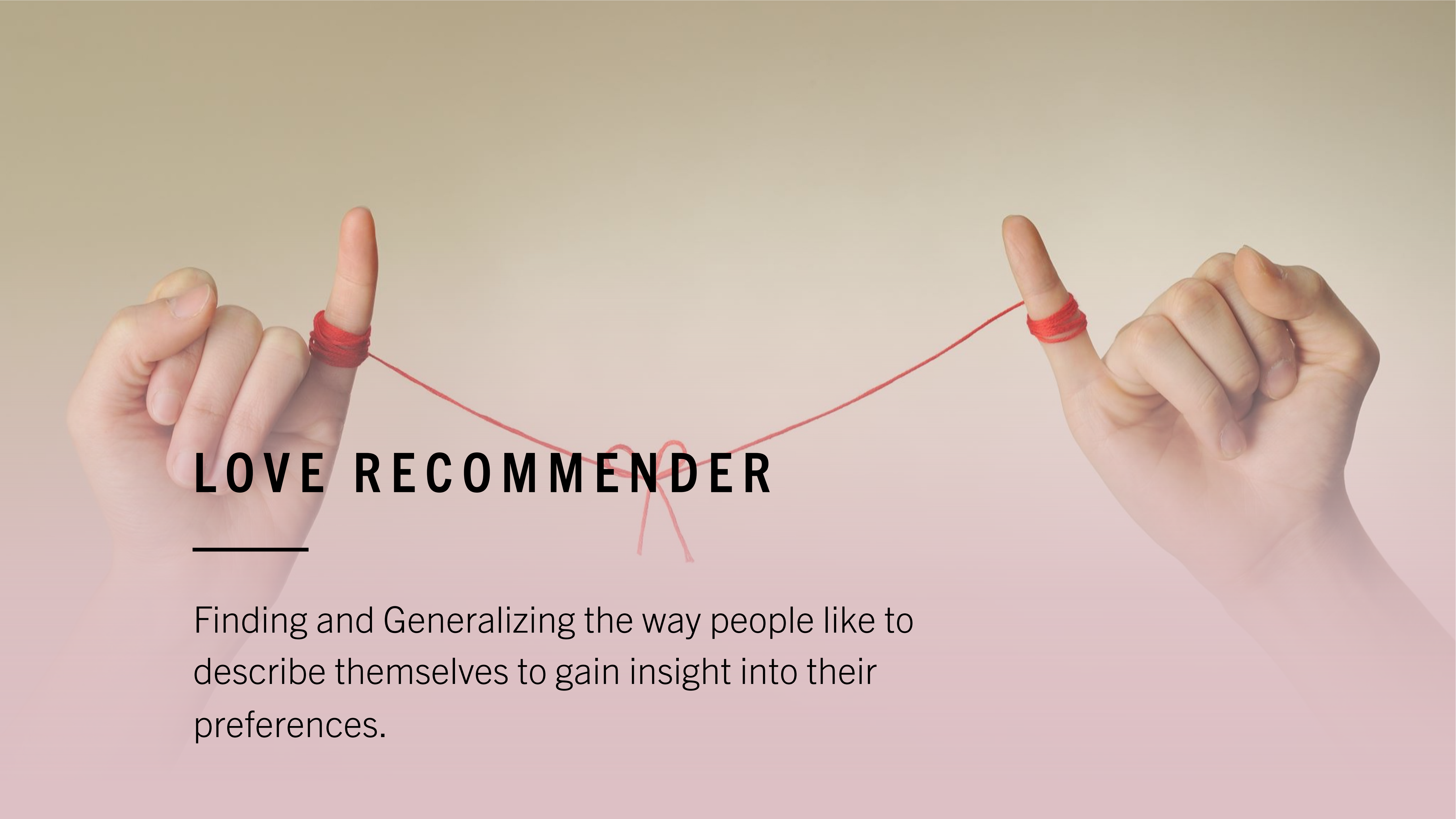


LOVE RECOMMENDER

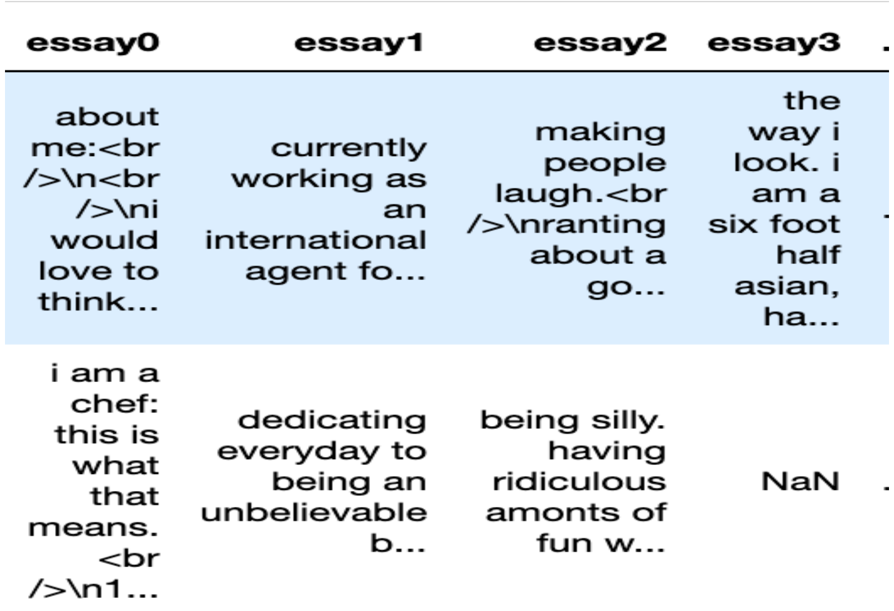
Finding and Generalizing the way people like to describe themselves to gain insight into their preferences.

PROJECT GOAL

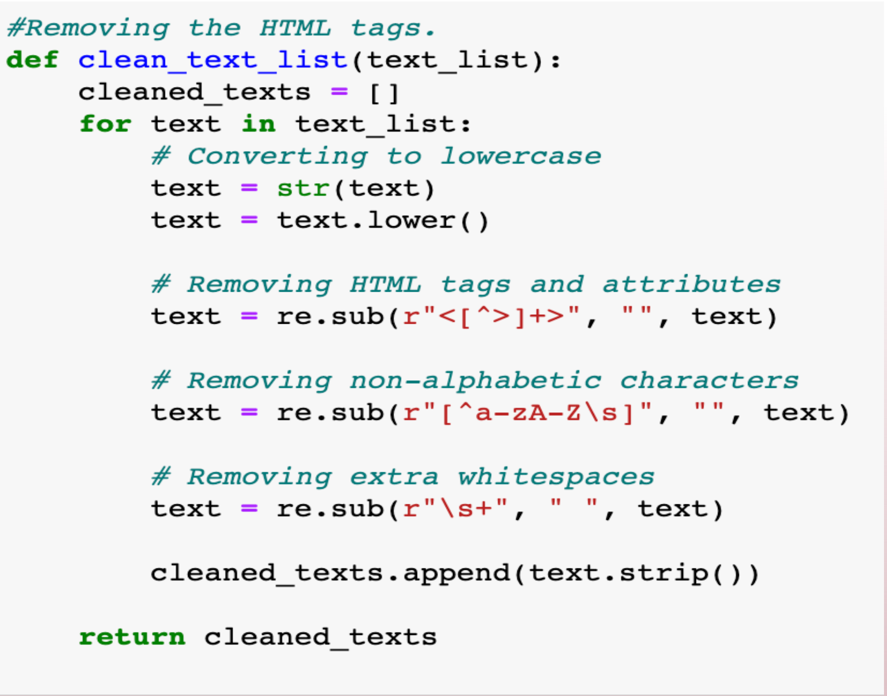
- Analyzing the descriptions of people available on okcupid website's dataset and finding out how they like to describe themselves most commonly.
- Creating a Recommendation system that asks a user to enter their description and find the most similar descriptions available on the dataset.

DATA CLEANING

