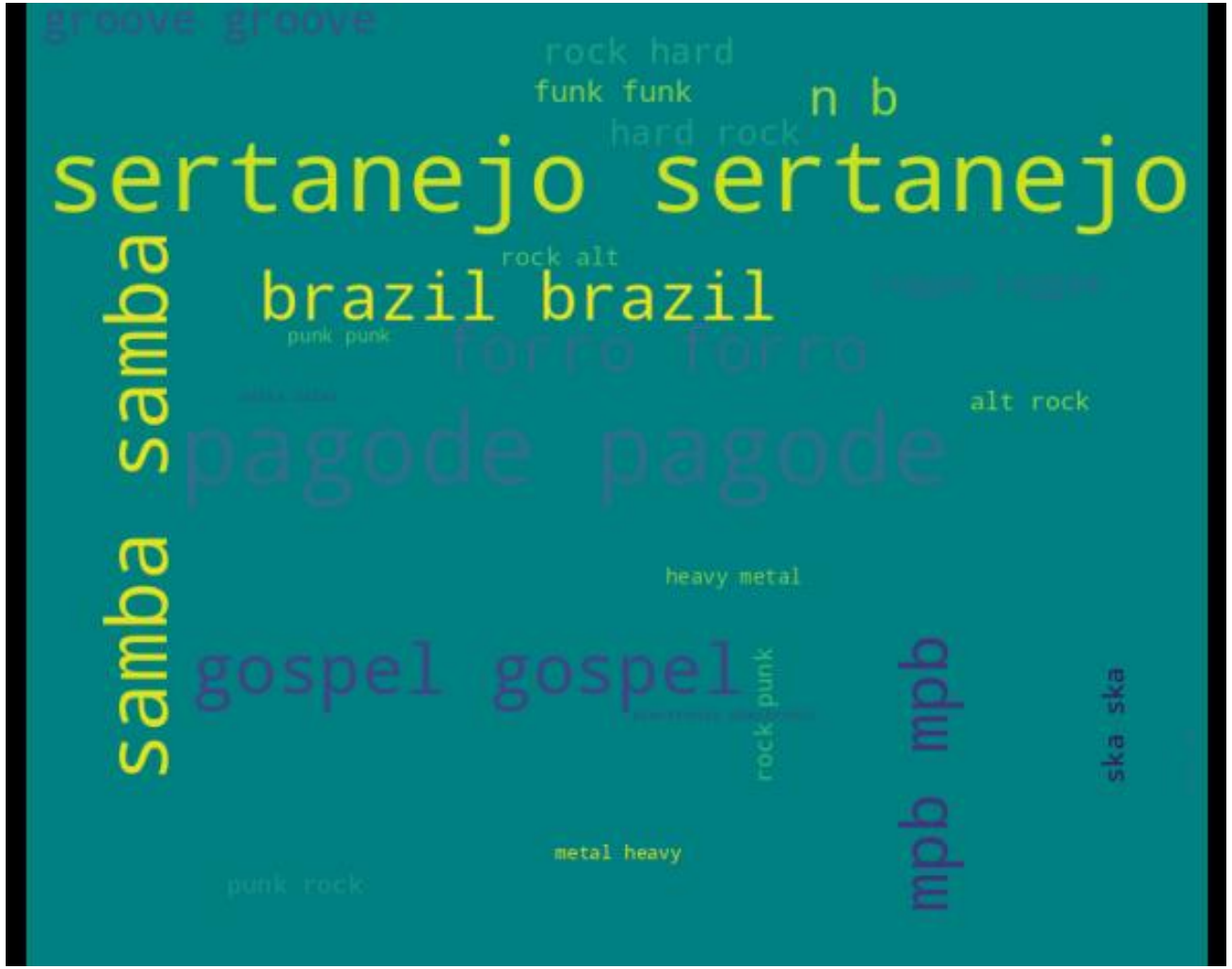
Calculado por**Clúster K-means** (K = 6)



