

Generación automática de playlist de canciones mediante técnicas de minería de datos

Miguel Ángel Cantero Vllora
Grado en Ingeniería Informática



- Introducción
- Conjunto de datos
- Modelo de recomendación
- Resultados
- Conclusiones y propuestas




Contenido

Introducción

Introducción

¿Qué es una playlist?



LISTA

Happy Moments

Canciones felices para animar tu día! Feliz día! Happy day!

Creada por Filtr España • 124 canciones, 7 hr 29 min

REPRODUCIR

SEGUIDORES
41.325

Q Filtrar

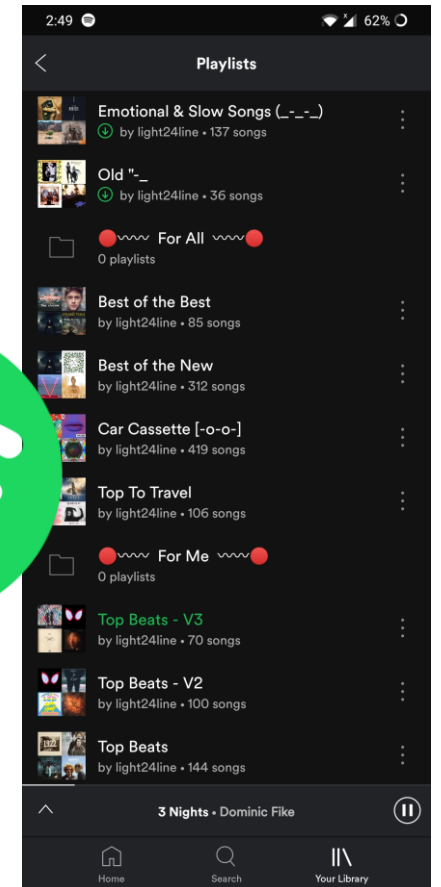
	TÍTULO	ARTISTA	ÁLBUM	
♡	SexyBack (feat. Timbaland) <small>EXPLICIT</small>	Justin Timberl...	FutureSex/Lo...	2019-09-13
♡	Single Ladies (Put a Ring on It)	Beyoncé	I AM...SASHA ...	2019-09-13
♡	Hey Ya!	OutKast	The Way You ...	2019-09-13
♡	Happy	Pharrell Willia...	Despicable M...	2019-09-13
♡	Bring a Little Lovin'	Los Bravos	Dame un Poc...	2019-09-13
♡	I Gotta Feeling	The Black Eye...	The Beginning...	2019-09-13
♡	Summer Days (feat. Mackl... <small>EXPLICIT</small>	Martin Garrix,...	Summer Days ...	2019-09-13
♡	Canyon Moon	Harry Styles	Fine Line	2019-12-13
♡	CAN'T STOP THE FEELING! (Original ...	Justin Timberl...	CAN'T STOP ...	2019-09-13

- Lista de canciones definida por un usuario.
- Acompañada por un título
- Se establece un orden, aunque puede reproducirse de manera aleatoria.
- Los servicios también crean sus propias playlists para ofrecer a los suscriptores.

Introducción

¿En qué consiste nuestro proyecto?

- Construir un sistema de creación o continuación de playlists.
- A partir de un nombre y/o una lista de canciones.
- Nuestro sistema también deberá ser capaz de solventar el problema del “arranque en frío”.



Introducción

Objetivos



Conjunto de datos

Million Playlist Dataset (MPD)



- 2018 ACM RecSys Spotify Playlist Challenge
- *MPD* → 1.000.000 de playlists creadas por usuarios
- *Challenge Dataset* → 10.000 playlists incompletas
- Predecir las canciones que faltan

Million Playlist Dataset (MPD)