- The main part of data cleaning was removing the html tags from the available descriptions in the dataset.
- After obtaining the cleaned descriptions, I saved just the descriptions in a .txt file since I was concerned only with text analysis.



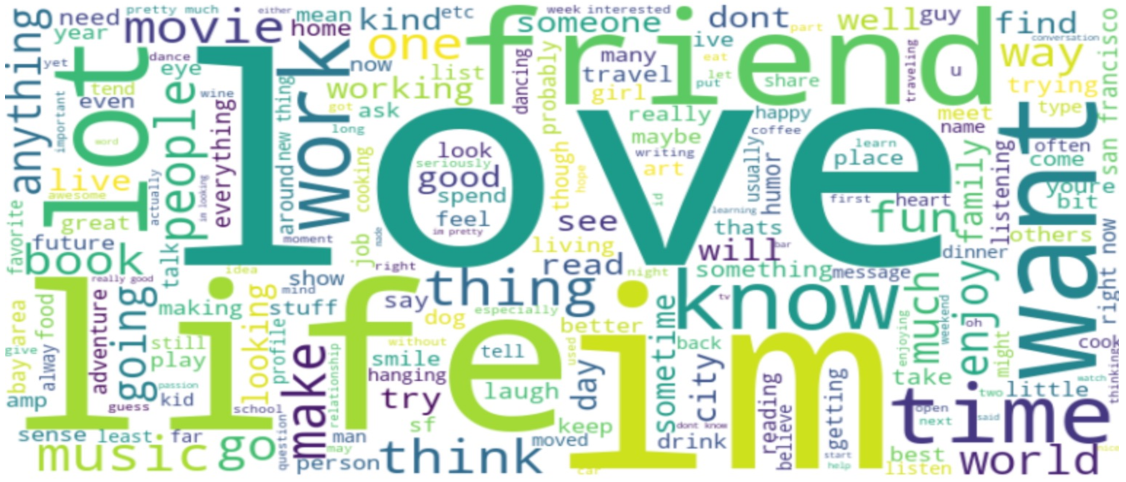
essay0	essay1	essay2	essay3	.
about me:<br />\n<br />\ni would love to think...	currently working as an international agent fo...	making people laugh.<br />\nranting about a go...	the way i look. i am a six foot half asian, ha...	-
i am a chef: this is what that means.<br />\n1...	dedicating everyday to being an unbelievable b...	being silly. having ridiculous amonts of fun w...	NaN	.