PalabraNube

Comúngéneros musicalesdel3er clúster

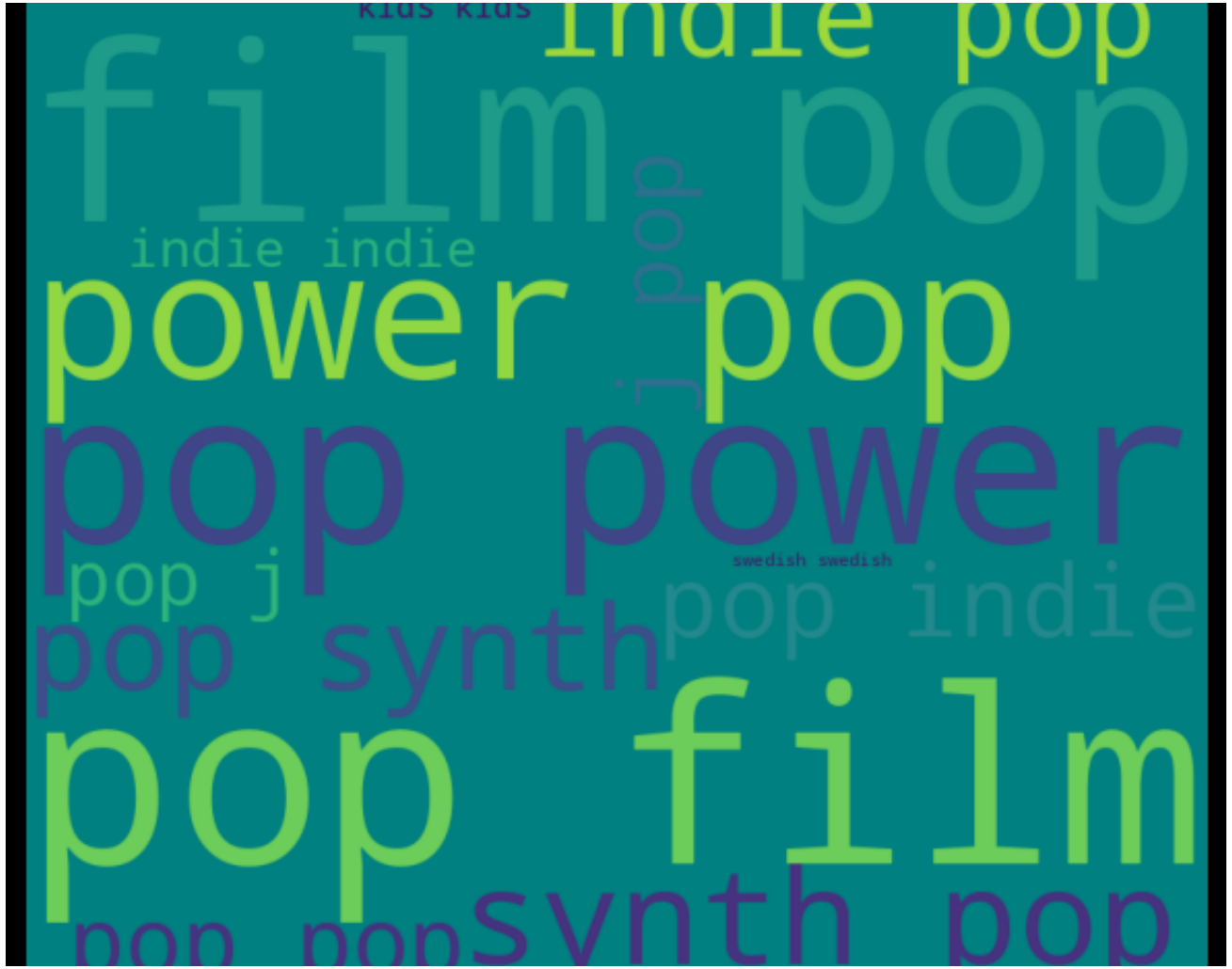
Calculado por**Clúster K-means** (K = 6)



PalabraNube

Comúngéneros musicalesdel4to grupo

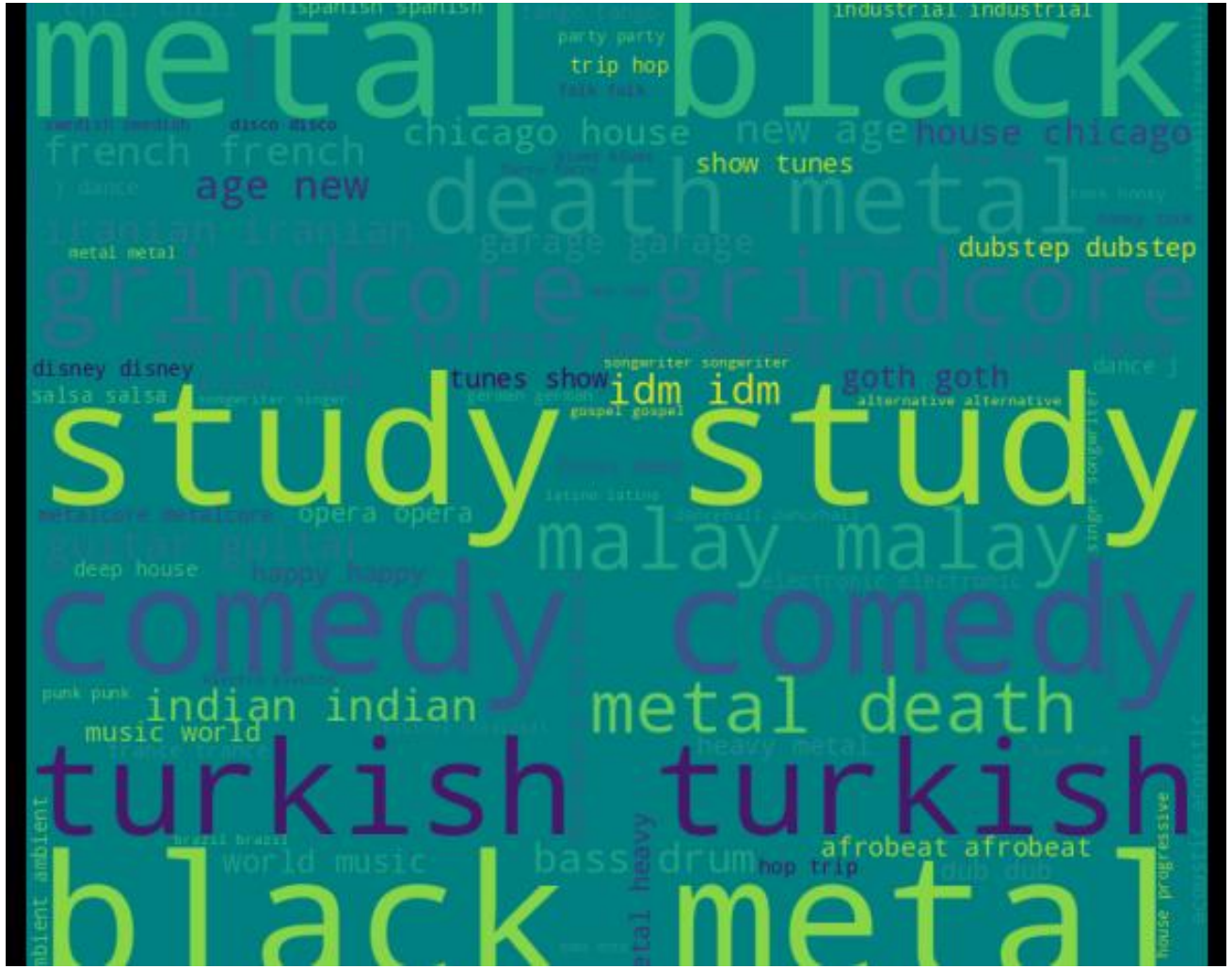
Calculado por**Clúster K-means** (K = 6)