- Una playlist se considera válida para *MPD* sí:
 - Creada por un usuario residente en USA y mayor de 13 años.
 - Lista pública.
 - Entre 5 y 250 pistas.
 - Contiene al menos 3 artistas diferentes.
 - Contiene al menos 2 álbumes diferentes.
 - No contiene pistas locales.
 - Tienen al menos un seguidor.

```
{
  "name": "musical",
  "collaborative": "false",
  "pid": 5,
  "modified_at": 1493424000,
  "num_albums": 7,
  "num_tracks": 12,
  "num_followers": 1,
  "num_edits": 2,
  "duration_ms": 2657366,
  "num_artists": 6,
  "tracks": [
    {
      "pos": 0,
      "artist_name": "Degiheugi",
      "track_uri": "spotify:track:7vqa3sDmtEaVJ2gcvxtRiD",
      "artist_uri": "spotify:artist:3V2paBXEoZIAhfZRJmo2jL",
      "track_name": "Finalement",
      "album_uri": "spotify:album:2KrRMJ9z7Xjoz1Az406UML",
      "duration_ms": 166264,
      "album_name": "Dancing Chords and Fireflies"
    },
    // 10 tracks omitted
  ],
}
```

Conjunto de datos

¿Cómo se ha creado?



Búsqueda

Recopilar un conjunto de
playlists



Descarga

Descarga de la
información de las
playlists



Filtrado

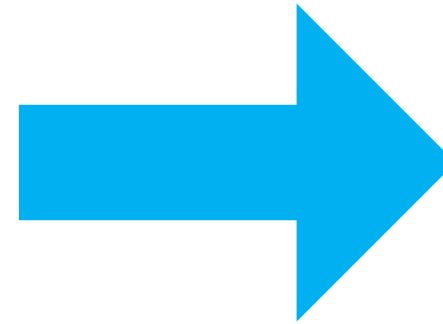
Reducir el número de
playlists a 1 millón



Construcción

Crear conjunto en
formato *JSON*

- 3.000 palabras más comunes (Inglés)
- Géneros (Pop, Rock, RnB, Hip-Hop, ...)
- Períodos (80's, 90's, 00's, ...)
- Actividades (swimming, running, ...)
- Eventos (Halloween, Christmas, ...)
- Estados de ánimo (happy, sad, tired, ...)

