Proyecto: Sistema de emparejamiento para estudiantes de la UNAM

1 Introducción

El presente proyecto explora cómo los métodos de análisis de similitud pueden aplicarse al diseño de sistemas de emparejamiento basados en afinidades declaradas por los usuarios.

A partir de un conjunto de datos recolectado mediante un formulario respondido por 545 personas de la comunidad universitaria de la UNAM, se busca analizar tanto las respuestas estructuradas (intereses, actividades, tipo de relación deseada) como un campo de texto libre en el que los participantes escribieron un mensaje breve dirigido a su posible pareja ideal.

El objetivo principal es diseñar un sistema que identifique emparejamientos coherentes en función de la similitud entre las características de los usuarios, y, de forma complementaria, explorar el contenido lingüístico de los textos para comprender qué expresan las personas sobre sí mismas o sobre lo que buscan.

2 Preguntas de investigación

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

De forma complementaria: ¿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?

3 Hipótesis

Al utilizar los intereses en común y las preferencias declaradas por los participantes, es posible identificar personas con gustos más afines y, por tanto, con mayor probabilidad de establecer una conexión o relación compatible.

4 Planteamiento del problema

Los sistemas de emparejamiento actuales suelen basarse en modelos de recomendación que dependen de grandes volúmenes de datos históricos. Sin embargo, en contextos controlados y de pequeña escala, puede ser más apropiado recurrir a métodos de similitud directa entre respuestas, sin requerir retroalimentación continua de los usuarios.

El problema central consiste en determinar qué métricas y ponderaciones entre preguntas permiten obtener emparejamientos que reflejen una afinidad real entre los individuos.

5 Objetivos

Objetivo general

Diseñar y evaluar un sistema de emparejamiento entre usuarios utilizando métricas de similitud aplicadas a las respuestas del formulario, integrando variables categóricas y textuales mediante un esquema ponderado.

Objetivos específicos

- 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 (por ejemplo, Jaccard para categóricas).
- 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.

6 Justificación

El proyecto permite integrar herramientas de análisis de datos estructurados y no estructurados dentro de un contexto social y lúdico. Más allá de su aplicación inmediata, el ejercicio contribuye a comprender cómo distintos métodos de similitud afectan el agrupamiento y emparejamiento de individuos.

El estudio es de carácter exploratorio y académico, enfocado en la comparación metodológica más que en la implementación de un sistema funcional a gran escala.

7 Metodología

Descripción general de los datos

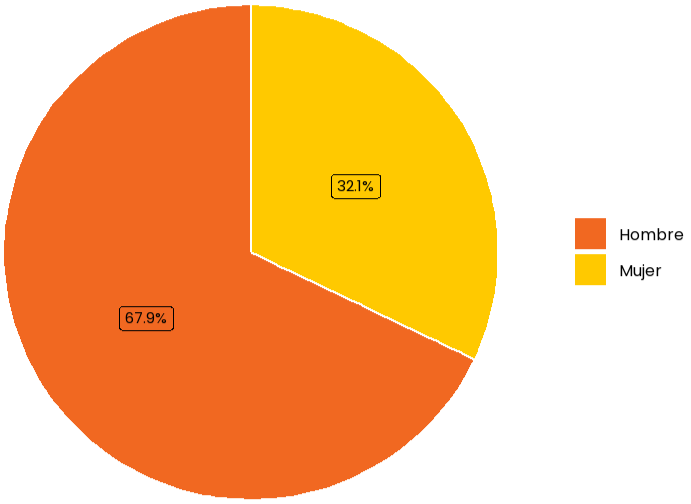
El conjunto consta de 557 registros correspondientes a 545 personas (algunos respondieron más de una vez). Para cada usuario se consideraron las respuestas a las siguientes preguntas:

1. ¿Cómo te identificas? (Hombre/Mujer)
2. ¿Qué te gusta? (Hombres/Mujeres/Ambos)
3. ¿Principalmente qué estás buscando?
4. Hobbies
5. Lugares preferidos/Indiferentes/Malos para la primera cita
6. Lugares favoritos de la facultad
7. Escribe algo para tu match

Todas las preguntas fueron cerradas excepto la última.

