TaD_Project_How Gender and Education Relate to Writing Style on Dating Profiles

December 6, 2025

1 How Gender and Education Relate to Writing Style on Dating Profiles?

1.1 Introduction

1.1.1 Problem statement

- *What is your topic and why is it important and interesting to social scientists?*

Our topic examines how gender and education level shape the writing styles people use on dating profiles. Dating profiles are places where individuals intentionally construct how they want to be perceived by potential partners. Writing style becomes a subtle form of social signaling, allowing people to reveal values, expectations, and preferences without saying them directly. Since gender and education level already influence self-presentation in everyday life, studying how these factors appear in online dating helps us understand how people communicate identity, status, and compatibility in digital spaces. Analyzing these language patterns connects to broader social science themes such as social stratification, mate selection, assortative mating, and the ways people navigate social boundaries when forming relationships.

There has been a variety of research on assortative mating, which is the idea that people often find partners of similar backgrounds and/or partners who possess similar traits. This concept is applicable to many aspects of a person, but has been found to be particularly common in terms of educational background. Scholars Stephen Whyte and Benno Torgler's work titled "Things change with age: Educational assortment in online dating, Personality and Individual Differences", explores assortative mating based on education in the world of online dating. They concluded that "more educated online daters are consistently likely to assort positively (homogamy) meaning that they are more likely to contact potential mates with the same level of education." (Whyte and Torgler, 2017). The concept is interesting because it often happens subconsciously. Typically, people aren't exclusively searching for a partner of a similar background, but the ways in which their background has shaped them creates a foundation for forming relationships with people of a similar background. David M. Buss takes a more broad approach in his article "Human Mate Selection: Opposites Are Sometimes Said to Attract, but in Fact We Are Likely to Marry Someone Who Is Similar to Us in Almost Every Variable", where education is considered in relation to other variables of assortative mating such as socioeconomic background, personality traits, religion, and more. He points out that "education is also an interesting factor as it is commonly used in human mating behaviour as a proxy for resources and future provision helping to gain reproductive or (economic) advantages" (Buss, 1985). This article was published in American Scientist years before online dating sites had been created, so it is interesting to see how these ideas hold up in a vastly different setting. Overall,

there is a solid foundation of existing literature surrounding assortative mating, but writing styles are not examined in these works.

- *Review relevant literature.*

    1. Buss, David M. "Human Mate Selection: Opposites Are Sometimes Said to Attract, but in Fact We Are Likely to Marry Someone Who Is Similar to Us in Almost Every Variable." American Scientist, vol. 73, no. 1, 1985, pp. 47–51. JSTOR, http://www.jstor.org/stable/27853061.
    2. Stephen Whyte, Benno Torgler, Things change with age: Educational assortment in online dating, Personality and Individual Differences, Volume 109, 2017, Pages 5-11, ISSN 0191-8869, https://doi.org/10.1016/j.paid.2016.12.031.

Note: This and other sections should include full sentences and paragraphs with proper citations.

1.1.2 Research Question/s

How do gender and education level influence the writing styles people use when signaling preferences for potential partners on dating profiles?

1.1.3 Data overview and analysis plan

- *What data do you plan to use? This can be very general.*

We plan to use the OkCupid Dating Profiles dataset from Kaggle. It includes about 60,000 profiles with both demographic information (age, gender, orientation, and education level) and several open-ended text prompts. The text responses are what we are going to do the analysis on, since they give us enough writing samples to compare how men and women present themselves and examine whether their languages suggest a preference for partners with similar educational backgrounds.

- *What methods do you plan to us to answer your questions and address your hypotheses?*

We plan to analyze dating profile text using descriptive text analysis, including measures of text length and word usage (such as word frequency distributions and word clouds). We will also apply standard readability measures to capture differences in sentence structure and writing complexity. These methods allow us to compare writing styles across gender and education levels and assess whether individuals with similar educational backgrounds tend to present themselves in similar ways.

- *Discuss papers that use similar methods to the ones you plan on using.*

There have been many studies that use readability measures to compare how different groups write. In an article by Erin Hengel, "Publishing While Female", Hengel uses Flesch Reading Ease and Flesch–Kincaid Grade Level to study gender differences in academic writing and finds consistent differences in text complexity across male and female authors (Hengel, 2017). This research provides background for any differences in the writing of male and female users that may show up in the text analysis. While the research found that women's writing had better readability scores, this study was done using formal and academic writing, so it will be interesting to explore readability of informal and personal writing. In a different setting, Michael D. Aldridge's work "Writing and Designing Readable Patient Education Materials", applies similar readability formulas to patient education materials and shows that the reading level of many texts does not

match the education levels of their intended audiences, highlighting how education shapes written communication (Aldridge, 2004). These findings are particularly interesting regarding the chosen research question, given that an online dating site is a fairly informal setting. More recently, Sánchez-Hernández et al. (2025) examined the readability of sustainability reports and related linguistic complexity to gender and experience at the organizational level. Although these studies focus on very different types of writing, they all use readability metrics to compare writing styles across groups. Our project uses the same general approach but applies it to dating profiles, where differences in readability may reflect informal communication norms linked to gender and education rather than formal writing standards.