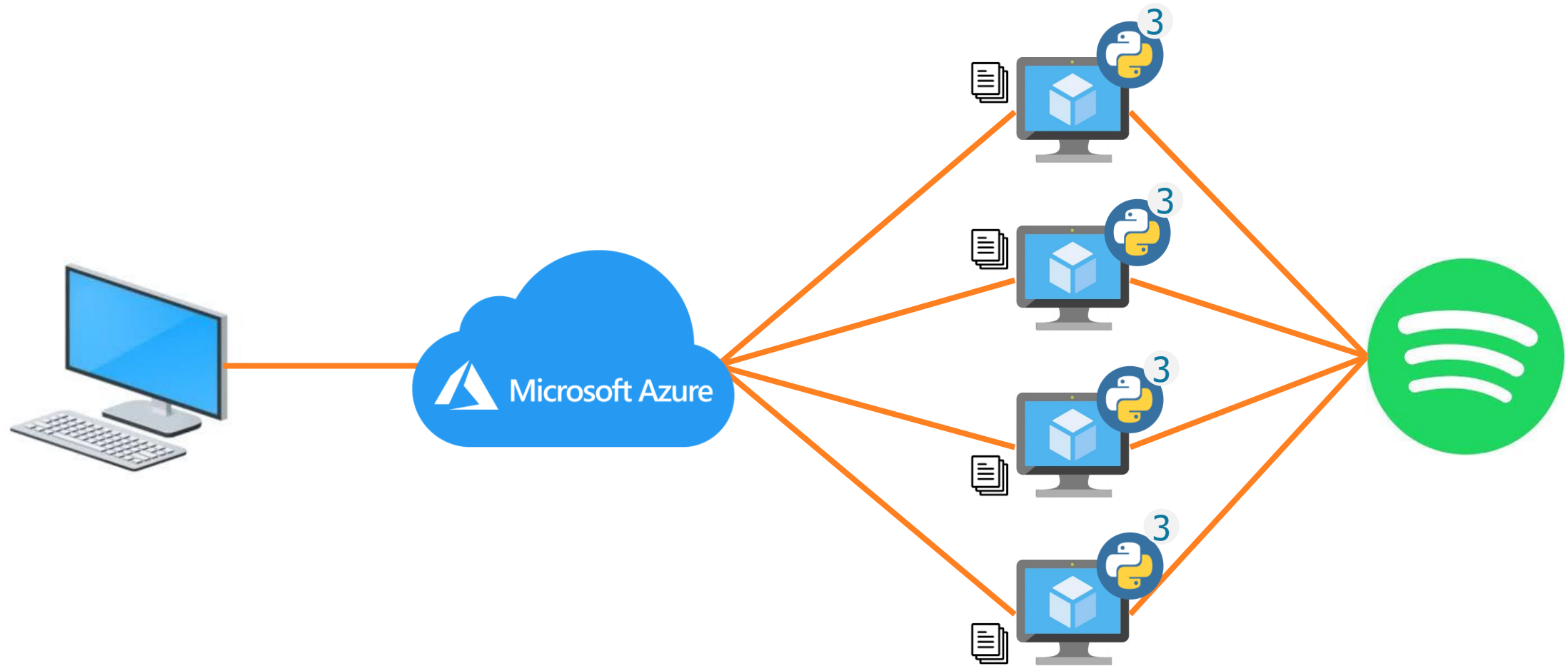




○ *Spotipy*

- Identificador
- Título
- Número de pistas

Búsqueda



Resultados

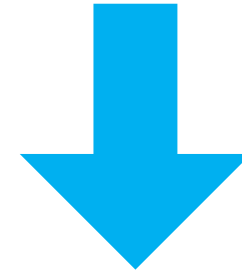
Número de términos buscados	3.154
Playlists obtenidas tras la búsqueda de términos	15.797.992
Tiempo empleado	16,5 horas
Espacio ocupado en memoria	715 MB

Descarga

Preparación

- Eliminación de playlists (mismo identificador *Spotify*).
- Filtrado por número de pistas:
 - [5, 250] pistas
- Filtrado por título:
 - [2, 250] caracteres
 - < 10 palabras

15.797.992

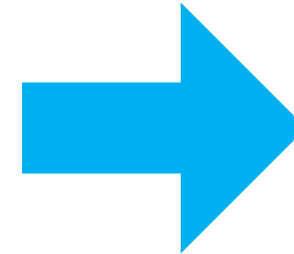


10.394.619

○ *Spotipy*

- Información playlist
- Pistas que contiene

3JwPVKISB9IBlE2RST1MVn
01DbkmjFPYPeZyw7MxBal5
19Hsw3I1EALtwkdimI80UK
26nuxjbRxPo8oguHqcxZaf
605B7FEV5ecYmNP6BKIWbb
4vldFTvc5ckbj4p5Z9s2G2
5HrfgcBGkqEDsUK2Svhqkf
0qE4T7evcQWakZLMX0D4FA
1rmsEzwr6ZmRNzCUph24vZ
28oY1vSsipRE5V0yLDQqed



Requisitos



- La playlist debe tener al menos *1 seguidor*.
- En el momento de descarga, la playlist debe ser *pública*.
- La playlist *no contiene pistas locales*.

Resultados

Playlists descargadas	3.122.640
Tiempo empleado	152 horas (\approx 6'3 días)
Espacio ocupado en memoria	124 GB

1. Número de pistas, artistas y
álbumes.

2. Idénticas características.

3. Número de ediciones.

4. Títulos.

5. Criterios adicionales

	name	num_tracks	num_followers	num_artists	num_albums	duration_ms	num_edits
id							
9hUw5qK0K2GDwH	Love Letter	17	65	16	16	3601989	5
t8mB0CayapCcRr	skin deep	35	1	32	35	7146193	16
QuXhb3Yq024HHw	junior year	250	1	113	184	57370653	113
Q33CvNMaMiZ1uH	BOOM BOOM	20	1	16	20	4120665	2
HLbXILXIhYoxF6	Born in the Wrong Era	168	1	112	146	36941201	86

Grupos

Grupo 1

- Número de pistas comprendido entre 5 y 250.
- 3 artistas diferentes.
- 2 álbumes diferentes.

