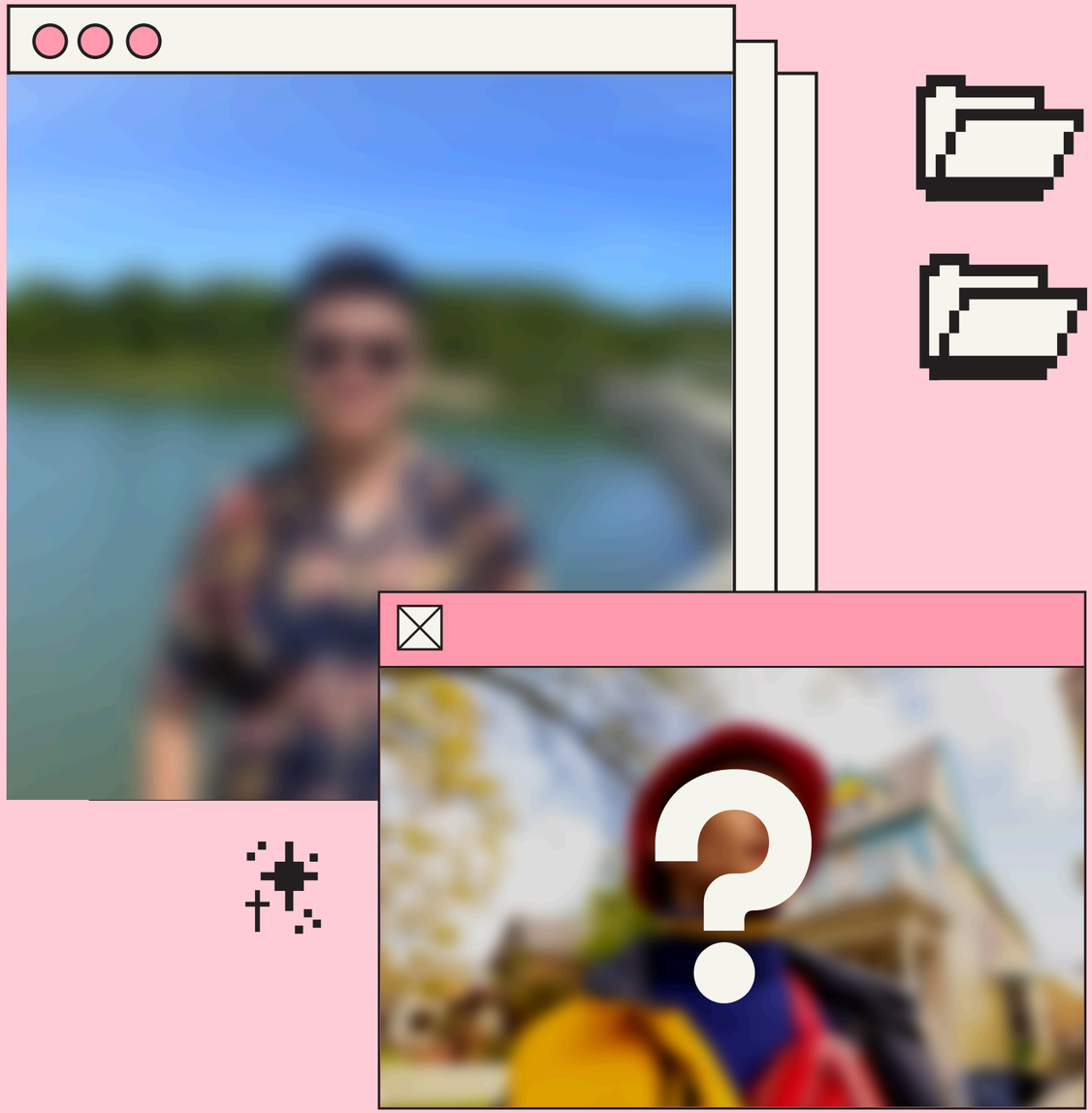
HAPPY Valentine's DAY

by Omakase:D


Damn Single Dog



Find Me a Date

A Dating Recommendation System

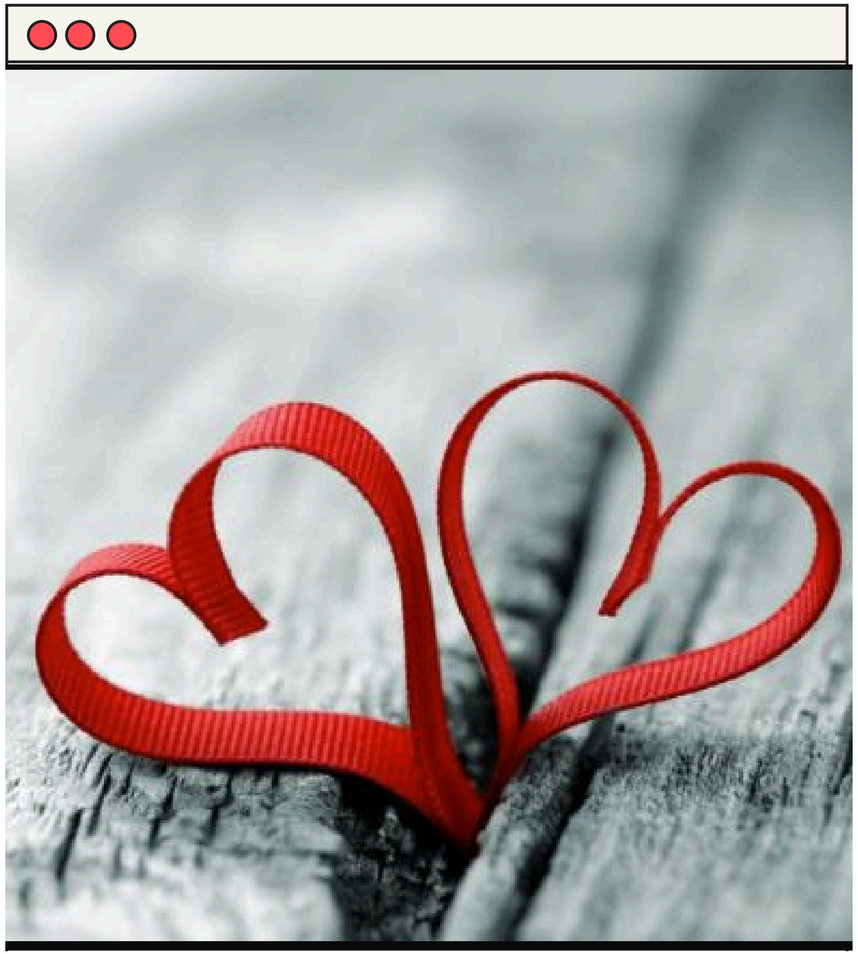
by Omakase:D

Background

- Similarity-Attraction Theory
  - Ellen Berscheid, Elaine Hatfield, 1969
  - Like People Who Are Sililar With Us
- Why Similar? Implies Similiar Intrinsic Values!

- Everyone Needs Love
- People are Looking for the Half One

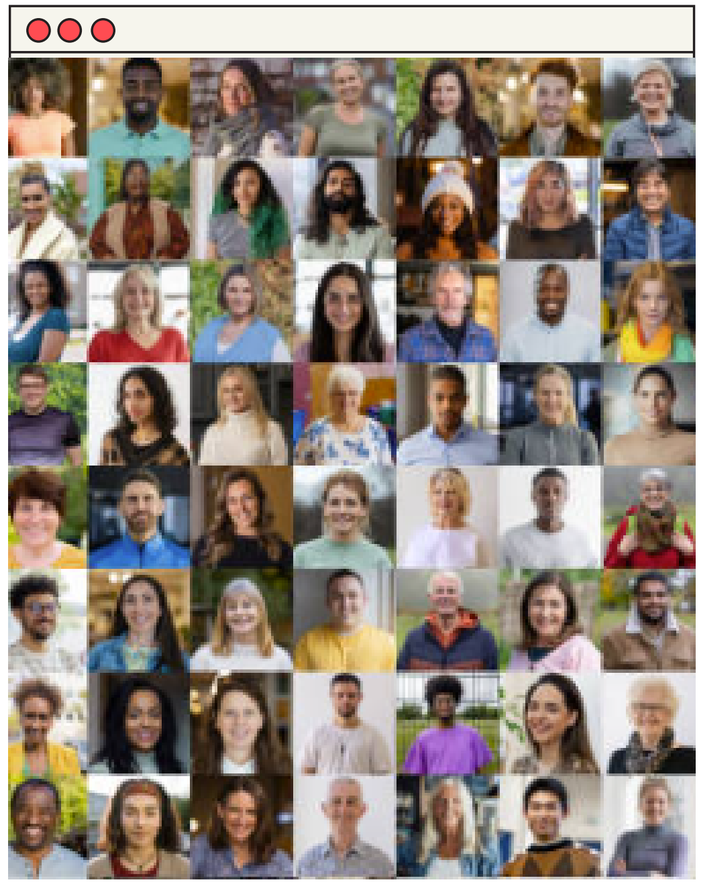
- Develop Dating Recommendation System
  - Pairing Most Similiar Candidate



Introduction

- Data Description:
  - User Profile Data in Dating App (OKCoupid)
  - Source: Kaggle
  - ~60,000(59,946) Row of Records with 31 Columns
    - Demographic & 10 Essay Questions About Themselves
  - Mainly US Users
  - Dataset Created Around 2020

- Procedure
  - Data Cleansing
  - Transformers (all-MiniLM-L6-v2)
    - Sentence Embedding (Text data)
    - Standardizing (Non-text data)
  - Similarity
  - Deployment



Data Cleansing

- Fill NA
- Remove Url in Essay Questions
- Use KNN to Fill
  - Encode to Numerical Format
  - KNN = 5
  - Convert Back to Original Values

```python
# col to fillna: body_type, education, ethnicity, job
# Function to fill missing values with meaningful text
def fill_missing_values(df):
    df = df.copy()
    df['diet'] = df['diet'].fillna("anything")
    df['drinks'] = df['drinks'].fillna("not at all")
    df['drugs'] = df['drugs'].fillna("never")
    df['height'] = df['height'].fillna(df['height'].median())
    df['job'] = df['job'].fillna('other')
    df['offspring'] = df['offspring'].fillna("no kids and neutral to kids")
    df['pets'] = df['pets'].fillna("no pets and neutral to pets")
    df['religion'] = df['religion'].fillna("irreligion")
    df['sign'] = df['sign'].fillna("unknown zodiac sign")
    df['smokes'] = df['smokes'].fillna("no")
    df['speaks'] = df['speaks'].fillna("english")
    df['essay_all'] = df.loc[:,"essay0":"essay9"].apply(lambda x: ','.join(x.astype(str)),axis=1)
    return df.drop(columns=['essay0', 'essay1', 'essay2', 'essay3', 'essay4', 'essay5', 'essay6',
       'essay7', 'essay8', 'essay9'])
```