1.1.4 Expected findings/hypotheses

Drawing on the idea of assortative mating, we hypothesize that gender and education level are related to differences in how people write about themselves on dating profiles. In particular, people with higher levels of education may show more similar writing styles, such as comparable readability levels or sentence structures. These similarities may reflect shared communication habits or norms rather than intentional strategies. Overall, these writing patterns may act as subtle signals of background or compatibility to potential partners, even if they do not directly determine dating outcomes.

1.2 Data description/Overview

- *Detailed description of the dataset.*

The dataset comes from Kaggle and is called OKCupid Profiles. It contains 59,946 observations, and each row represents one real user profile from the OKCupid dating website. There are 30 columns in total. Some of these columns are regular demographic variables, such as age, status, sex, orientation, body_type, diet, drinks, drugs, education, ethnicity, height, income, job, last_online, location, offspring, pets, religion, sign, smokes, and speaks. The rest of the columns are essay0 to essay9, which are long, free-text answers that users wrote about themselves. These essays cover things like self-summary, hobbies, favorite books and movies, what they are doing with their life, and other personal details. This dataset was created from profiles in the San Francisco area during the 2010s, so it reflects online dating culture from that time period. The OKCupid site had several open-ended writing prompts, and users could write as much or as little as they wanted. Because of this, the text in the dataset is very natural, messy, and personal. Some people wrote long paragraphs, while others wrote only a few words or simple lists. The dataset is also interesting because it combines structured data (things like age and education) with unstructured text (the essays), which makes it useful for many types of analysis.

This dataset has several limitations. First, the writing quality is very inconsistent. Some essays are long and detailed, while others are only a few words, have no punctuation, or are written as lists instead of real sentences. This makes readability analysis harder. Second, the dataset is old, from a period in the 2010s, so it may not show how people write dating profiles today. Third, many users did not fill out all ten essays, so a lot of profiles have missing or incomplete text. Another limitation is that the dataset does not include the original essay prompts. The columns are named essay0 to essay9, but we do not know which section each one belongs to, so we had to combine them into one all_essays column for analysis. There are also limitations in the education variable. People on dating sites can lie or exaggerate, so someone might say they "graduated" even if they are still in school. The education field also contains some categories that are too vague or not

useful for our analysis, such as "college/university," "space camp," "graduated from space camp," or "working on space camp." We had to drop these categories because they are ambiguous and do not clearly indicate a real education level. Finally, OKCupid users are not a random sample of the population. They tend to be younger, more urban, and more active online, so the results cannot represent everyone. These limitations mean we need to be careful about how far we generalize our findings.

Ethical Considerations: This dataset comes from OKCupid and was made publicly available for research purposes. According to the dataset description, permission to use this data was explicitly granted by OKCupid. The dataset does not include usernames, profile pictures, or any direct identifiers, which helps protect user privacy. All of the analysis in this project is done at an aggregate level, and we do not attempt to identify or focus on individual users. However, while the platform granted permission for the dataset to be used, we do not have information about whether individual users explicitly consented to their profiles being included for research. Because of this, we are careful to treat the data as sensitive and to focus only on general writing patterns rather than personal details or individual behavior. We also avoid making judgments about specific people and instead examine broader trends related to gender and education.

- *Where did you get it?*

The dataset was sourced from Kaggle, where it is publicly available under the title "OKCupid Profiles." It consists of user profiles scraped from OKCupid and reflects how individuals presented themselves on the platform during that period.

- *How many observations does it contain?*

The dataset contains 59,946 user profiles and 30 total variables. Each row represents a single user profile, and each column captures a specific attribute or written response.

- *What other metadata is present?*

The dataset contains a range of structured demographic and lifestyle variables that serve as metadata. These include age, relationship status, sex, sexual orientation, body type, diet, drinking habits, drug use, smoking status, education level, ethnicity, height, income, job, and astrological sign. It also includes information about pets, whether the user has or wants children, their religion, and the languages they speak. Two additional fields, last_online and location, provide time-stamped activity data and geographic information, since users list cities across the San Francisco Bay Area with dates mostly from 2011 and 2012. Alongside these structured fields, the dataset contains ten unstructured text entries labeled essay0 through essay9, which include users' self-descriptions and personal statements.