Grupo 2

- Duración de la playlist.
- Número de canciones.
- Número de artistas diferentes.
- Número de álbumes diferentes.

-- Iguales --

Grupo 3

- Editada, como mínimo, 2 veces

(Una ventana de 2 horas corresponden a una sesión de edición)

Grupos

Grupo 4

- 5 y 50 caracteres (sin espacios en blanco).
- Menos de 10 palabras.
- Presencia de emoticonos:
 - Texto + Emoticonos: Máximo de 10.
 - Emoticonos: Máximo de 4.
- Caracteres pertenecen al alfabeto latino, al conjunto de caracteres comunes, o son emoticonos.
- Idioma (Inglés)
- Contenido ofensivo.

Grupo 5

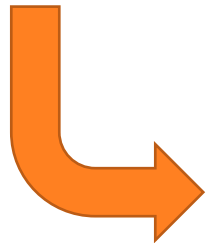
- Número de seguidores inferior a 2.
- Artistas poco frecuentes.

Resultados	
Playlists disponibles antes de aplicar los filtros	3.122.640
Playlists cuyo número de pistas no es valido	8.051
Playlists cuyo número de artistas y álbumes no es valido	159.638
Playlists repetidas	3.772
Playlists editadas menos de 2 veces	183.152
Playlists eliminadas aplicando los filtros para títulos	375.618
Playlists con un número de seguidores inferior a 2	1.117.885
Playlists que contienen artistas poco frecuentes	255.176
Playlists disponibles tras aplicar los filtros	1.019.348

Conjunto de Entrenamiento

1.000.000 playlists

- Formato *JSON*
- 1.000 archivos
- 1.000 playlists / archivo



Canciones



Conjunto de Test

10.000 playlists

- Incompletas
- Categorías:
 1. Playlists dado sólo su título.
 2. Playlists dado su título y la primera pista.
 3. Playlists dado su título y las primeras 5 pistas.
 4. Playlists dadas las primeras 5 pistas (sin título).
 5. Playlists dado su título y las primeras 10 pistas.
 6. Playlists dadas las primeras 10 pistas (sin título).
 7. Playlists dado su título y las primeras 25 pistas.
 8. Playlists dado su título y 25 pistas aleatorias.
 9. Playlists dado su título y las primeras 100 pistas.
 10. Playlists dado su título y 100 pistas aleatorias.



Browse



Radio

Albums

Artists

Local Files

Podcasts

PLAYLISTS

2019 Great alb...

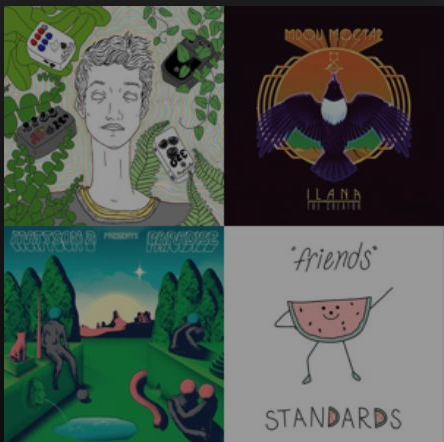
2018 Great Albums

70s Punk

70s Funk



New Playlist



PLAYLIST

2019 Great albums

Created by Sam Moore • 29 songs, 1 hr 45 min

PAUSE



Go to Playlist Radio

Collaborative Playlist

Make Secret

Edit Details

Report

Delete

Create Similar Playlist

Download

Share

FOLLOWERS
0

Download

Filter

TITLE



Take It Away There



Shrugging Match



Yaskool



Chicken Sized Nugget



Salmon Rushdie



Johnny Bravo



Kamane Tarhanin



Asshet Akal

Mdou Moctar

Mdou Moctar

ALBUM



Thank You for Singing ...

2019-07-17

2:03

Thank You for Singing ...

2019-07-17

3:58

Thank You for Singing ...

2019-07-17

3:04

Thank You for Singing ...

2019-07-17

3:04

Thank You for Singing ...

2019-07-17

1:34

Thank You for Singing ...