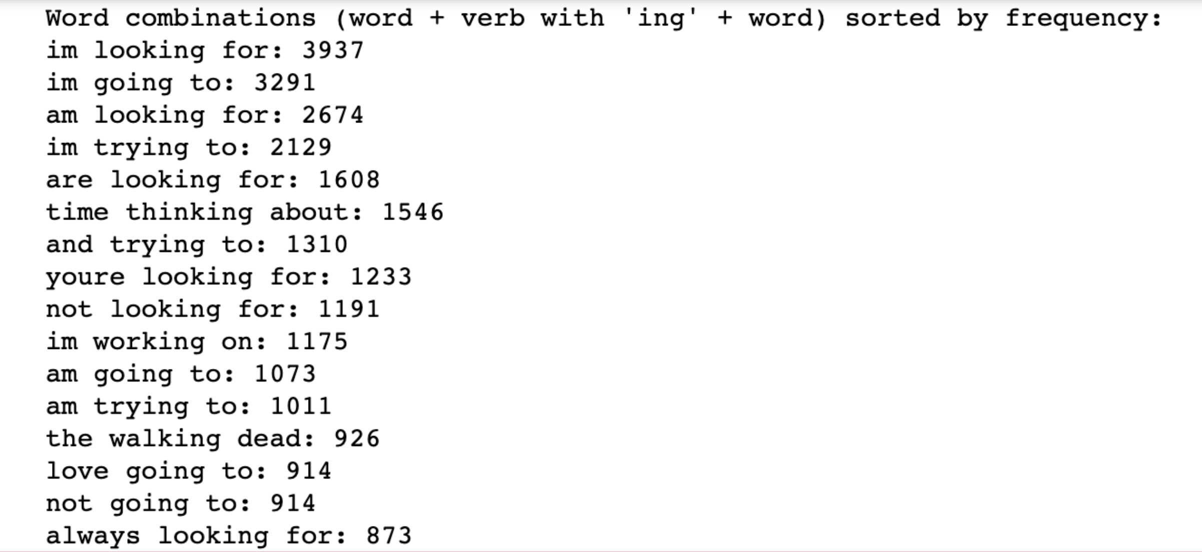


```
#Removing the HTML tags.
def clean_text_list(text_list):
    cleaned_texts = []
    for text in text_list:
        # Converting to lowercase
        text = str(text)
        text = text.lower()

        # Removing HTML tags and attributes
        text = re.sub(r"<[^>]+>", "", text)

        # Removing non-alphabetic characters
        text = re.sub(r"[^a-zA-Z\s]", "", text)

        # Removing extra whitespaces
        text = re.sub(r"\s+", " ", text)

        cleaned_texts.append(text.strip())

    return cleaned_texts
```

INITIAL ANALYSIS

- Firstly, I separated all the descriptions into single words and analyzed the most common words using wordcloud. The most commonly used words were not insightful as they contained lots of common articles.
- Secondly, I tried to filter the words by common present-tense action verbs, having a word+ing-verb+ word as a unit and examine the most common form of them. The result were still typical and uninsightful.
- Finally, I decided to import and use natural language processing library to gain more insightful data as the library would filter our common articles and extract common but useful data.