- *What summary measures are helpful to introduce a reader to this data?*

Helpful summary measures for this dataset include basic information about the size and structure of the data, such as the number of rows, the number of columns, and the amount of missing information. It is also useful to look at how the main demographic variables are distributed, particularly education level and gender, since these define the groups used in the analysis. The bar chart showing the number of males and females in each education category helps illustrate how many users fall into each group.

For the writing portion of the dataset, summary measures like the average essay length within each gender and education group, as well as the variation around those averages, help introduce how

much users typically write. Readability scores and simple word-frequency patterns also provide an initial view of the general style and content of the essays.

It may be helpful to answer these questions in the context of running blocks of code.

1.2.1 Load Libraries & Data set

[1]:
```
import os
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import nltk
from nltk.tokenize import word_tokenize, sent_tokenize
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
from nltk import FreqDist
import re
import string
import matplotlib.pyplot as plt
import seaborn as sns
import textstat
from wordcloud import WordCloud
os.getcwd()
os.listdir()
nltk.download('punkt')
nltk.download('stopwords')
nltk.download('wordnet')
```

```
[nltk_data] Downloading package punkt to /home/jupyter-
[nltk_data]     huang68y/nltk_data…
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package stopwords to /home/jupyter-
[nltk_data]     huang68y/nltk_data…
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package wordnet to /home/jupyter-
[nltk_data]     huang68y/nltk_data…
[nltk_data]   Package wordnet is already up-to-date!
```

[1]:
```
True
```

[2]:
```
%%capture
#Load data set
df_cupid = pd.read_csv('okcupid_profiles.csv', encoding="latin1")
df_cupid.head(50)
```

2 Pre-processing steps

2.1 Part 1

In the first part of our preprocessing, we focused on preparing the dataset so it matched the needs of our research question. Because our project compares writing style across education levels, we first removed rows where education was missing or uninformative (for example, "space camp"), and we filtered out categories that could not be meaningfully grouped into "Advanced" or "Not Advanced." We also removed certain education labels that were too vague to classify, such as "college/university," because this could mean anything from someone who is still in college/university to someone who finished a full degree, making it unclear whether they should be placed in the Advanced or Not Advanced group. In contrast, we kept labels like "high school," even though they are broad, because they clearly belong in the Not Advanced group. This approach allowed us to remove only labels that were genuinely ambiguous while keeping those with consistent meaning. We also changed the gender labels from "f" and "m" to "female" and "male" for easier presentation and cleaner graphs. After this, we gathered all of the user essay fields into one combined text column. The dataset does not include information about which essay corresponds to which prompt, so even though the columns are labeled essay0 to essay9, we do not know which specific question each text belongs to. Because of this, we combined all available essays into a single all_essays column. We then skipped missing essays on a per-person basis instead of dropping entire rows, since many users only answered some of the prompts. Because we wanted our readability analysis to stay as close as possible to the original writing, we kept the preprocessing minimal at this stage and only normalized spacing while keeping punctuation, emojis, and the natural sentence structure. This gave us a stable text column that preserved the writer's authentic style. We then created two separate versions of the cleaned text: a readability-safe version with only light cleaning, and a second version for word-level analysis that required lowercasing, removing punctuation and stopwords, and lemmatizing each word.

[3]:
```
%%capture
# Drop rows where education is missing
# We do this first because the project focuses
# on comparing writing style across education levels.
df_cupid = df_cupid.dropna(subset=['education'])


# Rename sex categories for readability
df_cupid['sex'] = df_cupid['sex'].replace({'m': 'male', 'f': 'female'})


# Identify all essay columns
# Looks for any column name that starts with 'essay'
# (e.g., essay0, essay1, ..., essay9)
essay_cols = [col for col in df_cupid.columns if col.startswith('essay')]


# Combine all essay columns into a single text column
# Remove missing essays *per person*
# so that NaN essays are simply skipped,
```

```
# rather than being converted to "nan" or dropped globally.
df_cupid['all_essays'] = df_cupid[essay_cols].apply(
    lambda row: " ".join(
        [str(x) for x in row if pd.notna(x)]
    ),
    axis=1
)


# Drop the original individual essay columns
# Since all text is now in 'all_essays', the separate
# essay0-essay9 columns are no longer needed.
df_cupid.drop(columns=essay_cols, inplace=True)

# Preview the result to confirm the new column exists
# and that missing essays were removed per-person.
df_cupid.head(40)
```

Reasoning: The essay fields in the dataset are labeled generically from essay0 to essay9, and the dataset does not include information about which specific prompt each essay corresponds to. Because of this, we could not analyze prompt-specific writing styles. To address this issue, we combined all available essay responses into a single text column called all_essays for each user.