Análisis Descriptivo

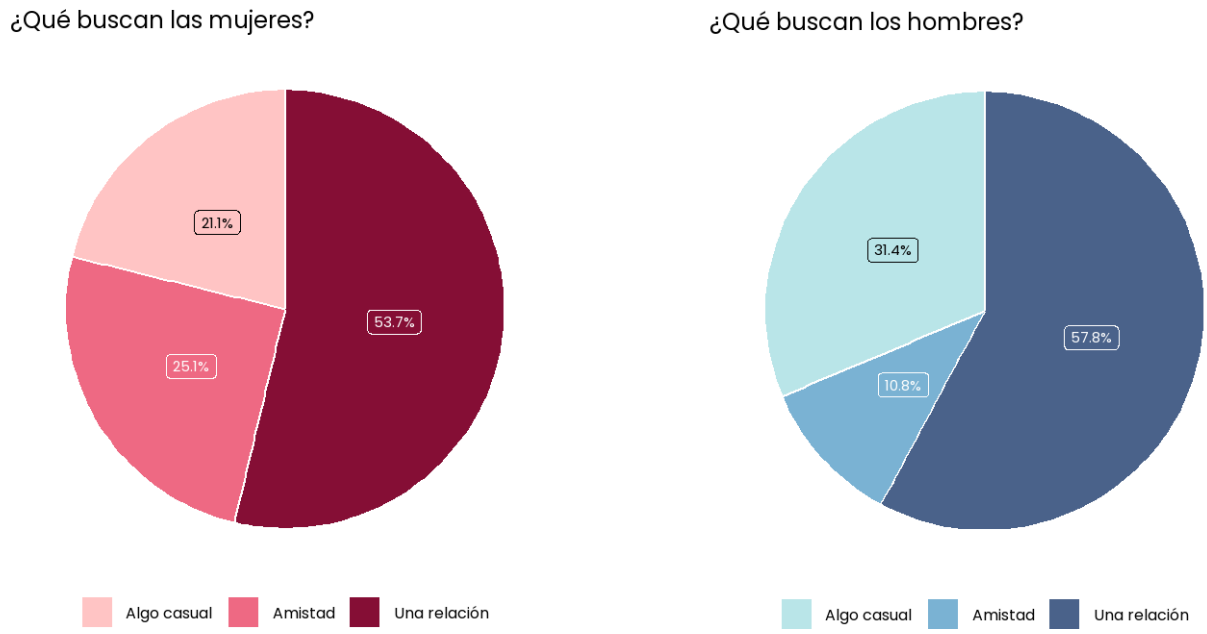
Género de los participantes



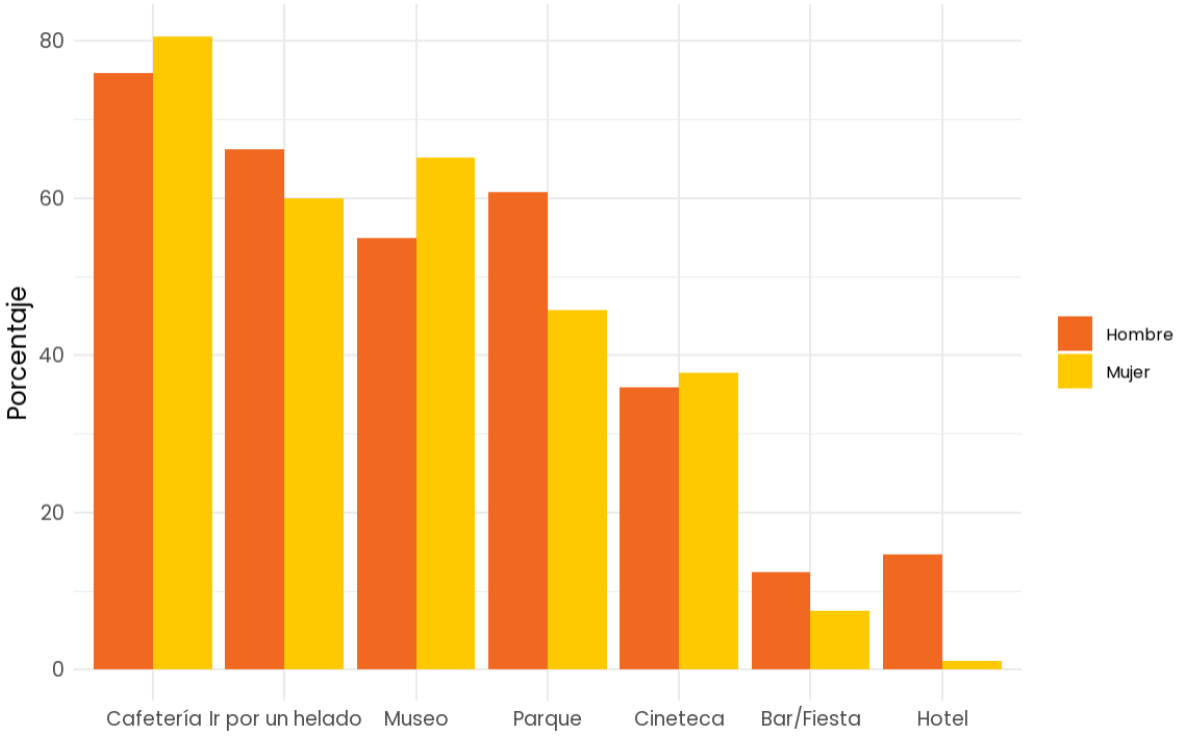
Gustos de los participantes