✓ 0.0s

```python
def fill_missing_values_knn(df, col_fill, list_col_ref, list_col_num):
    df = df.copy()

    # Encode categorical columns
    label_encoder = LabelEncoder()

    # Encode body_type column
    df_non_ (variable) df_non_null: Any
    df.loc[df_non_null.index, col_fill +'_encoded'] = label_encoder.fit_transform(df_non_null)

    # Fill missing values with NaN
    df[col_fill + '_encoded'] = df[col_fill + '_encoded'].astype(float)
    df.loc[df[col_fill].isnull(), col_fill + '_encoded'] = np.nan

    print("Label Encoding Mapping: ", dict(zip(label_encoder.classes_, label_encoder.transform(label_encoder.classes_))))


    # Convert categorical variables (sex, drinks, diet) into numerical format
    for col in list_col_ref:
        df[col + '_encoded'] = LabelEncoder().fit_transform(df[col].astype(str))  # Encode as numbers

    # Select numerical features for KNN
    knn_features = list_col_num + [col + '_encoded' for col in list_col_ref] + [col_fill+'_encoded']

    # Ensure only numerical columns are passed to KNN
    df_knn = df[knn_features]

    # Initialize KNN Imputer with 5 neighbors
    imputer = KNNImputer(n_neighbors=5, weights='distance')
    df_knn_imputed = imputer.fit_transform(df_knn)  # Impute missing values

    # Replace original dataframe values with imputed values
    df[knn_features] = df_knn_imputed

    # Round to nearest integer before inverse transforming
    df[col_fill + '_encoded'] = df[col_fill + '_encoded'].round()

    # Convert back to original body_type values
    df[col_fill + '_imputed'] = label_encoder.inverse_transform(d

    df[col_fill] = df[col_fill + '_imputed']
    df = df.drop(columns=[col_fill + '_encoded', col_fill + '_imp

    return df

def fill_missing_values_knn_multi_col(df, col_fillna):
    df = df.copy()
    for col_fill, list_col_ref in col_fillna.items():
        df = fill_missing_values_knn(df, col_fill, list_col_ref[0
    return df
```

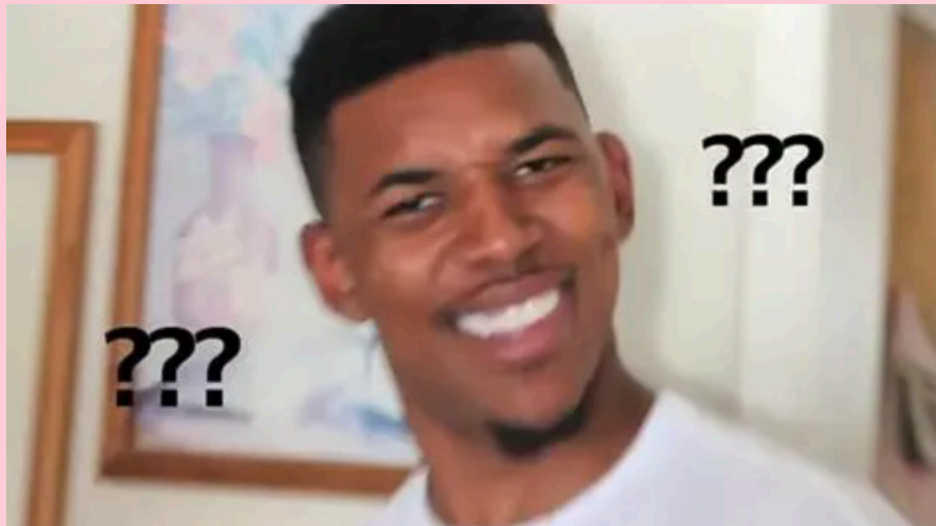
0.0s

```python
# Function to extract only valid URLs from text
def extract_urls(text):
    if not isinstance(text, str):
        return []
    # extract URLs
    url_pattern = r'\b(?=https?:\/\/|www\.[-a-zA-Z0-9@
    # Extract matches
    matches = re.findall(url_pattern, text)
    return matches

def remove_url(row):
    temp = row['essay_all']
    for url in row['url']:
        temp = temp.replace(url, '')
    return temp

def remove_url_col(df):
    df = df.copy()
    df['url'] = df['essay_all'].apply(extract_urls)
    df['essay_all'] = df.apply(remove_url, axis=1)
    return df
```

EDA: Shown in Tableau

What is

Embedding? Similarity?

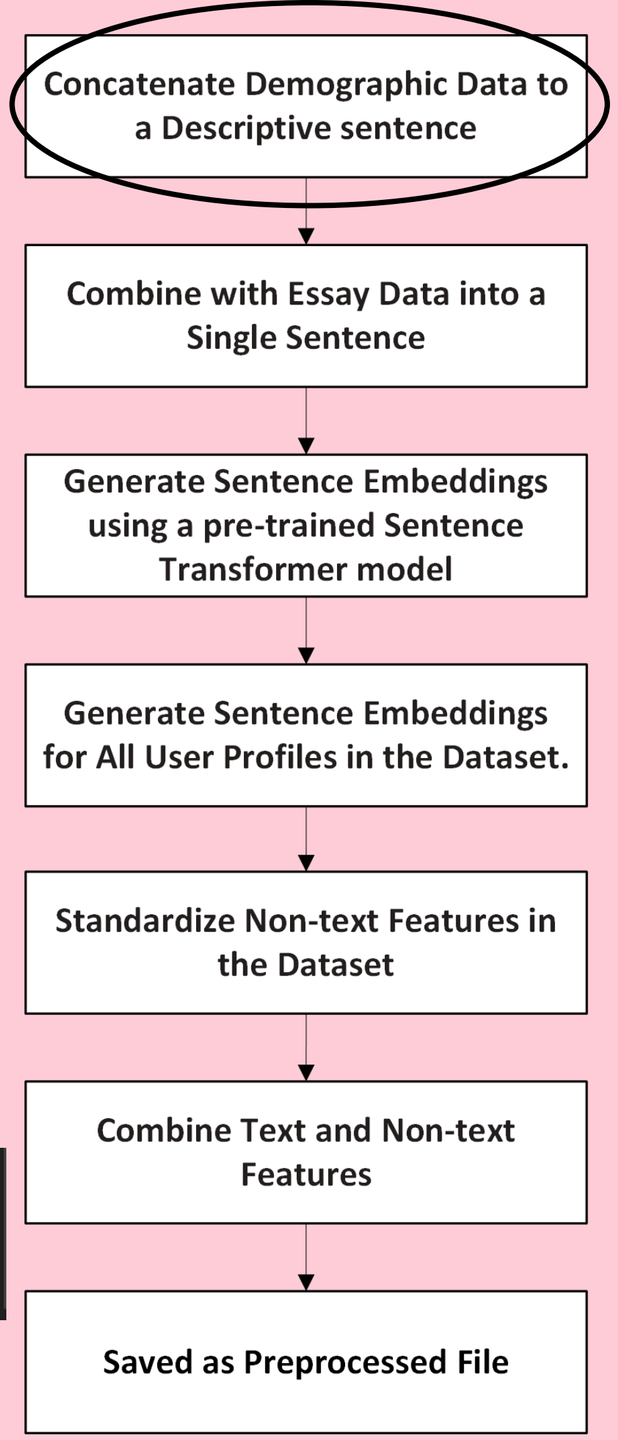


Embedding

22.0, single, male, straight, a little extra, strictly anything.....