PalabraNube

Comúngéneros musicalesdelQuinto grupo

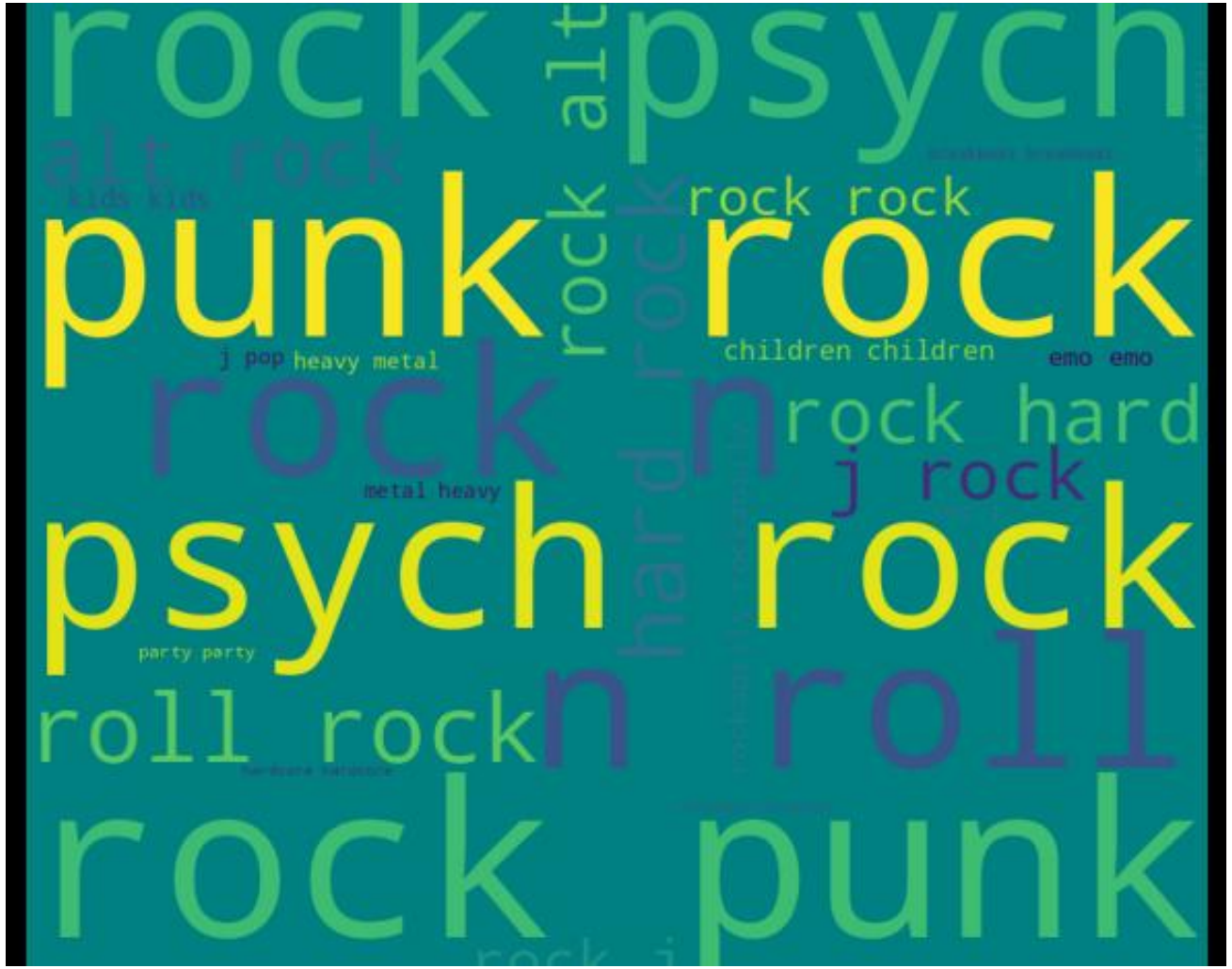
Calculado por**Clúster K-means** (K = 6)