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

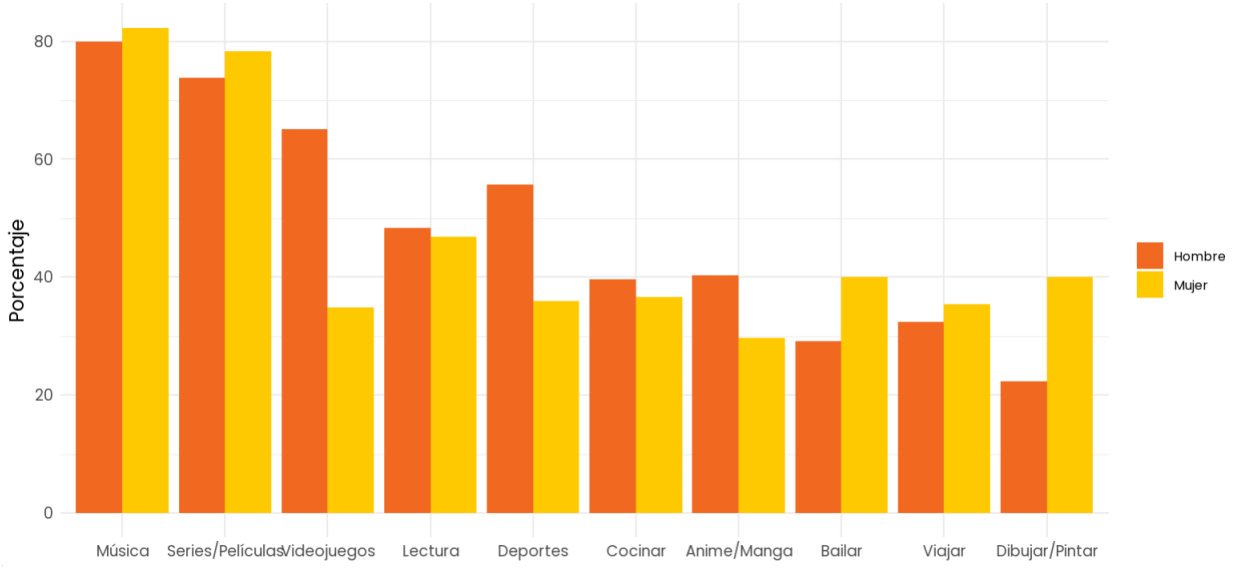
¿Qué están buscando los participantes?



