SISTEMA DE EMPAREJAMIENTO

Para estudiantes de la facultad de ciencias, UNAM

Preguntas de investigación

- **¿Qué combinación de métricas de similitud y ponderaciones entre variables produce los emparejamientos más coherentes entre usuarios, considerando sus preferencias?**

- **¿Qué patrones, temas o tonos predominan en los mensajes escritos por los participantes, y qué pueden revelar sobre las motivaciones y expectativas al participar en el formulario?**

Objetivos

- Preprocesar y limpiar el conjunto de datos, asegurando una única respuesta por participante
- Calcular la similitud entre usuarios para cada pregunta del formulario, utilizando la métrica más adecuada según el tipo de variable
- Integrar las distintas matrices de similitud en una sola matriz global mediante la asignación de pesos diferenciados a cada pregunta
- Implementar una estrategia de emparejamiento basada en la minimización de la distancia total ponderada entre usuarios, utilizando variantes del algoritmo de Gale-Shapley
- Analizar de forma exploratoria el texto libre, identificando patrones temáticos, tonos y características lingüísticas recurrentes entre los participantes

Metodología

01 Base de datos

557 registros correspondientes a 545
Para cada usuario se consideraron respuestas a preguntas cerradas sobre actividades, intereses y preferencias, así como un texto libre mínimo de 30 palabras dirigido a su “match” ideal

02 Análisis descriptivo

- Gráficas estadísticas con grupos/preferencias
- Análisis de texto libre:
  - Análisis de frecuencia de términos y nubes de palabras
  - Análisis de tono o sentimiento (positivo, negativo, neutro)
  - Agrupamiento semántico para identificar posibles perfiles de lenguaje (por ejemplo, romántico, humorístico o descriptivo)

Metodología

03 Cálculo de similitud

1. Cálculo de matrices de similitud independientes para cada pregunta, utilizando métricas específicas (por ejemplo, Jaccard para variables categóricas o similitud de coseno para representaciones textuales).
2. 2. Integración de todas las matrices en una sola matriz global mediante un esquema de ponderación que asigna diferentes pesos a las preguntas según su relevancia

Emparejamiento con el Algoritmo de Galey -Shapley

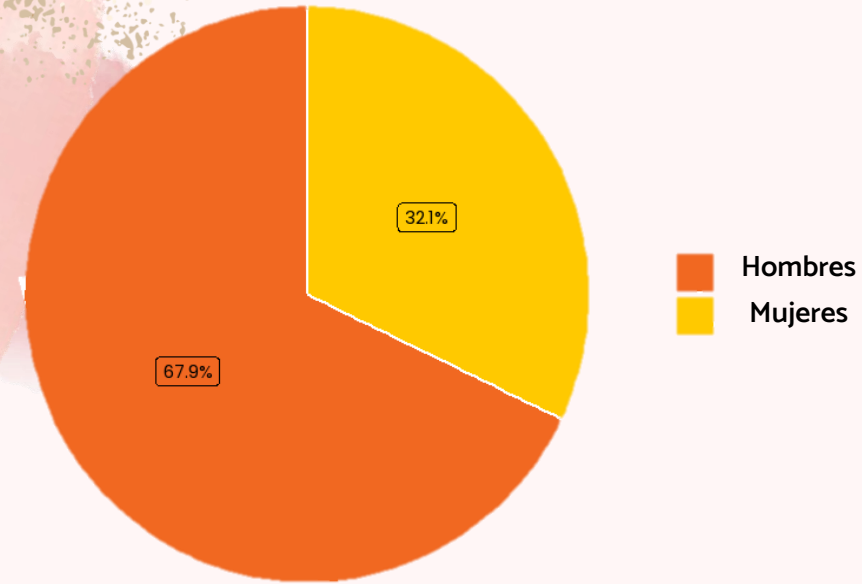
Parejas heterosexuales – Variante de Universidad/Estudiantes