We confirmed that the new column that combines all essays exists, so we can now move on to the next step.

2.1.1 Education Data Cleaning and Creation of an Education × Gender Group

[4]:
```
# We first remove the education entries that are either ambiguous (e.g.,␣
 ↪"college/university")
#or clearly non-academic (e.g., "graduated from space camp").
# These categories can't be meaningfully grouped into Advanced vs. Not␣
 ↪Advanced, so we drop them before applying our classification.
drop_values = [
    "college/university",
    "graduated from space camp",
    "working on space camp",
    "dropped out of space camp",
    "space camp"
]

# Remove rows where the education label is in our drop list
df_cupid = df_cupid[~df_cupid['education'].isin(drop_values)].copy()


# Categorize the remaining education labels into Advanced vs. Not Advanced
# Cutoff: anything at or above "graduated from college/university" counts as␣
 ↪Advanced
```

```
def categorize_education(level):
    text = str(level).lower().strip()

    # Education labels that qualify as Advanced:
    # graduated from univeristy/college, masters, PhD,
    # law school, and med school (including dropped-out cases for masters, PhD,␣
 ↪law school, and med school)
    advanced_terms = [
        "graduated from college/university",
        "graduated from masters program",
        "working on masters program",
        "masters program",
        "dropped out of masters program",
        "graduated from ph.d program",
        "working on ph.d program",
        "ph.d program",
        "dropped out of ph.d program",
        "graduated from law school",
        "working on law school",
        "law school",
        "dropped out of law school",
        "graduated from med school",
        "working on med school",
        "dropped out of med school"
    ]
    # If the label is one of the Advanced terms, return Advanced
    if text in advanced_terms:
        return "Advanced"

    # Otherwise, classify as Not Advanced
    return "Not Advanced"


# Apply the education grouping
df_cupid['education_group'] = df_cupid['education'].apply(categorize_education)


# Create the education × gender group variable
# Combines education level and sex into one category, such as:
#   Advanced_male
#   Advanced_female
#   Not Advanced_male
#   Not Advanced_female

df_cupid['edu_gender_group'] = df_cupid['education_group'] + "_" +␣
 ↪df_cupid['sex']
```

```
# 5. Reset index for clean display
df_cupid = df_cupid.reset_index(drop=True)


# 6. View a random sample of 20 rows
df_cupid[['sex', 'education', 'education_group', 'edu_gender_group']].sample(20)
```

[4]:

	sex	education	education_group	edu_gender_group
16854	male	working on college/university	Not Advanced	Not Advanced_male
27602	male	graduated from masters program	Advanced	Advanced_male
18247	female	graduated from college/university	Advanced	Advanced_female
1535	female	graduated from masters program	Advanced	Advanced_female
41203	male	graduated from masters program	Advanced	Advanced_male
10781	male	graduated from masters program	Advanced	Advanced_male
28130	male	graduated from two-year college	Not Advanced	Not Advanced_male
16278	male	graduated from masters program	Advanced	Advanced_male
44081	female	graduated from law school	Advanced	Advanced_female
46687	male	graduated from masters program	Advanced	Advanced_male
28299	female	graduated from masters program	Advanced	Advanced_female
20736	male	graduated from ph.d program	Advanced	Advanced_male
26855	female	working on college/university	Not Advanced	Not Advanced_female
5817	male	graduated from college/university	Advanced	Advanced_male
9068	female	graduated from college/university	Advanced	Advanced_female
50450	female	working on two-year college	Not Advanced	Not Advanced_female
1292	male	graduated from college/university	Advanced	Advanced_male
31822	female	graduated from college/university	Advanced	Advanced_female
29451	female	working on college/university	Not Advanced	Not Advanced_female
47295	male	graduated from college/university	Advanced	Advanced_male

```
1292          Advanced_male
31822       Advanced_female
29451   Not Advanced_female
47295         Advanced_male
```

2.1.2 Pre-processing redability:

[5]:
```python
def prepare_for_readability(text):
    # Keep punctuation and emojis
    # Only normalize whitespace
    return " ".join(text.split())

df_cupid['clean_readability_text'] = (
    df_cupid['all_essays']
    .astype(str)
    .apply(prepare_for_readability)
)
```

Interpretation: For readability analysis, we minimally cleaned the essays by normalizing whitespace while preserving punctuation, emojis, and sentence structure, since readability metrics depend on intact grammatical features.