2019-07-17

3:02

Ilana (The Creator)

13 days ago

5:08

Ilana (The Creator)

13 days ago

4:51

Modelo de recomendación

Sistemas de recomendación

Problema

- Gran cantidad de información.
- Dificultad de elección.
- Auge de servicios como:
 - Comercio electrónico
 - Streaming multimedia
 - Buscadores
 - Agregadores de contenido



Sistemas de recomendación

¿Qué son?

- Sistemas capaces de predecir el grado de preferencia de un usuario para un conjunto de items.
- Ordenados por relevancia para el usuario
- Tipos:
 - Filtrado colaborativo
 - Filtrado basado en contenido
 - Sistemas híbridos

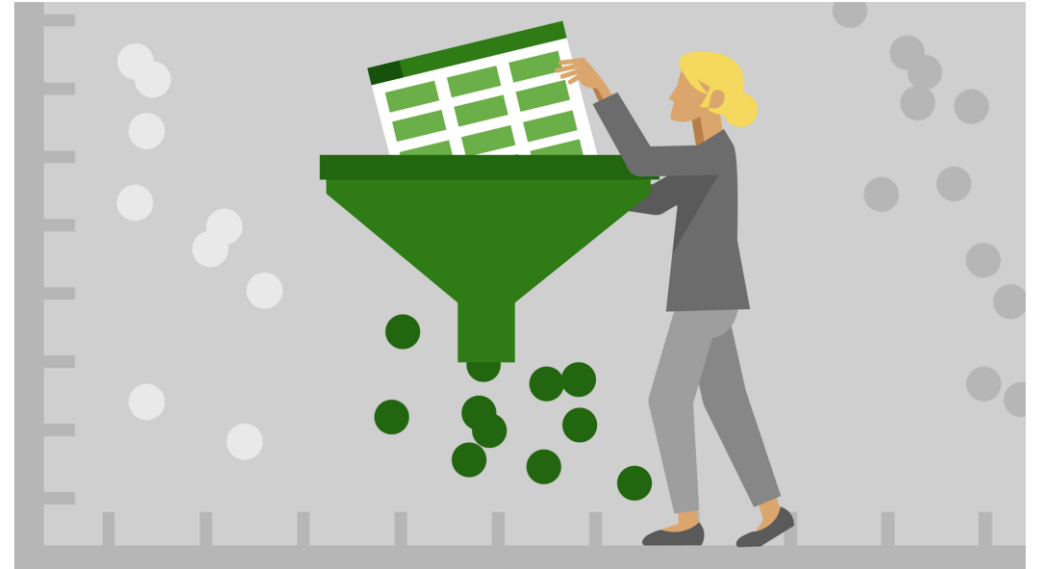




- Implementación en Python
- Modelo híbrido de recomendación
 - Filtrado colaborativo
 - Filtrado basado en contenido
- Soluciona el problema del “arranque en frío”

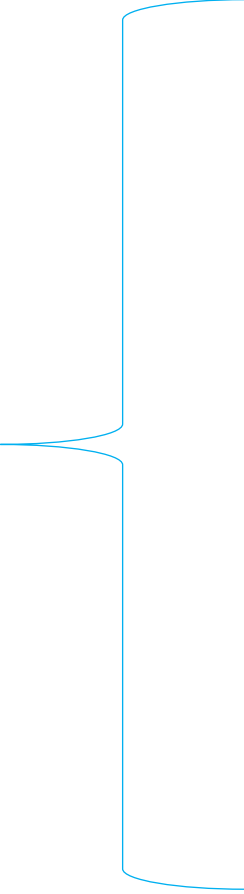
Preprocesamiento

1. Conversión de formato JSON a CSV
2. Creación de características de usuario
(playlists)
3. Creación de matrices de expansión



CSV

1.000 ficheros *JSON*

- 
- Álbumes
 - Artistas
 - Pistas
 - Información sobre las playlists
 - Información sobre las playlists del conjunto de prueba
 - Pistas que conforman las playlists
 - Artistas que conforman las playlists

Preprocesamiento

Creación de características de usuario (playlist)