Parejas homosexuales – Variante de Roomies

02

Análisis descriptivo

Género de los participantes

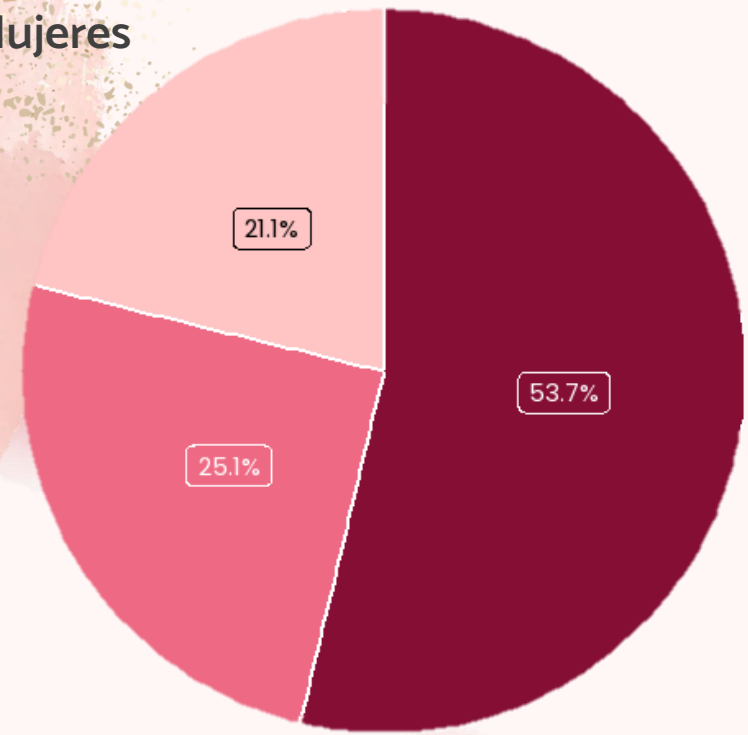


Preferencias de los participantes

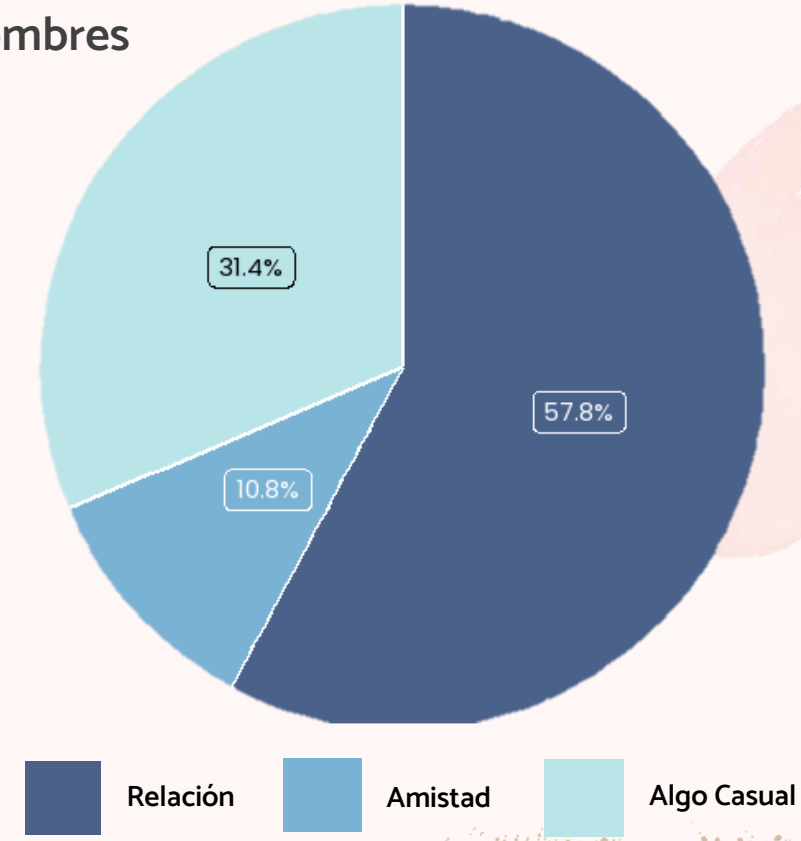
Género	Gustos	Total	Porcentaje
Hombres	Ambos	13	3.5%
Hombres	Hombres	33	8.9%
Hombres	Mujeres	324	87.6%
Mujeres	Ambos	59	33.7%
Mujeres	Hombres	102	58.3%
Mujeres	Mujeres	14	8%

¿Qué buscan los participantes?