2.1.3 Pre-processing Word-Level Analysis:

[6]:
```python
%%capture
# Prepare stopwords and lemmatizer
stop_words = set(stopwords.words('english'))
lemm = WordNetLemmatizer()

# Function for NLP word-level cleaning
def prepare_for_word_analysis(text):
    # Lowercase
    text = text.lower()

    # Remove apostrophes inside contractions BEFORE punctuation removal
    text = re.sub(r"'", "", text)

    # Remove all punctuation
    text = text.translate(str.maketrans('', '', string.punctuation))

    # Tokenize
    tokens = word_tokenize(text)

    # Keep alphabetic tokens and remove stopwords
    tokens = [t for t in tokens if t.isalpha() and t not in stop_words]

    # Lemmatize tokens
```

```
    tokens = [lemm.lemmatize(t) for t in tokens]

    return tokens

# Create NEW column with cleaned tokens for word-level analysis
df_cupid['clean_word_tokens'] = (
    df_cupid['all_essays']
    .astype(str)
    .apply(prepare_for_word_analysis)
)

# Create a string version of the tokens for readability/display
df_cupid['clean_word_string'] = df_cupid['clean_word_tokens'].apply(lambda x: "␣
 ↪".join(x))

# Preview the first rows
pd.set_option("display.max_colwidth", None)
df_cupid[['all_essays', 'clean_word_tokens', 'clean_word_string']].head()
```

Interpretation: For word-level NLP analysis, we removed punctuation, normalized casing, filtered out stopwords, kept only alphabetic tokens, and lemmatized the remaining words to obtain a consistent set of meaningful tokens for frequency and pattern analysis.

3 EDA

[7]:
```
# Before analyzing writing patterns, we check the structure
# of the dataset: column types, missing values, and basic layout.
# This helps confirm the data is complete and formatted properly.

# Look at the number of rows, columns, and overall structure
df_cupid.info()

# Check the percentage of missing values by column
df_cupid.isna().mean().sort_values(ascending=False)

# Display the first few rows to visually inspect the data
df_cupid.head(5)
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 50834 entries, 0 to 50833
Data columns (total 27 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   age                     50834 non-null  int64
 1   status                  50834 non-null  object
 2   sex                     50834 non-null  object
 3   orientation             50834 non-null  object
```

```
 4   body_type                50605 non-null  object
 5   diet                     31203 non-null  object
 6   drinks                   49168 non-null  object
 7   drugs                    39046 non-null  object
 8   education                50834 non-null  object
 9   ethnicity                46614 non-null  object
 10  height                   50834 non-null  float64
 11  income                   50834 non-null  int64
 12  job                      45979 non-null  object
 13  last_online              50834 non-null  object
 14  location                 50834 non-null  object
 15  offspring                21305 non-null  object
 16  pets                     35262 non-null  object
 17  religion                 35175 non-null  object
 18  sign                     42538 non-null  object
 19  smokes                   47108 non-null  object
 20  speaks                   50797 non-null  object
 21  all_essays               50834 non-null  object
 22  education_group          50834 non-null  object
 23  edu_gender_group         50834 non-null  object
 24  clean_readability_text   50834 non-null  object
 25  clean_word_tokens        50834 non-null  object
 26  clean_word_string        50834 non-null  object
dtypes: float64(1), int64(2), object(24)
memory usage: 10.5+ MB
```

[7]:

	age	status	sex	orientation	body_type	diet	drinks	drugs	education	ethnicity	…	religion
0	22	single	male	straight	a little extra	strictly anything	socially	never	working on college/university	asian, white	…	agnosticism and very serious about it
1	38	available	male	straight	thin	anything	socially	NaN	graduated from masters program	NaN	…	NaN
2	23	single	male	straight	thin	vegetarian	socially	NaN	working on college/university	white	…	NaN
3	29	single	male	straight	athletic	NaN	socially	never	graduated from college/university	asian, black, other	…	NaN
4	29	single	male	straight	average	mostly anything	socially	NaN	graduated from college/university	white	…	atheism

```
                                sign     smokes  \
0                              gemini  sometimes
1  pisces but it doesn&rsquo;t matter         no
2                              pisces         no
3                            aquarius         no
4                              taurus         no

                                speaks  \
0                              english
1                 english, french, c++
2             english, german (poorly)
3                              english
4  english (fluently), chinese (okay)

           all_essays  \
0
about me:  i would love to think that i was some some kind of intellectual:
either the dumbest smart guy, or the smartest dumb guy. can't say i can tell the
difference. i love to talk about ideas and concepts. i forge odd metaphors
instead of reciting cliches. like the simularities between a friend of mine's
house and an underwater salt mine. my favorite word is salt by the way (weird
choice i know). to me most things in life are better as metaphors. i seek to
make myself a little better everyday, in some productively lazy way. got tired
of tying my shoes. considered hiring a five year old, but would probably have to
tie both of our shoes… decided to only wear leather shoes dress shoes.  about
you:  you love to have really serious, really deep conversations about really
silly stuff. you have to be willing to snap me out of a light hearted rant with
a kiss. you don't have to be funny, but you have to be able to make me laugh.
you should be able to bend spoons with your mind, and telepathically make me
smile while i am still at work. you should love life, and be cool with just
letting the wind blow. extra points for reading all this and guessing my
favorite video game (no hints given yet). and lastly you have a good attention
span. currently working as an international agent for a freight forwarding
company. import, export, domestic you know the works. online classes and trying
to better myself in my free time. perhaps a hours worth of a good book or a
video game on a lazy sunday. making people laugh. ranting about a good salting.
finding simplicity in complexity, and complexity in simplicity. the way i look.
i am a six foot half asian, half caucasian mutt. it makes it tough not to notice
me, and for me to blend in. books: absurdistan, the republic, of mice and men
(only book that made me want to cry), catcher in the rye, the prince.  movies:
gladiator, operation valkyrie, the producers, down periscope.  shows: the
borgia, arrested development, game of thrones, monty python  music: aesop rock,
hail mary mallon, george thorogood and the delaware destroyers, felt  food: i'm
down for anything. food. water. cell phone. shelter. duality and humorous things
trying to find someone to hang out with. i am down for anything except a club. i
am new to california and looking for someone to wisper my secrets to. you want
to be swept off your feet! you are tired of the norm. you want to catch a coffee
```

or a bite. or if you want to talk philosophy.
1 i'm not ashamed of much, but writing public text on an online dating site makes me pleasantly uncomfortable. i'll try to be as earnest as possible in the noble endeavor of standing naked before the world. i've lived in san francisco for 15 years, and both love it and find myself frustrated with its deficits. lots of great friends and acquaintances (which increases my apprehension to put anything on this site), but i'm feeling like meeting some new people that aren't just friends of friends. it's okay if you are a friend of a friend too. chances are, if you make it through the complex filtering process of multiple choice questions, lifestyle statistics, photo scanning, and these indulgent blurbs of text without moving quickly on to another search result, you are probably already a cultural peer and at most 2 people removed. at first, i thought i should say as little as possible here to avoid you, but that seems silly. as far as culture goes, i'm definitely more on the weird side of the spectrum, but i don't exactly wear it on my sleeve. once you get me talking, it will probably become increasingly apparent that while i'd like to think of myself as just like everybody else (and by some definition i certainly am), most people don't see me that way. that's fine with me. most of the people i find myself gravitating towards are pretty weird themselves. you probably are too. i make nerdy software for musicians, artists, and experimenters to indulge in their own weirdness, but i like to spend time away from the computer when working on my artwork (which is typically more concerned with group dynamics and communication, than with visual form, objects, or technology). i also record and deejay dance, noise, pop, and experimental music (most of which electronic or at least studio based). besides these relatively ego driven activities, i've been enjoying things like meditation and tai chi to try and gently flirt with ego death. improvising in different contexts. alternating between being present and decidedly outside of a moment, or trying to hold both at once. rambling intellectual conversations that hold said conversations in contempt while seeking to find something that transcends them. being critical while remaining generous. listening to and using body language--often performed in caricature or large gestures, if not outright interpretive dance. dry, dark, and raunchy humor. my large jaw and large glasses are the physical things people comment on the most. when sufficiently stimulated, i have an unmistakable cackle of a laugh. after that, it goes in more directions than i care to describe right now. maybe i'll come back to this. okay this is where the cultural matrix gets so specific, it's like being in the crosshairs. for what it's worth, i find myself reading more non-fiction than fiction. it's usually some kind of philosophy, art, or science text by silly authors such as ranciere, de certeau, bataille, baudrillard, butler, stein, arendt, nietzche, zizek, etc. i'll often throw in some weird new age or pop-psychology book in the mix as well. as for fiction, i enjoy what little i've read of eco, perec, wallace, bolao, dick, vonnegut, atwood, delilo, etc. when i was young, i was a rabid asimov reader. directors i find myself drawn to are makavejev, kuchar, jodorowsky, herzog, hara, klein, waters, verhoeven, ackerman, hitchcock, lang, gorin, goddard, miike, ohbayashi, tarkovsky, sokurov, warhol, etc. but i also like a good amount of "trashy" stuff. too much to name. i definitely enjoy the character development that happens in long form episodic

