NLP dating

Introduction

В рамках проекта предполагается создание телеграмм бота для знакомств. Приложение будет запрашивать информацию у пользователя (как табличные данные вроде возраста, пола и тд, так и текстовое описание в свободной форме). Далее пользователю будет предоставлен список других юзеров, которые наилучшим образом подходят для знакомства.

Team

Oleg Bobrov, M4150

Related Work

MATCHING THEORY-BASED RECOMMENDER SYSTEMS IN ONLINE DATING [Tomita et Al., 2022] (https://arxiv.org/pdf/2208.11384.pdf)

В статье обсуждаются рекомендательные системы, основанные на теории соответствия в онлайн-знакомствах. В нем освещаются проблемы, с которыми сталкиваются системы взаимных рекомендаций (RRS) при учете взаимных интересов и возможностей пользователей. В статье исследуется применение теории соответствия для решения этих проблем. В документе также описывается текущий проект по развертыванию рекомендательной системы, основанной на теории соответствия (MTRS), на реальной платформе онлайн-знакомств. Обсуждаются проблемы масштабируемости и алгоритмической справедливости в MTRS, а также потенциальные направления будущих исследований.

Данная статья дейтсвительно поднимает важную проблему взаимного сопоставления, однако гипотеза состоит в том, что большая языковая модель способна по предоставленным описаниям учесть потребности обоих сторон приложения знакомств.

I Used Machine Learning NLP on Dating Profiles (https://medium.com/swlh/using-nlp-machine-learning-on-dating-profiles-1d9328484e85)

Автор предоставляет статистику после анализа более 2000 профилей в приложении для знакомств. Можно заметить, что пользователя в своих коротких описаниях действительно указыват как свои потребности в партнере, так и указывают свои личные качества. Эти данные могут позволить решить проблему взаимного

матчинга.

Model Description

Планируется разработать двухуровневую рекоммендательную систему - на первом этапе будут отбираться кандидаты при помощи поиска наиболее релевантных юзеров по текстовому описанию. На второй стадии данный список будет ранжирован с учетом различных признаков и агрегатов из анкет. В финальной выдаче списки будут отфильтрованы на основаниии табличных запросов пользователя. Для получения текстовых эмбедингов будем использовать предобученный BERT, будем искать ближайших соседей с помощью ANN поиска и отдавать полученый списко id кандидатов. По полученым id достанем признаки для ранжирования и отранжируем кандидатов с помощью CatBoost - затем отдадим полученную выдачу пользователю.

Dataset

Предпологается использовать уже предобученную модель, которая хорошо умеет суммаризировать тексты - тем самым мы получим эмбединги, которые содержат много полезной информации о пользователе. На этапе же ранжирования на основе большего числа признаков от юзера мы сможем сформировать более релевантную выдачу с учетом всех подробностей из анкеты.

Experiments

TODO

Results

TODO