Títulos → Etiquetas

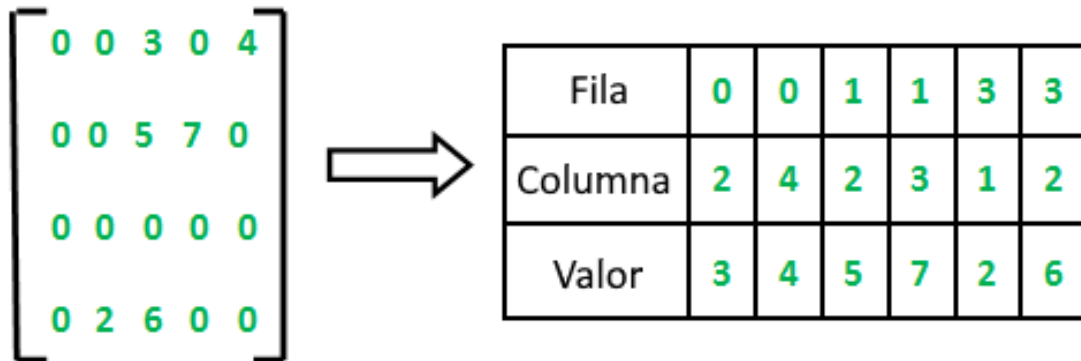
- Eliminación de signos de puntuación y caracteres especiales
- *Stemming* (obtención de la raíz de la palabra)
- Emoticonos a etiquetas



```
'💡': ['study', 'light', 'lit', 'yellow', 'lost']  
'👑': ['king', 'queen', 'girl', 'disney', 'rap']  
'🗣️': ['talk', 'yell', 'bake', 'sing', 'belt']  
'🔒': ['special', 'one', 'free', 'taz', 'hardcastle']  
'🥞': ['breakfast', 'bacon', 'wake', 'bake', 'vibe']  
'🥂': ['party', 'student', 'weekend', 'birthday', 'wine']  
'💛': ['yellow', 'happy', 'love', 'country', 'summer']  
'🖨️': ['office']  
'🏆': ['cool', 'beer']  
'🥗': ['salad', 'yo', 'mama']  
'🌟': ['jewish', 'rock', 'real']
```

	name	tags
pl_pid		
0	Low viscosity vibes	low viscos vibe
1	dalanda 🐉	imagine dragon game dalanda
2	freeze pops	pop freez
3	Golden Oldies	oldi golden

Preprocesamiento

Creación de matrices de expansión



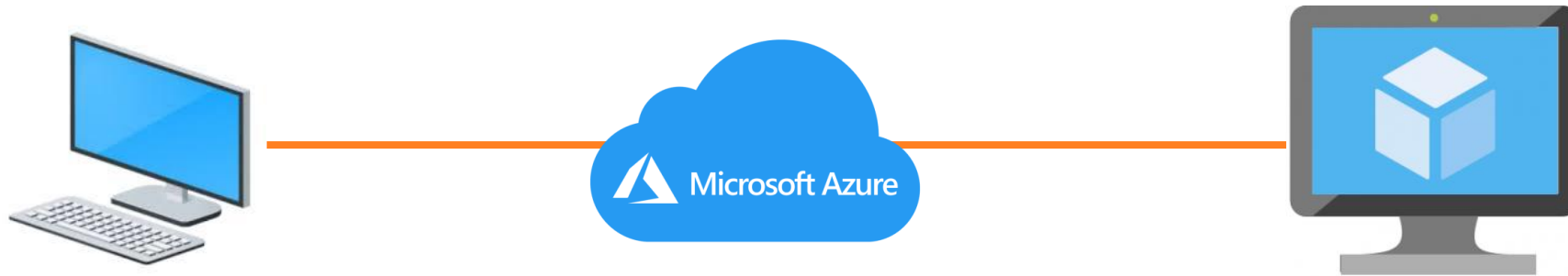
- Interacciones pista / playlists 
- Peso interacciones (en caso de existir)
- Características de usuarios (playlists) 
- Características de items (canciones)