PalabraNube

Comúngéneros musicalesdel6º grupo

Calculado por**Clúster K-means** (K = 6)



Spotify 音乐 类型 分析

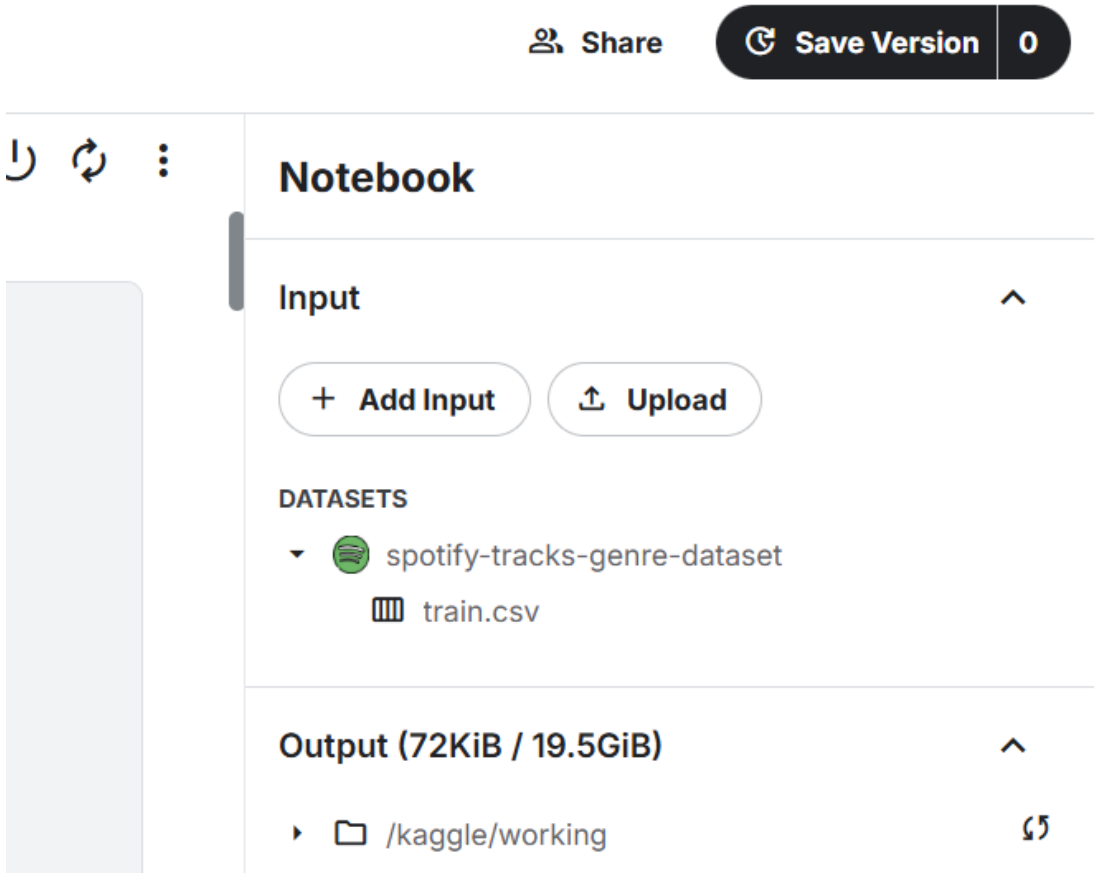
量化金融与经济学波恩

介绍

- 数据集：来自Spotify音乐 Kaggle和Hugging Face
- 目标：清洁凌乱的文本数据以准备分析
- 为什么清洁很重要：脏数据 - >**有缺陷**见解

环境向上

- 我们将使用Kaggle直接访问数据并实时运行代码
- 您将需要：
  - 注册Kaggle.com
  - 创建新笔记本
  - 将此数据集添加到您的输入中：kaggle.com/datasets/thedevastator/spotify-tracks-genre-dataset



数据检查

- 使用head(), shape, describe(), info(), nunique() 快速查看数据
- 检查是否有重复或任何列缺少值
- 尝试找出哪些数据条目缺少值
- 使用fillna() 替换缺失值(e.g. ‘’ or ‘Unknown’)
- 找出哪些列不是数值(‘object’)
- 找出哪个艺术家的曲目最多

数据可视化

- 使用plt and sns.histplot() 绘制轨道受欢迎程度的分布
- 使用groupby() 找到最高的十大流派意思是的受欢迎程度
- 使用sns.boxplot() 绘制流行的流行分布（前十种类型）

数据打扫

- 让我们创建一个新专栏 ‘clus_att’ 缩写聚类属性
- 现在，我们专注于清洁此新专栏：
  - 删除标点符号
  - 删除非ASCII字符
  - 删除停止单词
  - 删除重复项
  - 令牌单词
  - lemmatize动词

```
df['clustering_attributes'] = (df['artists'] + ' ' +
                               df['track_name'] + ' ' +
                                df['track_genre'])
```

非ASCII字符和停止文字

❌ Examples of Non-ASCII Characters

These characters are **not** part of the basic ASCII set:

Character	Description
é	Latin small e with acute
ñ	Latin small n with tilde
Ω	Greek capital omega
£	British pound sign
™	Trademark symbol
😊	Smiling face emoji
—	Em dash (long dash)

🛑 What Are Stop Words?