Male, single, living in south sanfrancisco, califirnia, sexual orientation is straight.....

age	status	sex	orientation	body_type	diet	drinks	drugs	education	ethnicity	height	job	location	offspring	pets	religion	sign	smokes	speaks	ess
22.0	single	male	straight	a little extra	strictly anything	socially	never	working on college/university	asian, white	190.0	transportation	south san francisco, california	doesn't have kids, but might want	likes dogs and likes cats	agnosticism and very serious about it	gemini	sometimes	english	lc



Embedding

Male, single, living in south sanfrancisco, califirnia, sexual orientation is straight.....

Male, single, living in south sanfrancisco, califirnia, sexual orientation is straight.....About me: i would love to think that i was some.....

age	status	sex	orientation	body_type	diet	drinks	drugs	education	ethnicity	height	job	location	offspring	pets	religion	sign	smokes	speaks	ess
22.0	single	male	straight	a little extra	strictly anything	socially	never	working on college/university	asian, white	190.0	transportation	south san francisco, california	doesn't have kids, but might want	likes dogs and likes cats	agnosticism and very serious about it	gemini	sometimes	english	lc

Concatenate Demographic Data to a Descriptive sentence

Combine with Essay Data into a Single Sentence

Generate Sentence Embeddings using a pre-trained Sentence Transformer model

Generate Sentence Embeddings for All User Profiles in the Dataset.

Standardize Non-text Features in the Dataset

Combine Text and Non-text Features

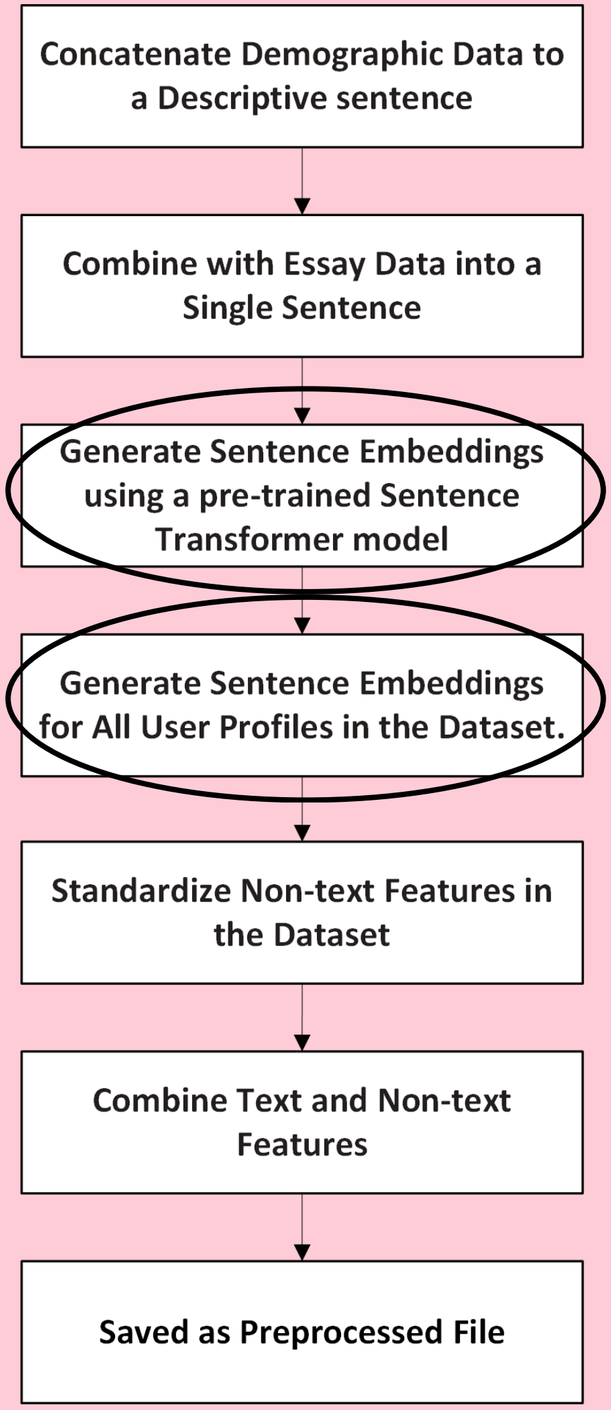
Saved as Preprocessed File

Embedding

sexual orientation is straight.....About me: i would love to think that i was some.....

```python
sbert_model = SentenceTransformer("all-MiniLM-L6-v2")
```

```python
def sentence_embeddings(texts, model):
    """
    Generate sentence embeddings using a pre-trained Sentence Transformer model.
    """
    embeddings = model.encode(texts, show_progress_bar=False)
    return embeddings

def generate_embeddings(df, model):
    """
    Generate sentence embeddings for all user profiles in the dataset.
    """
    # Combine demographic data and essay data into a single sentence
    sentences = df.apply(combine_demographics_essay, axis=1).tolist()

    # Generate sentence embeddings
    embeddings = sentence_embeddings(sentences, model)

    return embeddings
```