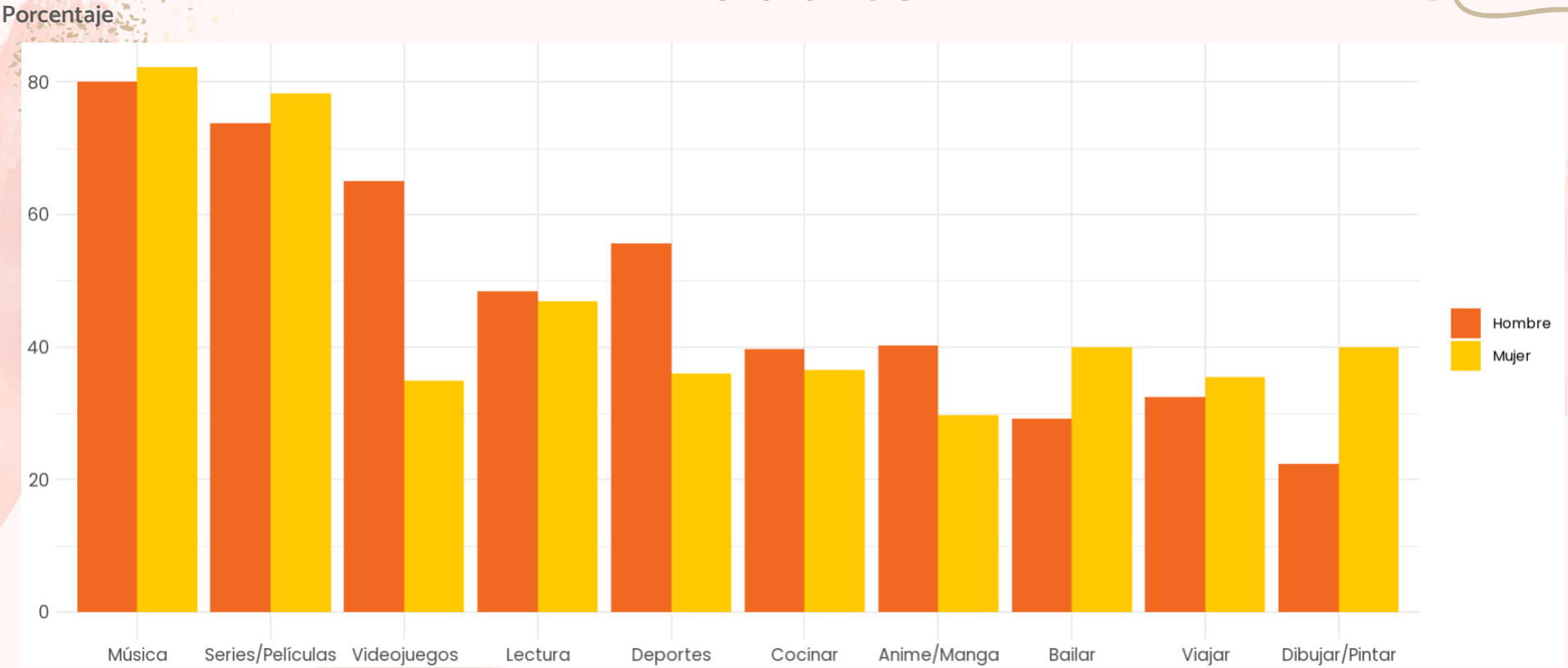
Mujeres



Hombres



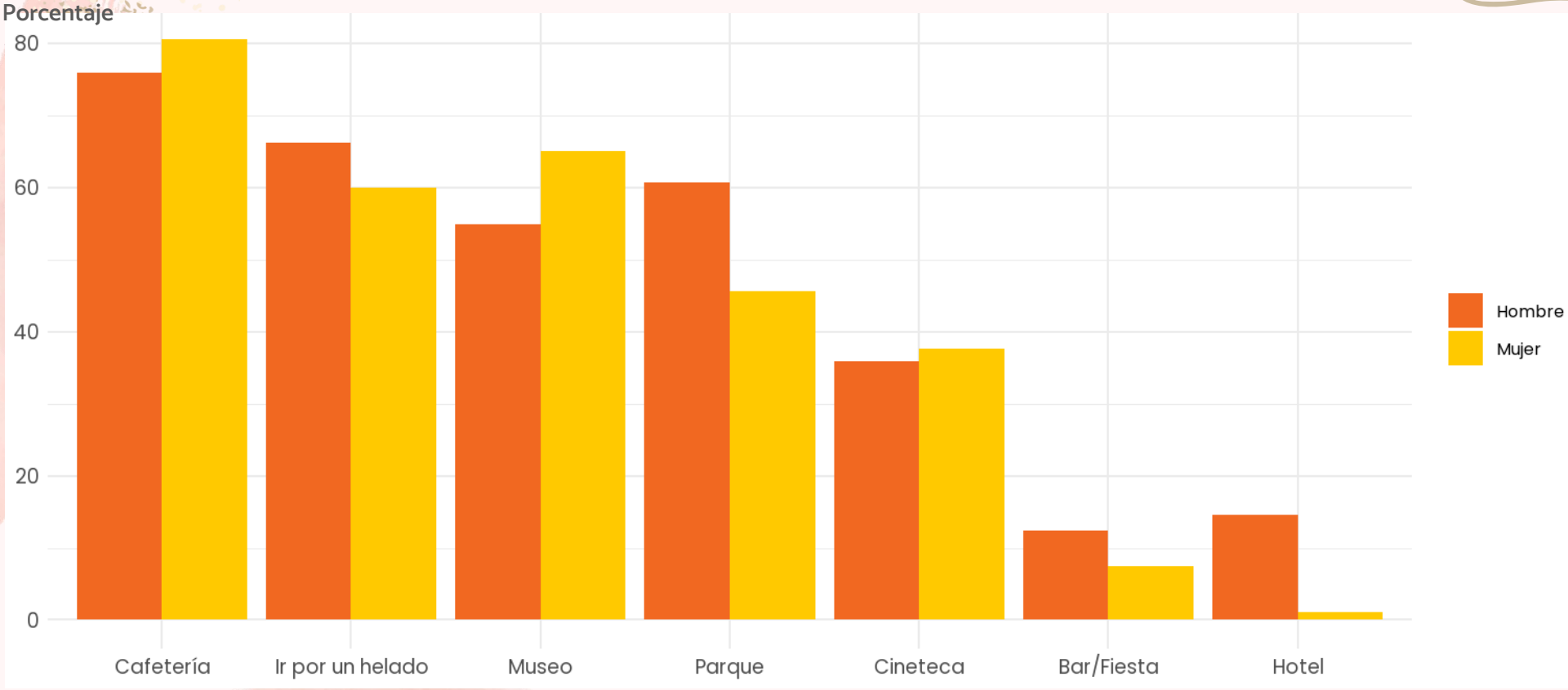
Hobbies



*Se puede seleccionar más de uno

** El porcentaje está calculado respecto al género con el que se identifican los participantes

Lugares favoritos para la primera cita



*Se puede seleccionar más de uno

** El porcentaje está calculado respecto al género con el que se identifican los participantes

Top de palabras más usadas en los comentarios

Analisis de los comentarios

Vectorización de Texto (TF-IDF) Esta técnica asigna un peso a cada palabra basado en su frecuencia dentro del documento y su rareza a través del corpus (TF-IDF), utilizando filtros estrictos (max df=0.95, min df=2).

Se aplicó el algoritmo de Factorización de Matrices No Negativas (NMF). Este modelo descompone la matriz en dos factores (W y H) no negativos, permitiendo identificar patrones. Se definió un total de k=5 temas latentes, de los cuales se extrajeron las 5 palabras más importantes y se obtuvieron los siguientes resultados:

Tema 1: conocer, salir, lugares, nuevos, personas
Tema 2: ciencias, estudio, ir, física, estudiar
Tema 3: busco, nada, hacer, amistad, serio
Tema 4: música, todo, escuchar, cosas, persona
Tema 5: terapia, fui, veterinaria, jajaja, jaja

03

Metodología

Distancia Jaccard

Ana:

- Leer
- Anime
- Cocinar
- Viajar

Luis:

- Cocinar
- Viajar
- Deporte
- Música

Leer	Cocinar	Viajar
Deporte	Música	Anime

Unión = 6

Cocinar	Viajar

Intersección = 2

$$1 - \frac{\#\ Elementos\ en\ la\ intersección}{\#\ Elementos\ en\ la\ unión}$$

Ana y Luis están a 2/3 de distancia

Obtención de matrices de distancias

1. ¿Cómo te identificas? (Hombre/Mujer)
2. **¿Qué te gusta? (Hombres/Mujeres/Ambos)** No negociable
3. **¿Principalmente qué estás buscando?**

Relación: 0.5
Amistad: 0
Casual: -0.5

} Distancia euclidiana

4. **Hobbies**
5. **Para la primera cita prefieres**
6. **Lugares favoritos de la facultad**

