# Subgroup Discovery Project Summary

Created by: Jason Wang

**Executive Summary:**

The overall objective of this project is to discover the subgroups with interesting properties using the survey data. The data set used in this project can be separated into two parts: attribute and response. For this part of the research, the survey response data is mainly used for the purpose of LDA topic modeling. The goal of LDA topic modeling is to search for the optimal number of topics which results in high entropy and low cosine-similarity. Detailed steps will be discussed below.

**File Used:**

face-rating-data-2.json

Subgroup Discovery.py

**Procedure:**

1. **Pre-processing:**

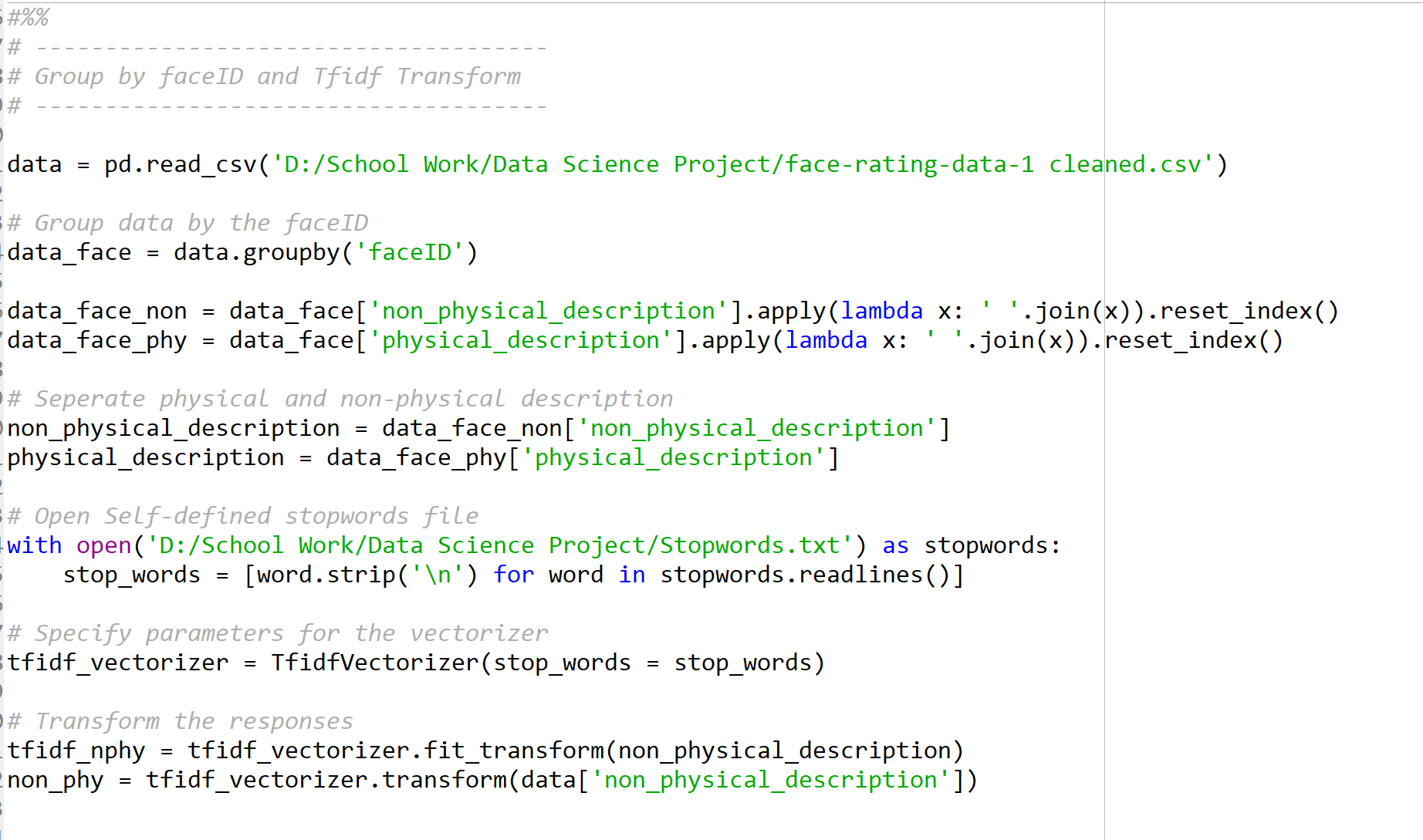
* Objective:
  + Parse out key values from the original .json file and convert it into a csv file. And then lower case all descriptions and remove all punctuations.
* Code:



* Integration in the future:
  + Use the code in the red box to parse out the desired new variable.

1. **Group Descriptions by faceID:**

* Group physical description and non-physical descriptions by each faceID.
* Combine all the responses with same faceID and treat them as one document.
* End up with 193 documents instead of 2491 documents.
* Convert them into a matrix of TF-IDF features with customized stop word list.
* Code:



1. **Grid Search for the Optimal Parameters:**

* Grid search LDA model to find the optimal hyperparameter that maximizes average entropy and average cosine similarity across all documents. Hyperparameter searched: # of topics, doc\_topic\_prior, topic\_word\_prior
* Append result in a text file.
* Visualize the result and choose an optimal hyperparameter range
* Grid search again around the optimal hyperparameter range
* Obtain the optimal result
* Functions Explanation:
  + Entropy: Used standard entropy formula and normalized by setting the base equal to the number of topics.
  + Cosine Similarity: Take one probability distribution vector at a time and compute cosine similarity against every other vectors in the matrix. Take the average of the results.

**Results:**

* Based on grid search results, the optimal hyperparameter occurred when number of topics is 9, document\_topic\_prior is 0.27 and topic\_word\_prior is 0.5