Stop words are **common words** in a language that are often **ignored** in text analysis or search engines.

📌 **Examples (in English):**

csharp Copy Edit

```csharp
a  an  the  and  or  but  is  are  was  were  in  on  at  with
```

📘 Why are they ignored?

They **don't add much meaning** and are used frequently, so removing them helps:

- Speed up processing
- Focus on important words

令牌和诱人的单词

🔤 What is Word Tokenization?

Word tokenization is the process of **splitting text into individual words**, called **tokens**.

📘 Example:

Text:

```sql
I love natural language processing.
```

Tokens:

```css
["I", "love", "natural", "language", "processing", "."]
```

🌱 What is Lemmatization?

Lemmatization is the process of reducing a word to its **base or dictionary form**, called a **lemma**.

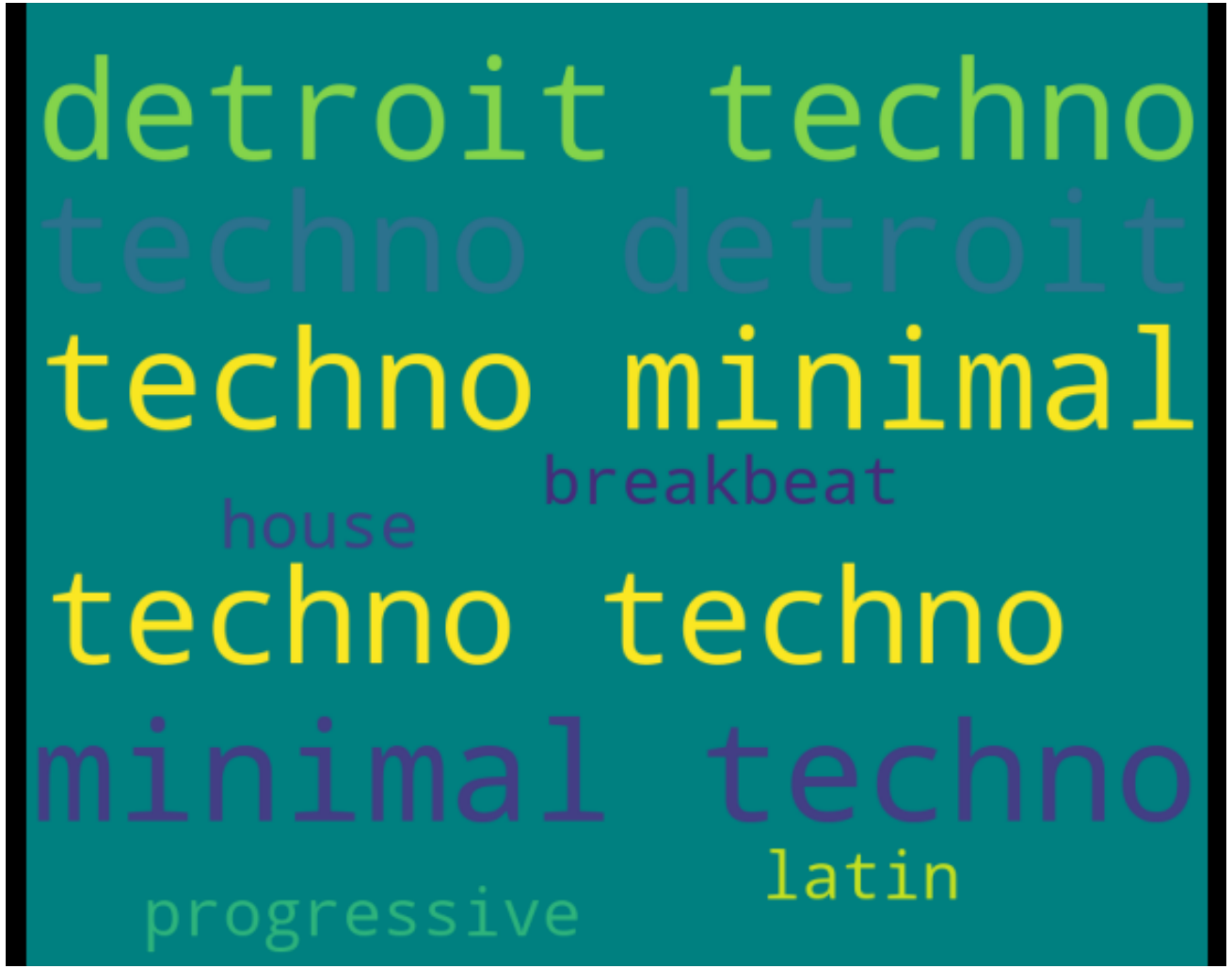
📘 Example:

Word	Lemma
running	run
better	good
studies	study
mice	mouse

单词云

常见的音乐流派的第一集群

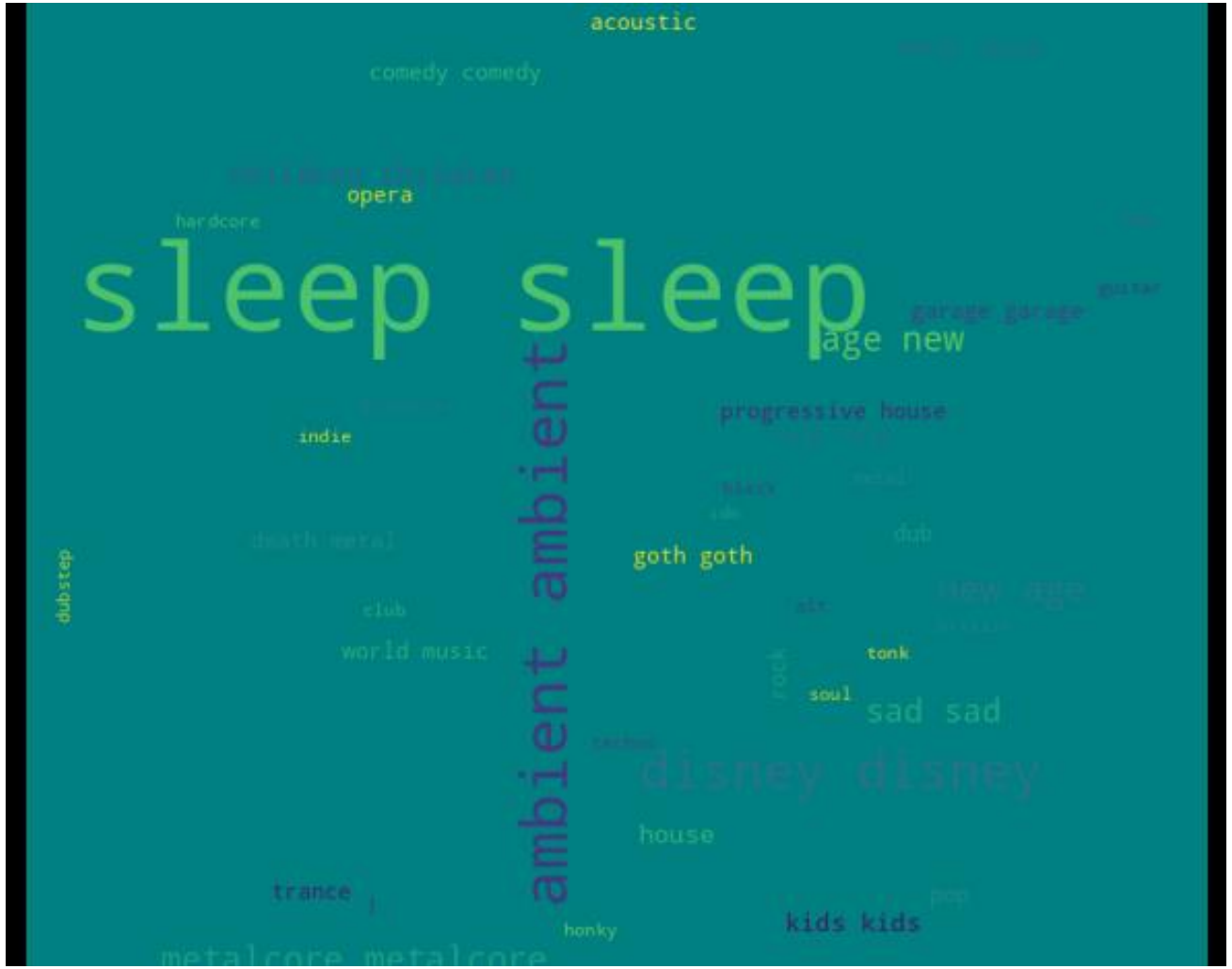
由**K-均值聚类** （k = 6）



单词云

常见的音乐流派的第二集群