} Distancia jaccard

7. Escribe algo a tu match

Matriz de distancia

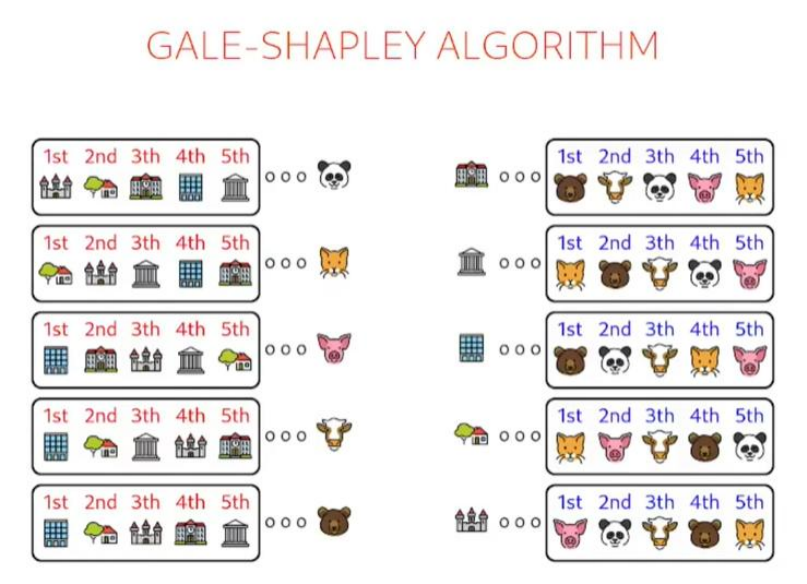
Cada pregunta generó una matriz de nxn de distancias, dentro del rango de 0 a 1, entre los participantes.

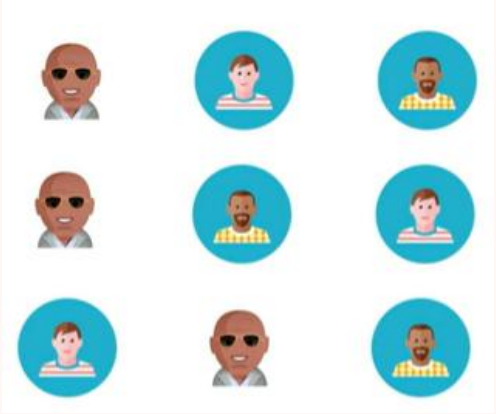
Como todas las matrices tienen las mismas dimensiones se le puede dar un peso distinto a cada pregunta y obtener una sola matriz sumando las entradas:

(pts_busca * busqueda) +
(pts_hobbies * hobbies) +
(pts_primeraCita * primeraCita) +
(pts_lugaresFac * lugaresFac)

Algoritmo de Galey Shapley

Para ejecutar el algoritmo se necesita una lista de preferencias de cada usuario y ésta se obtuvo de acuerdo a las distancias del participante con todos los demás





Modelos y evaluación

Se evaluaron distintos modelos y se conservó el que tenía en promedio menor distancia entre las parejas formadas:

(0.4 * busqueda) +
(0.3 * hobbies) +
(0.25 * primeraCita) +
(0.05 * lugaresFac)

El conjunto se dividió en heterosexuales y homosexuales/lesbianas. A cada mujer heterosexual le correspondieron dos hombres.

Adicionalmente se realizaron otros emparejamientos al azar, y se contrastaron las distancias de estos emparejamientos con los emparejamientos propuestos.

Modelos y evaluación

Distancias resultantes (máximo 1, mínimo 0)

Emparejamiento propuesto

	Lugares para la primera cita	Hobbies	Lugares favoritos de la facultad	Tipo de relación
Mediana	0.7000	0.8333	1.0000	0.500
Media	0.6773	0.8077	0.8966	0.5011

Emparejamiento al azar

Mediana	0.7029	0.8000	0.8889	0.5000
Media	0.6776	0.7833	0.8814	0.4118

Modelos y evaluación

Aunque parece ser que un modelo aleatorio puede obtener mejores resultados por tener distancias más cortas, podemos notar el siguiente problema:

Género p1	Gustos p1	Género p2	Gustos p2	Total
Hombre	Mujeres	Hombre	Mujeres	100
Hombre	Mujeres	Mujer	Hombres	31
Mujer	Hombres	Hombre	Mujeres	22
Hombre	Mujeres	Mujeres	Ambos	21
Mujer	Ambos	Hombre	Mujeres	18
Mujer	Hombres	Mujer	Hombres	12