Definición del modelo

- Función de pérdida (*loss*):
 - **WARP** (Weighted Approximate-Rank Pairwise)
 - Interacciones positivas
 - Optimizar la parte superior de la lista de recomendaciones.
- Número máximo de muestras (*max sampled*):
 - 20 → 30 (Mejorar la precisión del modelo)
- Número de componentes (*no components*):
 - 200 (Características para pistas)



Entrenamiento



- 16 Cores
- 32 GB de RAM
- Optimizada para procesos

Resultados	
Ciclos (<i>Epoch</i>)	150
Tiempo empleado	10 horas
Tamaño del modelo	12 GB

Resultados

Experimentos

Canciones similares

"Toxic" – *Britney Spears*

Hollaback Girl	<i>Gwen Stefani</i>
Toxic	<i>Britney Spears</i>
Crazy In Love	<i>Beyoncé</i>
Don't Stop the Music	<i>Rihanna</i>
Single Ladies	<i>Beyoncé</i>
Hollaback Girl	<i>Gwen Stefani</i>
Womanizer	<i>Britney Spears</i>
Don't Cha	<i>Pussycat Dolls</i>
Hips Don't Lie	<i>Shakira</i>

"Cool" – *Alesso*

Love Me Again*	<i>John Newman</i>
When The ... *	<i>Craig David</i>
Remind Me	<i>Jonas Blue</i>
Desire*	<i>Years & Years</i>
With Ever Heartbeat	<i>Robin</i>
Sorry	<i>Justin Bieber</i>
I'm In Control*	<i>AlunaGeorge</i>
Middle	<i>DJ Snake</i>
Love You Better	<i>Aston Powers</i>

"Six Feet Under" – *The Weeknd*

Party Monster	<i>The Weeknd</i>
Reminder	<i>The Weeknd</i>
Ordinary Life	<i>The Weeknd</i>
All I Know	<i>The Weeknd</i>
Acquainted	<i>The Weeknd</i>
Sidewalks	<i>The Weeknd</i>
Often	<i>The Weeknd</i>
Crew Love	<i>Drake</i>
Sameless	<i>The Weeknd</i>
Low Life	<i>Future</i>

Experimentos

Playlist (contenido conocido)

Título: **TEEN**

Canción	Artista
...Baby One More Time	<i>Britney Spears</i>
Sometimes	<i>Britney Spears</i>
Toxic	<i>Britney Spears</i>
Oops!...I Did It Again	<i>Britney Spears</i>
I'm a Slave 4 U	<i>Britney Spears</i>
Genie in a Bottle	<i>Christina Aguilera</i>
Rock Your Body	<i>Justin Timberlake</i>
Cry Me a River	<i>Justin Timberlake</i>
CAN'T STOP THE FEELING!	<i>Justin Timberlake</i>
Mirrors	<i>Justin Timberlake</i>

Canción	Artista
No Scrubs	<i>TLC</i>
Waterfalls	<i>TLC</i>
Kiss from a Rose	<i>Seal</i>
Fast Car	<i>Tracy Chapman</i>
I Want It That Way	<i>Backstreet Boys</i>
If I Ain't Got You	<i>Alicia Keys</i>
Always Be My	<i>Mariah Carey</i>
Baby I Want It That Way	<i>Backstreet Boys</i>
Torn	<i>Natalie Imbruglia</i>
Say My Name	<i>Destiny's Child</i>

Experimentos

Playlist (aleatoria)

Canción	Artista
Somethin' Stupid	<i>Frank Sinatra</i>
Je veux Zaz	<i>Zaz</i>
Les passants	<i>Zaz</i>
Qué vendrá	<i>Zaz</i>
Bonnie And	<i>Brigitte Bardot</i>
Clyde La Javanaise	<i>Serge Gainsbourg</i>
La Madrague	<i>Brigitte Bardot</i>
Comic Strip	<i>Serge Gainsbourg</i>
Je t'aime moi non plus	<i>Serge Gainsbourg</i>
Comic Strip	<i>Serge Gainsbourg</i>