television over the course of 10-100 episodes, which a 1-2hr movie usually can't compete with. some of my recent tv favorites are: breaking bad, the wire, dexter, true blood, the prisoner, lost, fringe. a smattered sampling of the vast field of music i like and deejay: art ensemble, sun ra, evan parker, lil wayne, dj funk, mr. fingers, maurizio, rob hood, dan bell, james blake, nonesuch recordings, omar souleyman, ethiopiques, fela kuti, john cage, meredith monk, robert ashley, terry riley, yoko ono, merzbow, tom tom club, jit, juke, bounce, hyphy, snap, crunk, b'more, kuduro, pop, noise, jazz, techno, house, acid, new/no wave, (post)punk, etc. a few of the famous art/dance/theater folk that might locate my sensibility: andy warhol, bruce nauman, yayoi kusama, louise bourgeois, tino sehgal, george kuchar, michel duchamp, marina abramovic, gelatin, carolee schneeman, gustav metzger, mike kelly, mike smith, andrea fraser, gordon matta-clark, jerzy grotowski, samuel beckett, antonin artaud, tadeusz kantor, anna halperin, merce cunningham, etc. i'm clearly leaving out a younger generation of contemporary artists, many of whom are friends. local food regulars: sushi zone, chow, ppq, pagolac, lers ros, burma superstar, minako, shalimar, delfina pizza, rosamunde, arinells, suppenkuche, cha-ya, blue plate, golden era, etc. movement conversation creation contemplation touch humor viewing. listening. dancing. talking. drinking. performing. when i was five years old, i was known as "the boogerman". you are bright, open, intense, silly, ironic, critical, caring, generous, looking for an exploration, rather than finding "a match" of some predetermined qualities. i'm currently in a fabulous and open relationship, so you should be comfortable with that.
2
i work in a library and go to school. . . reading things written by old dead people playing synthesizers and organizing books according to the library of congress classification system socially awkward but i do my best bataille, celine, beckett. . . lynch, jarmusch, r.w. fassbender. . . twin peaks & fishing w/ john joy division, throbbing gristle, cabaret voltaire. . . vegetarian pho and coffee cats and german philosophy you feel so inclined.
3
hey how's it going? currently vague on the profile i know, more to come soon. looking to meet new folks outside of my circle of friends. i'm pretty responsive on the reply tip, feel free to drop a line. cheers. work work work work + play creating imagery to look at: http://bagsbrown.blogspot.com/ http://stayruly.blogspot.com/ i smile a lot and my inquisitive nature music: bands, rappers, musicians at the moment: thee oh sees. forever: wu-tang books: artbooks for days audiobooks: my collection, thick (thanks audible) shows: live ones food: with stellar friends whenever movies > tv podcast: radiolab, this american life, the moth, joe rogan, the champs
4
i'm an australian living in san francisco, but don't hold that against me. i spend most of my days trying to build cool stuff for my company. i speak mandarin and have been known to bust out chinese songs at karaoke. i'm pretty cheeky. someone asked me if that meant something about my arse, which i find really funny. i'm a little oddball. i have a wild imagination; i like to think of the most improbable reasons people are doing things just for fun. i love to

laugh and look for reasons to do so. occasionally this gets me in trouble because people think i'm laughing at them. sometimes i am, but more often i'm only laughing at myself. i'm an entrepreneur (like everyone else in sf, it seems) and i love what i do. i enjoy parties and downtime in equal measure. intelligence really turns me on and i love people who can teach me new things. building awesome stuff. figuring out what's important. having adventures. looking for treasure. imagining random shit. laughing at aforementioned random shit. being goofy. articulating what i think and feel. convincing people i'm right. admitting when i'm wrong. i'm also pretty good at helping people think through problems; my friends say i give good advice. and when i don't have a clue how to help, i will say: i give pretty good hug. i have a big smile. i also get asked if i'm wearing blue-coloured contacts (no). books: to kill a mockingbird, lord of the rings, 1984, the farseer trilogy. music: the beatles, frank sinatra, john mayer, jason mraz, deadmau5, andrew bayer, everything on anjunadeep records, bach, satie. tv shows: how i met your mother, scrubs, the west wing, breaking bad. movies: star wars, the godfather pt ii, 500 days of summer, napoleon dynamite, american beauty, lotr food: thai, vietnamese, shanghai dumplings, pizza! like everyone else, i love my friends and family, and need hugs, human contact, water and sunshine. let's take that as given. 1. something to build 2. something to sing 3. something to play on (my guitar would be first choice) 4. something to write/draw on 5. a big goal worth dreaming about 6. something to laugh at what my contribution to the world is going to be and/or should be. and what's for breakfast. i love breakfast. out with my friends! i cried on my first day at school because a bird shat on my head. true story. you're awesome.