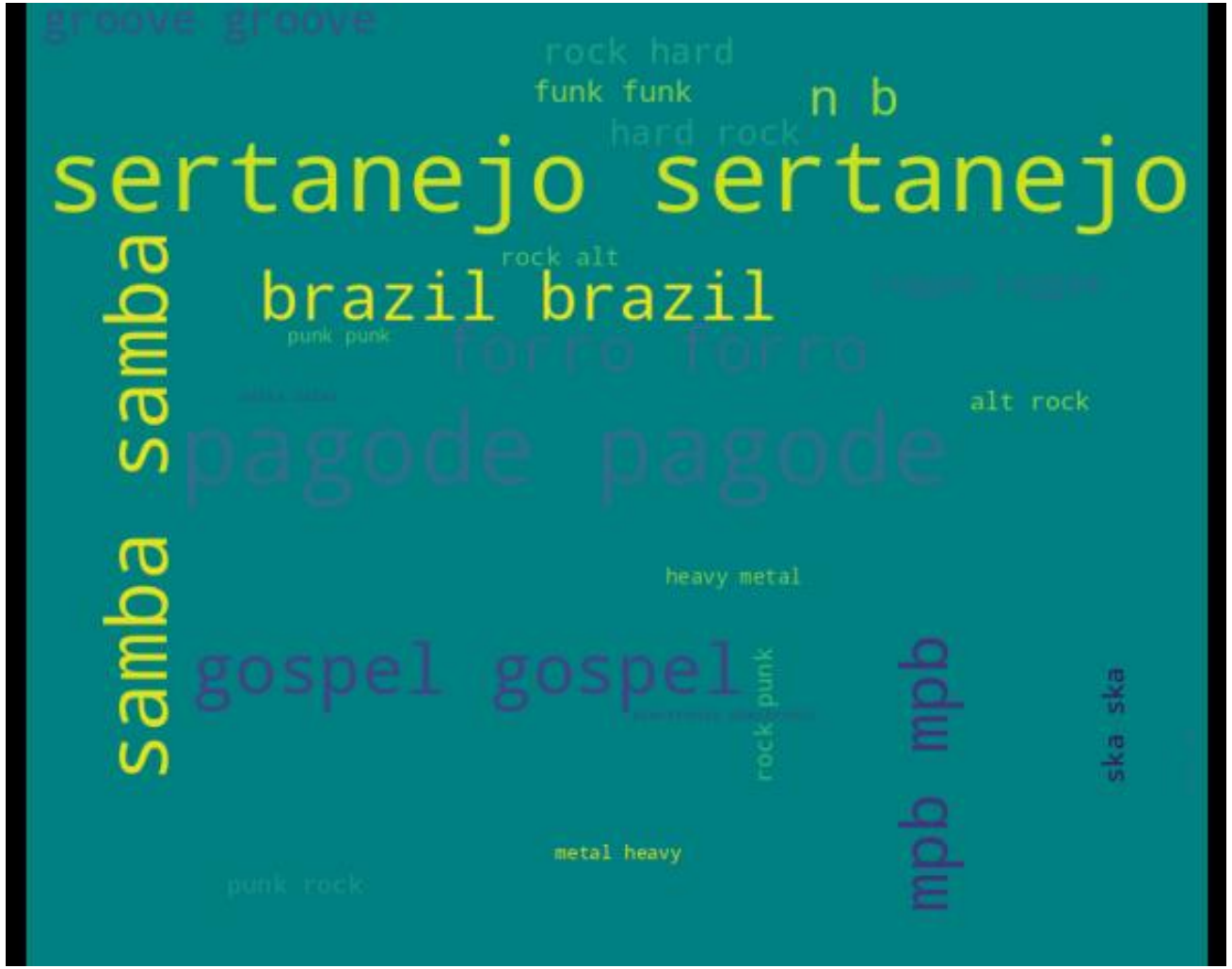
由**K-均值聚类** （k = 6）



单词云

常见的音乐流派的第三集群

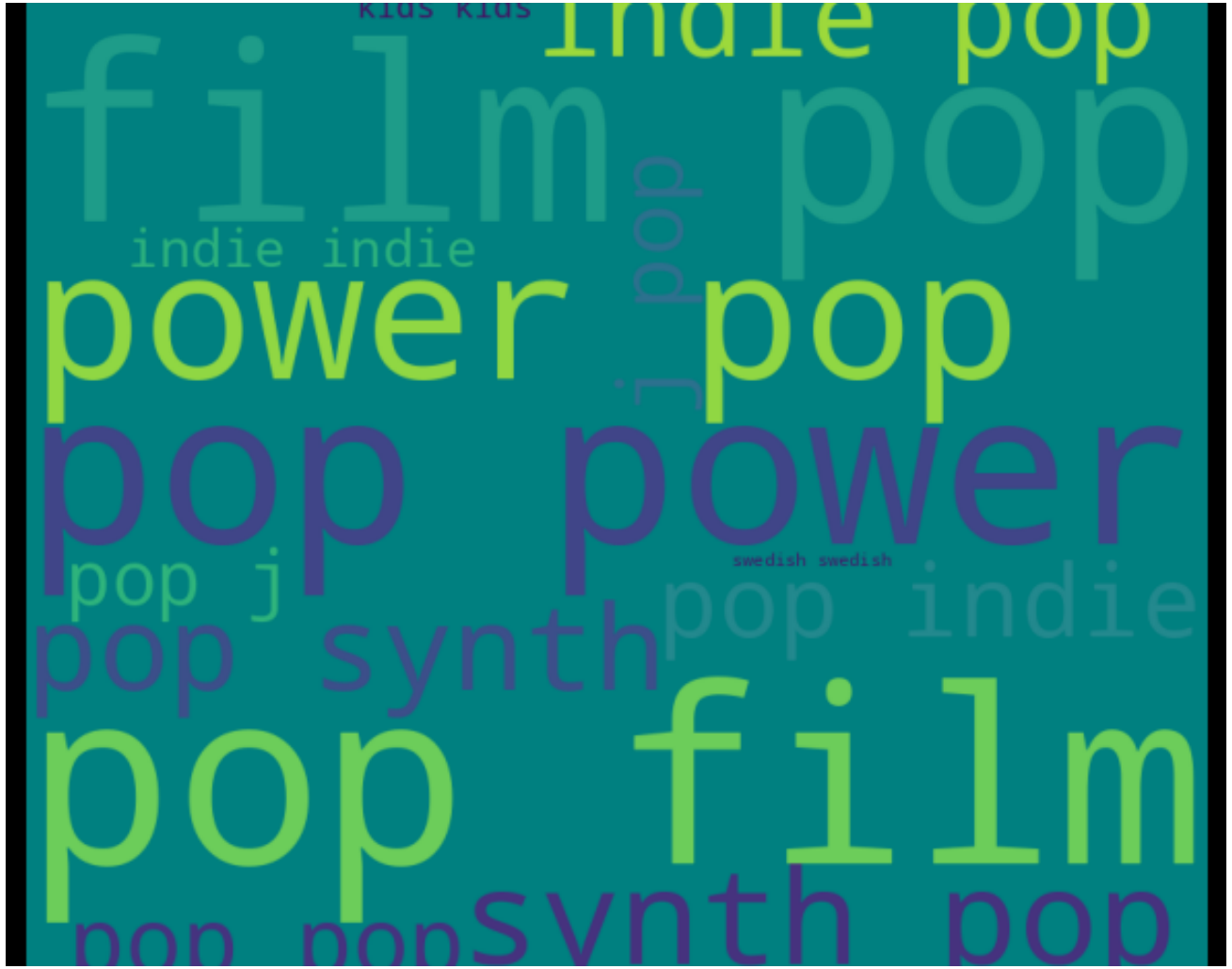
由**K-均值聚类** （k = 6）



单词云