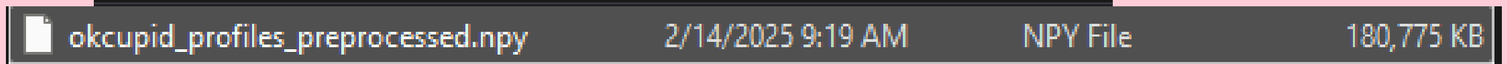


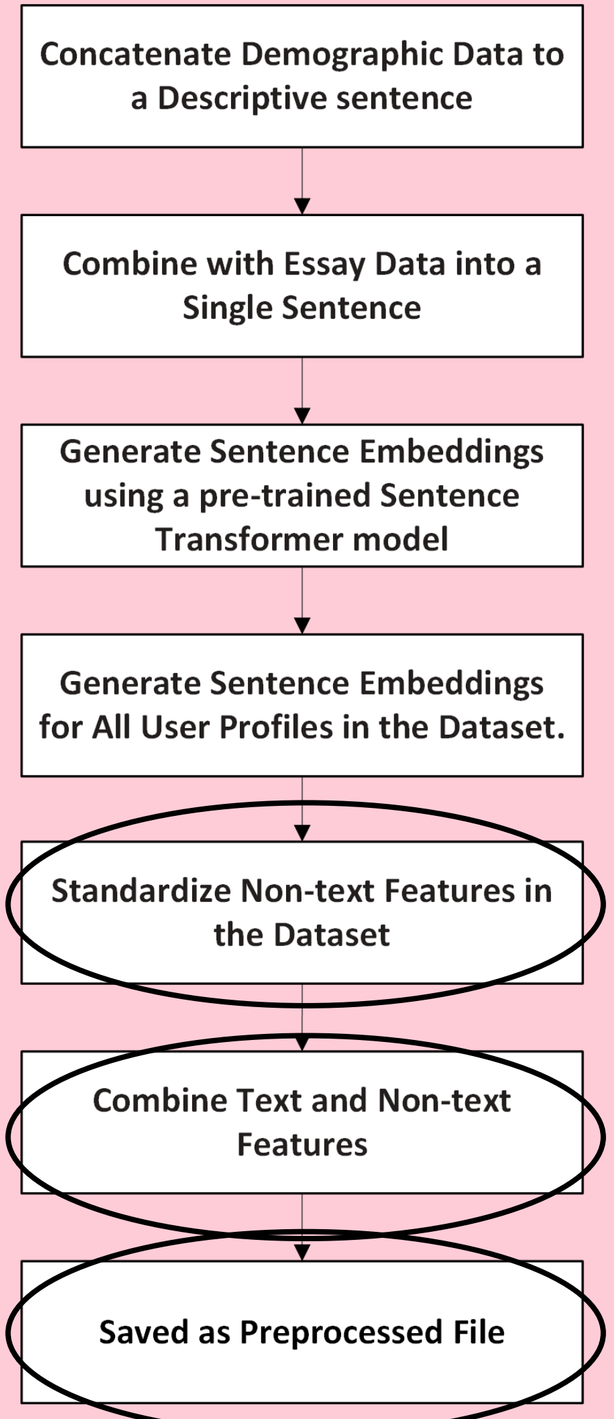
Embedding

```python
def standardize_numeric(df):
    """
    Standardize non-text features in the dataset.
    """
    df = df.copy()
    for col in df.columns:
        if (df[col].dtype == 'float64') or (df[col].dtype == 'int64'):
            df[col] = (df[col] - df[col].median()) / df[col].std()

    return df
```

```python
def preprocess_data(df, model):
    """
    Preprocess the dataset by generating embeddings and standardizing non-text features.
    """
    # Generate embeddings
    embeddings = generate_embeddings(df, model)

    # Standardize non-text features
    non_text_features = df.select_dtypes(include=['float64', 'int64'])
    non_text_features = standardize_numeric(non_text_features)

    # Combine text and non-text features
    X = np.concatenate([embeddings, non_text_features], axis=1)

    return X
```

```python
# Save preprocessed data to a file.
np.save("okcupid_profiles_preprocessed.npy", preprocess_data)
```





Similarity

- Use “cosine_similarity” from Sklearn
- Generate Similarity Matrix

```python
def generate_similarity_matrix(embeddings, df):
    """
    Generate similarity matrix from data
    :param data: pandas DataFrame
    :return: similarity matrix
    """
    user_ids = df.index

    # Assume user_embeddings is an (N, D) matrix, where:
    # - N = number of users
    # - D = embedding dimension

    similarity_matrix = cosine_similarity(embeddings)

    # Convert to DataFrame for easier lookup
    similarity_df = pd.DataFrame(similarity_matrix, index=user_ids, columns=user_ids)

    return similarity_df
```