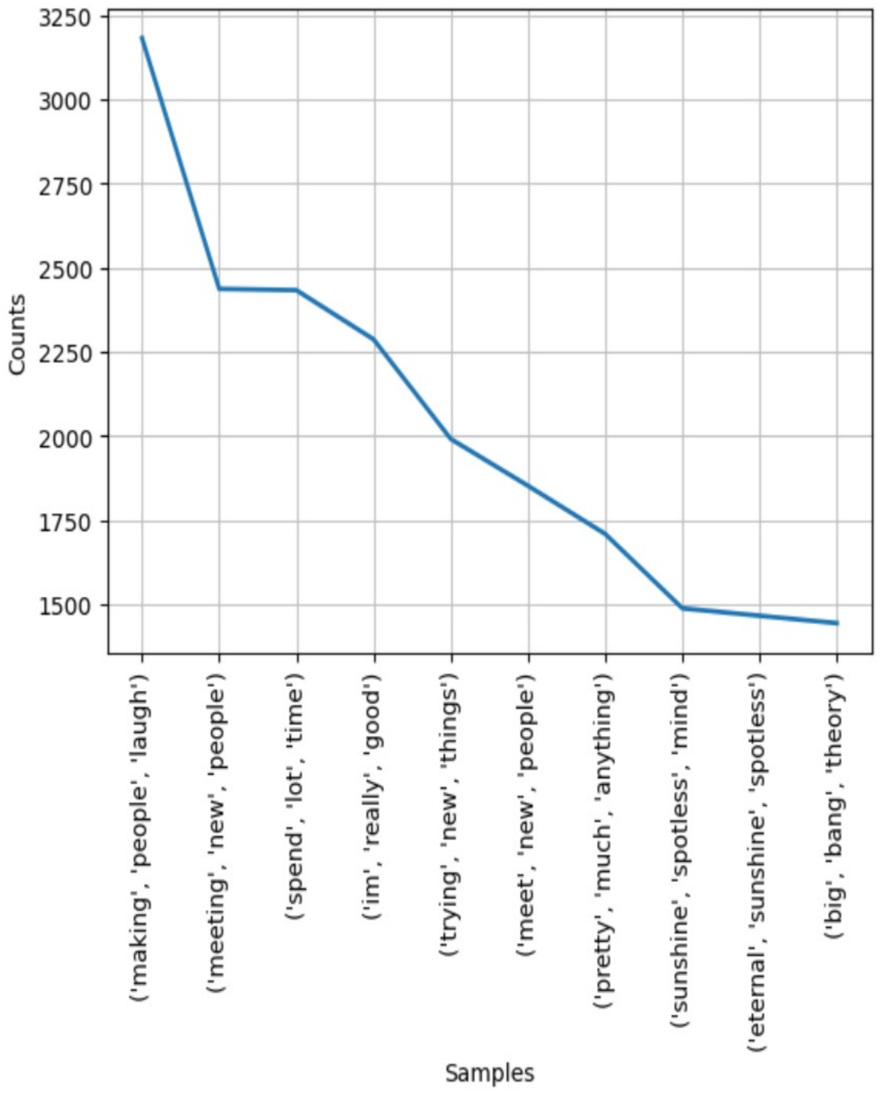




```
Word combinations (word + verb with 'ing' + word) sorted by frequency:
im looking for: 3937
im going to: 3291
am looking for: 2674
im trying to: 2129
are looking for: 1608
time thinking about: 1546
and trying to: 1310
youre looking for: 1233
not looking for: 1191
im working on: 1175
am going to: 1073
am trying to: 1011
the walking dead: 926
love going to: 914
not going to: 914
always looking for: 873
```

NLTK NGRAMS ANALYSIS

- After Importing the nltk library, I use its ngrams feature to properly analyze the descriptions. This method is superior to manual analysis as before since the library filters out commonly used articles to produce contiguous sequence of insightful words.
- After performing ngrams test with n=1,2,3,4&5, I found the words with n=3 to be most insightful.
- We find that people describe themselves as who like to meet new people, make people laugh, try new things whereas the most common movies and TV shows they like are eternal sunshine of the spotless mind and big bang theory.



RECOMMENDATION SYSTEM

- For the recommendation system I use the user input i.e. their own description and use TF-IDF vectorizing technique and cosine similarity to compare user's description to descriptions in the dataset and present the top 5 most common descriptions.
- The method works as long there are some common unique words in user's description and descriptions in the dataset, but further neural network and deep learning analysis would be necessary to refine the recommendations and provide more contextual recommendations.

CONCLUSION

- I have demonstrated proficient data cleaning and text analysis skills, effectively removing HTML tags and transforming descriptions for insightful analysis.
- Through the utilization of natural language processing (NLP) techniques, I uncovered meaningful phrases and patterns within the dataset, enhancing the quality of insights.
- I successfully developed a recommendation system using TF-IDF vectorization and cosine similarity, showcasing my ability to provide personalized content suggestions.
- While the current system performs well, I recognize the potential for further improvements through neural networks and deep learning to offer more contextually relevant recommendations.
- In conclusion, this project underlines my expertise in data analysis and NLP, setting the stage for the development of advanced recommendation systems in the future.