常见的音乐流派的第四集群

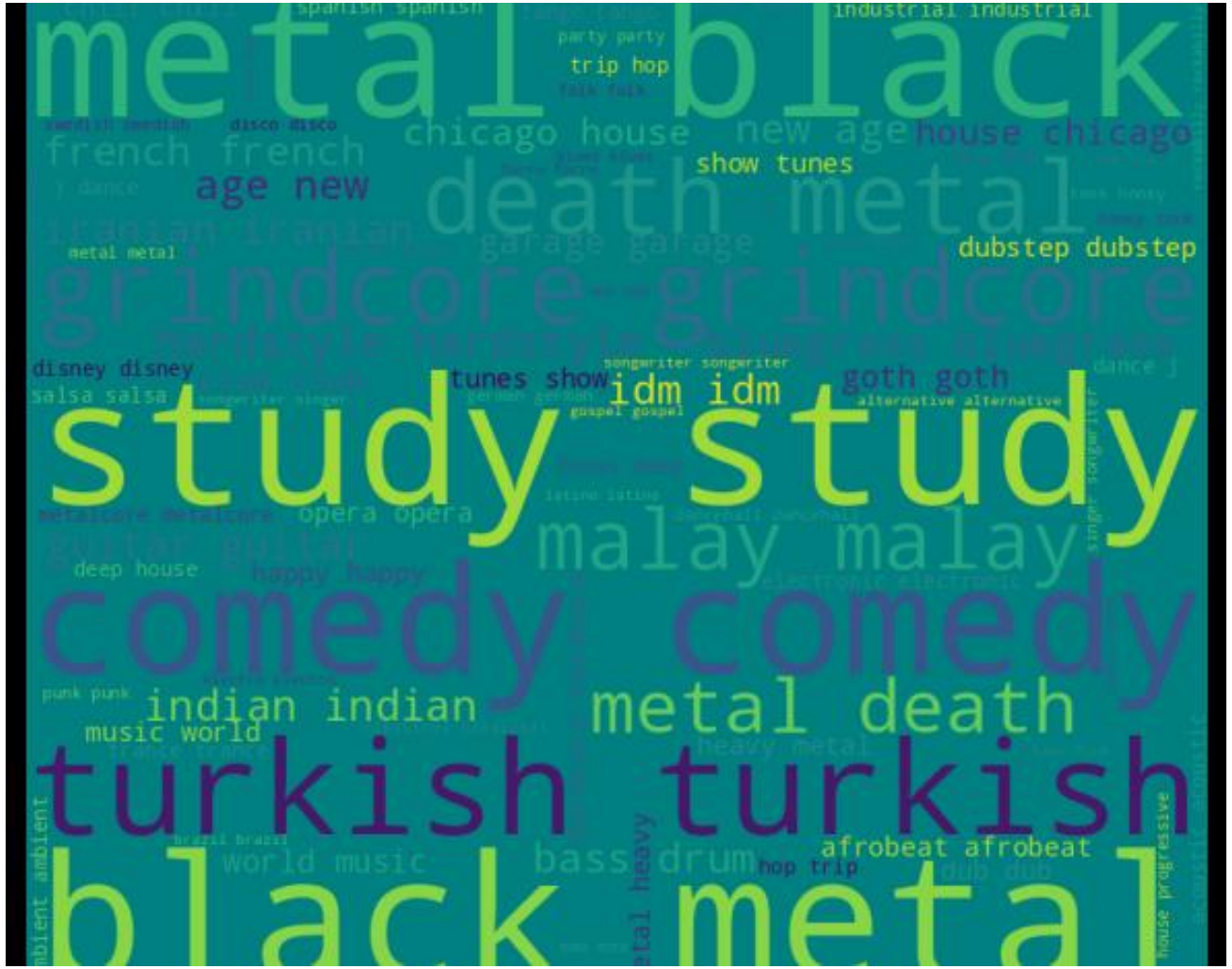
由**K-均值聚类** （k = 6）



单词云

常见的音乐流派的第五集群

由**K-均值聚类** （k = 6）



单词云

常见的音乐流派的第六集群

由**K-均值聚类** （k = 6）