Similarity

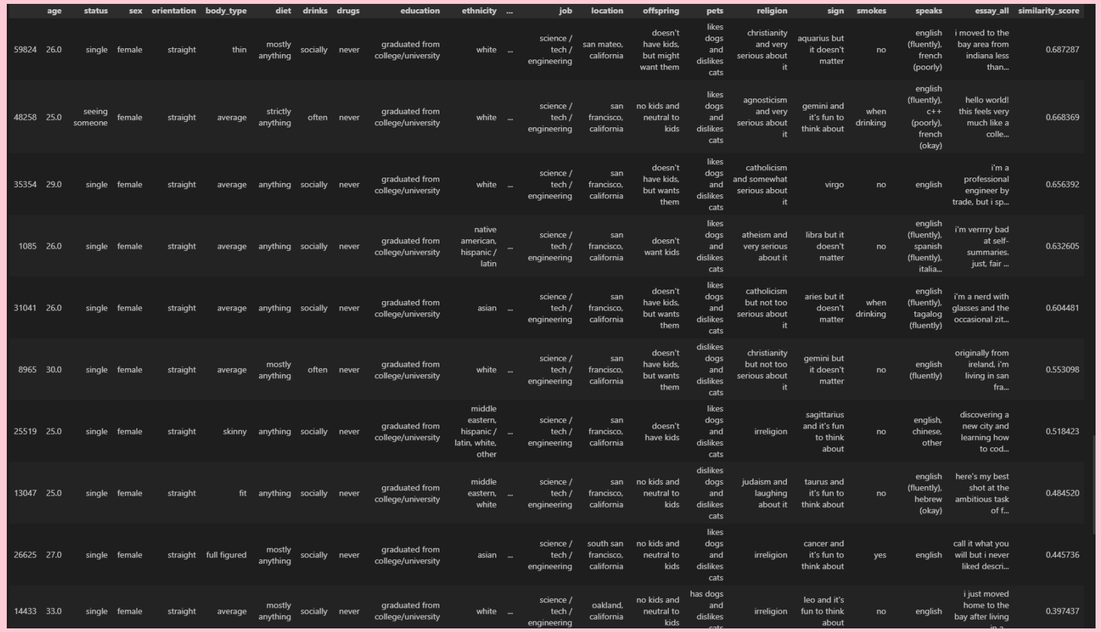
- Min -1: Perfectly Unsimilar
- 0: No Correlation Between 2 Persons
- Max 1: Perfectly Similar

```python
def generate_similarity_matrix(embeddings, df):
    """
    Generate similarity matrix from data
    :param data: pandas DataFrame
    :return: similarity matrix
    """
    user_ids = df.index

    # Assume user_embeddings is an (N, D) matrix, where:
    # - N = number of users
    # - D = embedding dimension

    similarity_matrix = cosine_similarity(embeddings)

    # Convert to DataFrame for easier lookup
    similarity_df = pd.DataFrame(similarity_matrix, index=user_ids, columns=user_ids)

    return similarity_df
```

	0	1	2	3	4	5	6	7	8	9	...	59936	59937	59938	59939	59940	59941	59942	59943	59944	59945
0	1.000000	0.402862	-0.057132	0.744822	-0.047582	0.092371	-0.301606	-0.252902	0.244712	-0.422904	...	-0.419586	0.351081	0.203758	-0.089630	-0.512311	-0.603729	0.881341	0.190992	0.884640	-0.086727
1	0.402862	1.000000	0.673523	0.251320	0.316383	0.394935	0.275800	0.265434	0.174826	0.435308	...	-0.190409	0.733489	0.785232	0.724801	0.024253	0.365281	0.475036	0.785896	0.563800	0.655437
2	-0.057132	0.673523	1.000000	-0.102936	0.396631	0.331526	0.481744	0.449703	0.024536	0.700343	...	0.053667	0.657962	0.703570	0.771470	0.311410	0.669914	0.038552	0.722382	0.134366	0.732014
3	0.744822	0.251320	-0.102936	1.000000	0.218367	0.224549	-0.119834	-0.024107	0.451047	-0.305540	...	-0.149203	0.285685	-0.026792	-0.248943	-0.252008	-0.562506	0.754436	0.008599	0.678194	-0.155555
4	-0.047582	0.316383	0.396631	0.218367	1.000000	0.669479	0.760354	0.813769	0.619304	0.650322	...	0.672918	0.500107	0.004472	0.088809	0.725776	0.290627	0.187130	0.077459	0.076768	0.283604
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
59941	-0.603729	0.365281	0.669914	-0.562506	0.290627	0.197476	0.564641	0.486804	-0.179641	0.799171	...	0.230997	0.268278	0.533709	0.762184	0.525000	1.000000	-0.513655	0.531318	-0.381359	0.696514
59942	0.881341	0.475036	0.038552	0.754436	0.187130	0.273466	-0.109767	-0.020456	0.379317	-0.215472	...	-0.230865	0.445759	0.211063	-0.064792	-0.304576	-0.513655	1.000000	0.223447	0.843424	-0.023088
59943	0.190992	0.785896	0.722382	0.008599	0.077459	0.130709	0.150237	0.135439	-0.196956	0.402727	...	-0.343759	0.539455	0.905891	0.867957	-0.064439	0.531318	0.223447	1.000000	0.379961	0.673854
59944	0.884640	0.563800	0.134366	0.678194	0.076768	0.197924	-0.171569	-0.148365	0.215013	-0.225393	...	-0.369054	0.503057	0.376140	0.122412	-0.403015	-0.381359	0.843424	0.379961	1.000000	0.122001
59945	-0.086727	0.655437	0.732014	-0.155555	0.283604	0.399724	0.428668	0.338667	-0.062797	0.636320	...	-0.021862	0.516881	0.702967	0.783893	0.235510	0.696514	-0.023088	0.673854	0.122001	1.000000

Similarity - Example

- User 12345 With His Requirements:
  - Age: 25-35
  - Height Range: 160-180
  - Lived in California
  - Possess University Degree
  - Work as Engineer
  - Like Dogs, Dislike Cats
- Generate top 10 Results With Highest Similarity