Cuáles son los lugares favoritos para la primera cita



Hobbies preferidos



Top de palabras más usadas en los comentarios

Temas de los comentarios

Para la extracción de los temas latentes a partir de los comentarios, el proceso se basó en el análisis de texto no supervisado, utilizando los siguientes pasos y modelos:

1. Limpieza de los datos: Se quitaron caracteres especiales y el texto se convirtió a minúsculas.
2. Vectorización de Texto (TF-IDF): La columna de comentarios fue transformada en una matriz numérica término-documento utilizando el modelo TfidfVectorizer. 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) para reducir el ruido y enfocarse en los 2000 términos más relevantes.
3. Modelado de Temas (NMF): Sobre la matriz TF-IDF (V), 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

Cálculo de similitud

El proceso se divide en dos etapas:

1. Cálculo de matrices de similitud independientes para cada pregunta, utilizando métricas específicas para cada pregunta (Jaccard para variables categóricas y euclidiana para lo que buscan las personas)
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.

Resultados

Emparejamientos

Para utilizar los algoritmos de emparejamiento se necesita tener una lista de preferencias de las personas; para esto se utilizó la matriz de distancias. Así, para cierta persona x sus preferencias se darán de acuerdo a sus distancias con cada persona, ubicando como sus personas *preferidas* a aquellas con las distancias más cercanas.

Para este paso se evaluaron 10 modelos, en dónde cada uno tenía diferentes pesos para cada pregunta y se conservó aquel que proporcionaba los emparejamientos con menor distancia en promedio.

Modelo	w(busca)	w(hobbies)	w(primera cita)	w(lugares facultad)	Promedio dist
1	0.4	0.3	0.25	0.05	0.6690
2	0.3	0.4	0.25	0.05	0.6797
3	0.25	0.25	0.25	0.25	0.6854
4	0.4	0.4	0.2	0	0.6815
5	0.3	0.3	0.2	0.2	0.6642
6	0.25	0.25	0.15	0.15	0.6748
7	0.5	0.3	0.2	0	0.6711
8	0.45	0.35	0.15	0.05	0.6724
9	0.35	0.35	0.25	0.05	0.6798
10	0.4	0.3	0.15	0.15	0.6719

Table 1: Tabla que muestra el rendimiento de los distintos modelos, donde w representa el peso de la variable.

Evaluación del modelo

Dado que no existe retroalimentación de los usuarios sobre la calidad de los emparejamientos, la evaluación se realizó comparando las distancias resultantes del modelo con las resultantes de un emparejamiento aleatorio.

Distancia	Propuestos	Aleatorios
Búsqueda	0.5011	0.4118
Hobbies	0.8077	0.7833
Primera cita	0.6773	0.6776
Lugares fac	0.8966	0.8814

Table 2: Comparación entre el promedio de las distancias de los emparejamientos propuestos y los aleatorios.

A pesar de que los emparejamiento aleatorios proporcionan distancias menores, cuando observamos los géneros y gustos de los participantes se obtiene lo siguiente.

género p1	gustos p1	género p2	gustos p2	n
Hombre	Mujeres	Hombre	Mujeres	100
Hombre	Mujeres	Mujer	Hombres	31
Mujer	Hombres	Hombre	Mujeres	22
Hombre	Mujeres	Mujer	Ambos	21
Mujer	Ambos	Hombre	Mujeres	18

Table 3: Tabla que muestra algunos emparejamientos, junto con sus gustos y géneros

Trabajo a futuro

- Optimizar los pesos de las variables mediante técnicas automáticas (búsqueda en rejilla, bayesiana o evolutiva) para mejorar la coherencia de los emparejamientos.
- Recolectar retroalimentación de usuarios para evaluar la calidad real de los emparejamientos y validar el modelo con métricas empíricas.