Título: **FAVOURITE FRENCH MUSIC** 

Canción	Artista
La mer	<i>Charles Trenet</i>
La vie en rose	<i>Édith Piaf</i>
Plus bleu que tes yeux	<i>Édith Piaf</i>
Sylvie	<i>Charles Aznavour</i>
Plus bleu que tes yeux	<i>Charles Aznavour</i>
Mon Raymond	<i>Carla Bruni</i>
Chez Keith Et Anita	<i>Carla Bruni</i>
Comment te dire adieu	<i>Françoise Hardy</i>
À quoi ça sert	<i>Françoise Hardy</i>
Etonnez-moi, Benoît	<i>Françoise Hardy</i>

Resultados

	precision@10	precision@20	precision@50	precision@100
<i>Título</i>	1,15 %	2,02 %	3,94 %	6,13 %
<i>Título + 1 Pista</i>	2,51 %	4,21 %	8,02 %	11,92 %
<i>Título + 5 Pistas</i>	3,80 %	6,14 %	11,68 %	17,38 %
<i>5 Pistas</i>	2,66 %	5,22 %	10,62 %	16,46 %
<i>Título + 10 Pistas</i>	4,10 %	6,77 %	13,19 %	19,64 %
<i>10 Pistas</i>	3,74 %	5,78 %	12,83 %	18,91 %
<i>Título + 25 Pistas</i>	4,29 %	7,21 %	13,66 %	20,64 %
<i>Título + 25 Pistas (A)</i>	6,20 %	10,39 %	18,63 %	26,98 %
<i>Título + 100 Pistas</i>	2,73 %	4,66 %	9,59 %	15,70 %
<i>Título + 100 Pistas (A)</i>	5,85 %	9,48 %	17,24 %	25,71 %

Conclusiones y propuestas

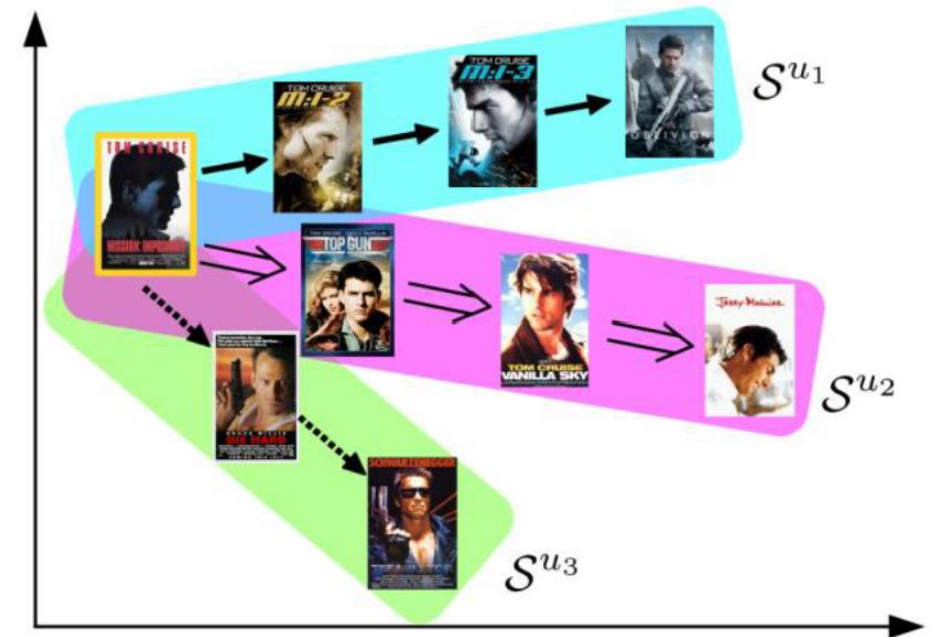
Conclusiones

- Tiempo
 - Conjunto de Datos >> Definición/entrenamiento del modelo
- Dificultad de predecir las canciones exactas
- Nuestro modelo es capaz de predecir canciones relacionadas



Propuestas

- Obtención de hiperparámetros óptimos
- Características para items (canciones)
- Modelo secuencial
- Despliegue de servicio web





Gracias

 Miguel Ángel Cantero Vállora

 miguelangel.cantero@alu.uclm.es