Proyecto: Sistema de emparejamiento basado en afinidad de respuestas y análisis textual exploratorio

1 Introducción

El presente proyecto explora cómo los métodos de análisis de similitud y procesamiento de lenguaje natural 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 Pregunta 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 respuestas a preguntas cerradas sobre actividades, intereses y preferencias, así como un texto libre mínimo de 30 palabras dirigido a su "match" ideal.

Se incluirá un análisis descriptivo de las variables, visualizando la distribución de género, tipo de relación buscada, principales intereses y lugares más mencionados. Además, el texto libre se explorará mediante conteo de frecuencia, nubes de palabras y medidas de longitud promedio.

Cálculo de similitud

El proceso se divide en dos etapas:

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. 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

A partir de la matriz de distancias global, se buscarán combinaciones de usuarios que minimicen la distancia total mediante variantes del algoritmo de Gale-Shapley, garantizando emparejamientos estables y coherentes con las afinidades estimadas.

Análisis del texto libre

Además de su uso dentro del cálculo de similitud, el texto será analizado de manera exploratoria para responder preguntas sobre su contenido y estilo. Se aplicarán las siguientes técnicas:

- Análisis de frecuencia de términos y nubes de palabras.
- Modelado de temas (LDA o BERTopic) para identificar motivos recurrentes.
- Análisis de tono o sentimiento (positivo, negativo, neutro).
- Medición de diversidad léxica, longitud promedio y presencia de pronombres personales.
- Agrupamiento semántico para identificar posibles perfiles de lenguaje (por ejemplo, romántico, humorístico o descriptivo).

Evaluación

Dado que no existe retroalimentación de los usuarios sobre la calidad de los emparejamientos, la evaluación se centrará en la coherencia interna: afinidad de respuestas y consistencia entre las variables más relevantes.

8 Resultados esperados

Se espera que la ponderación adecuada de las preguntas y el uso combinado de métricas de similitud produzcan emparejamientos coherentes entre usuarios con intereses afines. Asimismo, se prevé que las respuestas textuales reflejen patrones de lenguaje y temas comunes, permitiendo caracterizar los principales motivos o actitudes de quienes participaron.

9 Límites del proyecto

- **Muestra cerrada:** los datos provienen exclusivamente de estudiantes de la UNAM, por lo que los resultados no son generalizables a otras poblaciones.
- **Identificación controlada:** cada participante ingresó su correo institucional para evitar duplicados, pero no se validaron características demográficas adicionales.
- **Simulación teórica:** el sistema evalúa emparejamientos en función de similitud de respuestas, sin medir el éxito o satisfacción real de las parejas resultantes.
- **Falta de retroalimentación:** no se recopila información posterior al emparejamiento, por lo que no se ajustan los pesos o métricas con base en resultados observados.
- **Tamaño y extensión del texto:** los mensajes son relativamente breves y el corpus total es limitado, lo que restringe la capacidad de análisis semántico profundo.
- **Alcance computacional:** no se integran modelos de lenguaje avanzados ni aprendizaje supervisado; el análisis es exploratorio y comparativo.