	age	status	sex	orientation	body_type	diet	drinks	drugs	education	ethnicity	...	job	location	offspring	pets	religion	sign	smokes	speaks	essay_all	similarity_score
59824	26.0	single	female	straight	thin	mostly anything	socially	never	graduated from college/university	white	...	science / tech / engineering	san mateo, california	doesn't have kids, but might want them	likes dogs and dislikes cats	christianity and very serious about it	aquarius but it doesn't matter	no	english (fluently), french (poorly)	i moved to the bay area from indiana less than...	0.687287
48258	25.0	seeing someone	female	straight	average	strictly anything	often	never	graduated from college/university	white	...	science / tech / engineering	san francisco, california	no kids and neutral to kids	likes dogs and dislikes cats	agnosticism and very serious about it	gemini and it's fun to think about	when drinking	english (fluently), c++ (poorly), french (okay)	hello world! this feels very much like a colle...	0.668369
35354	29.0	single	female	straight	average	anything	socially	never	graduated from college/university	white	...	science / tech / engineering	san francisco, california	doesn't have kids, but wants them	likes dogs and dislikes cats	catholicism and somewhat serious about it	virgo	no	english	i'm a professional engineer by trade, but i sp...	0.656392
1085	26.0	single	female	straight	average	anything	socially	never	graduated from college/university	native american, hispanic / latin	...	science / tech / engineering	san francisco, california	doesn't want kids	likes dogs and dislikes cats	atheism and very serious about it	libra but it doesn't matter	no	english (fluently), spanish (fluently), italia...	i'm verrry bad at self-summaries. just, fair ...	0.632605
31041	26.0	single	female	straight	average	anything	socially	never	graduated from college/university	asian	...	science / tech / engineering	san francisco, california	doesn't have kids, but wants them	likes dogs and dislikes cats	catholicism but not too serious about it	aries but it doesn't matter	when drinking	english (fluently), tagalog (fluently)	i'm a nerd with glasses and the occasional zit...	0.604481
8965	30.0	single	female	straight	average	mostly anything	often	never	graduated from college/university	white	...	science / tech / engineering	san francisco, california	doesn't have kids, but wants them	dislikes dogs and dislikes cats	christianity but not too serious about it	gemini but it doesn't matter	no	english (fluently)	originally from ireland, i'm living in san fra...	0.553098
25519	25.0	single	female	straight	skinny	anything	socially	never	graduated from college/university	middle eastern, hispanic / latin, white, other	...	science / tech / engineering	san francisco, california	doesn't have kids	likes dogs and dislikes cats	irreligion	sagittarius and it's fun to think about	no	english, chinese, other	discovering a new city and learning how to cod...	0.518423
13047	25.0	single	female	straight	fit	anything	socially	never	graduated from college/university	middle eastern, white	...	science / tech / engineering	san francisco, california	no kids and neutral to kids	dislikes dogs and dislikes cats	judaism and laughing about it	taurus and it's fun to think about	no	english (fluently), hebrew (okay)	here's my best shot at the ambitious task of f...	0.484520
26625	27.0	single	female	straight	full figured	mostly anything	socially	never	graduated from college/university	asian	...	science / tech / engineering	south san francisco, california	no kids and neutral to kids	likes dogs and dislikes cats	irreligion	cancer and it's fun to think about	yes	english	call it what you will but i never liked descri...	0.445736
14433	33.0	single	female	straight	average	mostly anything	socially	never	graduated from college/university	white	...	science / tech / engineering	oakland, california	no kids and neutral to kids	has dogs and dislikes cats	irreligion	leo and it's fun to think about	no	english	i just moved home to the bay after living in a...	0.397437

```
Click to add a breakpoint
user_id = 12345
filtered_matches = get_top_matches(
    user_id, similarity_df, df, top_n=10,
    age_range=(25, 35),
    height_range=(160, 180),
    location="California",
    education="university",
    job="engineer",
    pets=["likes dogs", "dislikes cats"],
    # ethnicity="Asian",
    # diet="vegan",
    # offspring=["neutral to kids"],
    # keyword_filter="love hiking"
)

display(df.loc[user_id].to_frame().T)
display(filtered_matches)
```

	age	status	sex	orientation	body_type	diet	drinks	drugs	education	ethnicity	...	job	location	offspring	pets	religion	sign	smokes	speaks	essay_all	similarity_score
59824	26.0	single	female	straight	thin	mostly anything	socially	never	graduated from college/university	white	...	science / tech / engineering	san mateo, california	doesn't have kids, but might want them	likes dogs and dislikes cats	christianity and very serious about it	aquarius but it doesn't matter	no	english (fluently), french (poorly)	i moved to the bay area from indiana less than...	0.687287
48258	25.0	seeing someone	female	straight	average	strictly anything	often	never	graduated from college/university	white	...	science / tech / engineering	san francisco, california	no kids and neutral to kids	likes dogs and dislikes cats	agnosticism and very serious about it	gemini and it's fun to think about	when drinking	english (fluently), c++ (poorly), french (okay)	hello world! this feels very much like a colle...	0.668369
35354	29.0	single	female	straight	average	anything	socially	never	graduated from college/university	white	...	science / tech / engineering	san francisco, california	doesn't have kids, but wants them	likes dogs and dislikes cats	catholicism and somewhat serious about it	virgo	no	english	i'm a professional engineer by trade, but i sp...	0.656392
1085	26.0	single	female	straight	average	anything	socially	never	graduated from college/university	native american, hispanic / latin	...	science / tech / engineering	san francisco, california	doesn't want kids	likes dogs and dislikes cats	atheism and very serious about it	libra but it doesn't matter	no	english (fluently), spanish (fluently), italia...	i'm verrrry bad at self-summaries. just, fair ...	0.632605
31041	26.0	single	female	straight	average	anything	socially	never	graduated from college/university	asian	...	science / tech / engineering	san francisco, california	doesn't have kids, but wants them	likes dogs and dislikes cats	catholicism but not too serious about it	aries but it doesn't matter	when drinking	english (fluently), tagalog (fluently)	i'm a nerd with glasses and the occasional zit...	0.604481
8965	30.0	single	female	straight	average	mostly anything	often	never	graduated from college/university	white	...	science / tech / engineering	san francisco, california	doesn't have kids, but wants them	dislikes dogs and dislikes cats	christianity but not too serious about it	gemini but it doesn't matter	no	english (fluently)	originally from ireland, i'm living in san fra...	0.553098
25519	25.0	single	female	straight	skinny	anything	socially	never	graduated from college/university	middle eastern, hispanic / latin, white, other	...	science / tech / engineering	san francisco, california	doesn't have kids	likes dogs and dislikes cats	irreligion	sagittarius and it's fun to think about	no	english, chinese, other	discovering a new city and learning how to cod...	0.518423
13047	25.0	single	female	straight	fit	anything	socially	never	graduated from college/university	middle eastern, white	...	science / tech / engineering	san francisco, california	no kids and neutral to kids	dislikes dogs and dislikes cats	judaism and laughing about it	taurus and it's fun to think about	no	english (fluently), hebrew (okay)	here's my best shot at the ambitious task of f...	0.484520
26625	27.0	single	female	straight	full figured	mostly anything	socially	never	graduated from college/university	asian	...	science / tech / engineering	south san francisco, california	no kids and neutral to kids	likes dogs and dislikes cats	irreligion	cancer and it's fun to think about	yes	english	call it what you will but i never liked descri...	0.445736
14433	33.0	single	female	straight	average	mostly anything	socially	never	graduated from college/university	white	...	science / tech / engineering	oakland, california	no kids and neutral to kids	has dogs and dislikes cats	irreligion	leo and it's fun to think about	no	english	i just moved home to the bay after living in a	0.397437

Similarity - Example

- User 12345 With His Requirements:
  - Age: 25-35
  - Height Range: 160-180
  - Lived in California
  - Possess University Degree
  - Work as Engineer
  - Like Dogs, Dislike Cats

USER ID	Age	Status	Sex	Orientation	Body Type	Diet	Drinks	Drugs	Education	Ethnicity	Height	Job
59824	26	Single	Female	Straight	Thin	Mostly Anything	Socially	Never	Graduated from University	White	170	Science/Tech/Engineering




USER ID	Location	Offspring	Pets	Religion	Sign	Smokes	Speaks
59824	san mateo, california	doesn't have kids, but might want them	likes dogs and dislikes cats	christianity and very serious about it	aquarius but it doesn't matter	No	english (fluently), french (poorly)


Deployment

Streamlit

Q: What is deployment?

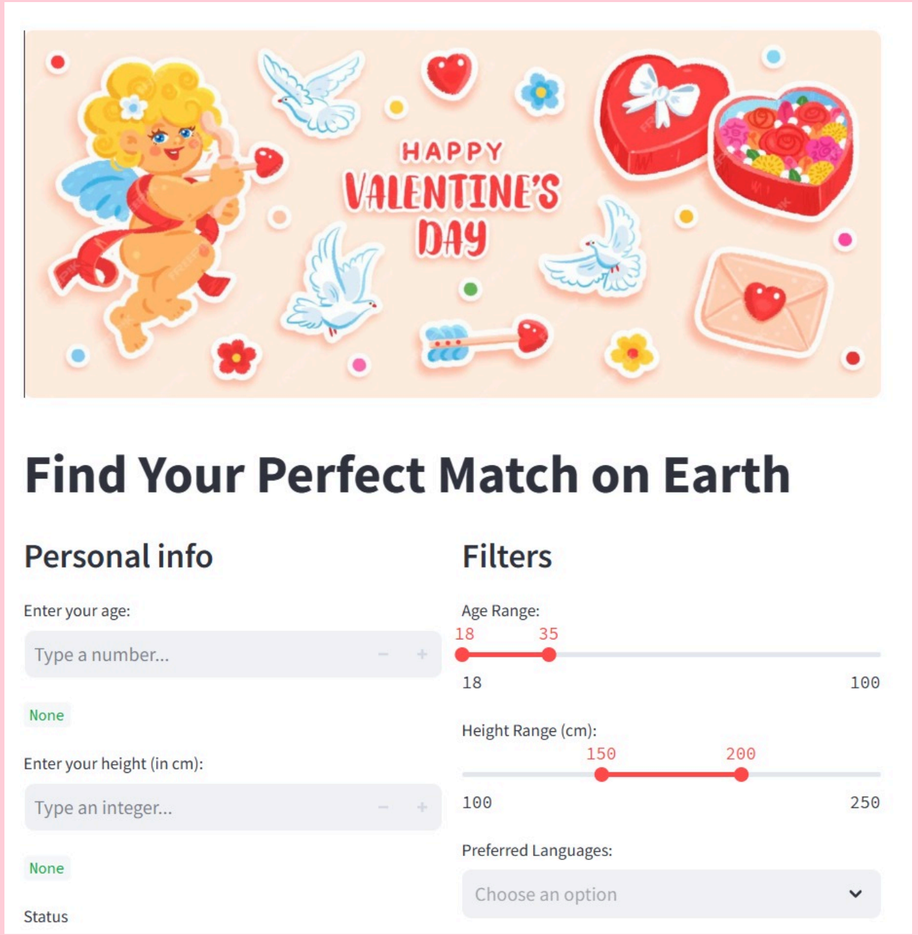
- How we integrate our machine learning model into a production environment

Q: What we use?

- Use **Streamlit** for deployment
- an open-source Python framework for delivering cool interactive web apps

Q: What we did for this part?

- write codes to set up a front-end web apps as a window for the users to input their personal info
- insert filters in the app



Limitation & Constraints

Dataset

- Geographic (US), Not Updated Dataset
- Tons of Cleansing Work
  - Null Value, Format, Categorizing User Responses
- Suspected Bias Sampling(e.g. Cat & Dog Lover)

Working

- Computation Power
- Alt: Google Cloud Platform
  - Lack of GPU Support

Time Constraint

- Online Available Dataset
- Model Selection
  - Compare Models
  - TF-IDF, Word2Vec, SBERT
- Waiting for Computing Result
- Compromise The Method to do the Project
  - Fill NA Method
  - ML Model
- Testing with Extra Examples

NOT ENOUGH SLEEP !!!

Future Work

- **Try to:**
  - Updated & Localized dataset
  - Web Scrapping from Dating Apps
  - Evaluate performance
  - User Profile With Personal Photo
  - If Deploy:
    - User Follow-up Decision to Enhance the Model
- **Different Fill NA Method**
  - SVD (Predict missing values)
- **Comparison of Different Machine Learning Model**
  - TF-IDF
  - Word2Vec
- **Fully Operate on Cloud Platform**
- **Storage in Database and Query by SQL**

Happy Valentine's Day