	education_group	edu_gender_group	\
0	Not Advanced	Not Advanced_male	
1	Advanced	Advanced_male	
2	Not Advanced	Not Advanced_male	
3	Advanced	Advanced_male	
4	Advanced	Advanced_male	

clean_readability_text \
0
about me: i would love to think that i was some some kind of intellectual: either the dumbest smart guy, or the smartest dumb guy. can't say i can tell the difference. i love to talk about ideas and concepts. i forge odd metaphors instead of reciting cliches. like the simularities between a friend of mine's house and an underwater salt mine. my favorite word is salt by the way (weird choice i know). to me most things in life are better as metaphors. i seek to make myself a little better everyday, in some productively lazy way. got tired of tying my shoes. considered hiring a five year old, but would probably have to tie both of our shoes… decided to only wear leather shoes dress shoes. about you: you love to have really serious, really deep conversations about really silly stuff. you have to be willing to snap me out of a light hearted rant with a kiss. you don't have to be funny, but you have to be able to make me laugh.

you should be able to bend spoons with your mind, and telepathically make me smile while i am still at work. you should love life, and be cool with just letting the wind blow. extra points for reading all this and guessing my favorite video game (no hints given yet). and lastly you have a good attention span. currently working as an international agent for a freight forwarding company. import, export, domestic you know the works. online classes and trying to better myself in my free time. perhaps a hours worth of a good book or a video game on a lazy sunday. making people laugh. ranting about a good salting. finding simplicity in complexity, and complexity in simplicity. the way i look. i am a six foot half asian, half caucasian mutt. it makes it tough not to notice me, and for me to blend in. books: absurdistan, the republic, of mice and men (only book that made me want to cry), catcher in the rye, the prince. movies: gladiator, operation valkyrie, the producers, down periscope. shows: the borgia, arrested development, game of thrones, monty python music: aesop rock, hail mary mallon, george thorogood and the delaware destroyers, felt food: i'm down for anything. food. water. cell phone. shelter. duality and humorous things trying to find someone to hang out with. i am down for anything except a club. i am new to california and looking for someone to wisper my secrets to. you want to be swept off your feet! you are tired of the norm. you want to catch a coffee or a bite. or if you want to talk philosophy.

1 i'm not ashamed of much, but writing public text on an online dating site makes me pleasantly uncomfortable. i'll try to be as earnest as possible in the noble endeavor of standing naked before the world. i've lived in san francisco for 15 years, and both love it and find myself frustrated with its deficits. lots of great friends and acquaintances (which increases my apprehension to put anything on this site), but i'm feeling like meeting some new people that aren't just friends of friends. it's okay if you are a friend of a friend too. chances are, if you make it through the complex filtering process of multiple choice questions, lifestyle statistics, photo scanning, and these indulgent blurbs of text without moving quickly on to another search result, you are probably already a cultural peer and at most 2 people removed. at first, i thought i should say as little as possible here to avoid you, but that seems silly. as far as culture goes, i'm definitely more on the weird side of the spectrum, but i don't exactly wear it on my sleeve. once you get me talking, it will probably become increasingly apparent that while i'd like to think of myself as just like everybody else (and by some definition i certainly am), most people don't see me that way. that's fine with me. most of the people i find myself gravitating towards are pretty weird themselves. you probably are too. i make nerdy software for musicians, artists, and experimenters to indulge in their own weirdness, but i like to spend time away from the computer when working on my artwork (which is typically more concerned with group dynamics and communication, than with visual form, objects, or technology). i also record and deejay dance, noise, pop, and experimental music (most of which electronic or at least studio based). besides these relatively ego driven activities, i've been enjoying things like meditation and tai chi to try and gently flirt with ego death. improvising in different contexts. alternating between being present and decidedly outside of a moment, or trying to hold both at once. rambling intellectual conversations that

hold said conversations in contempt while seeking to find something that transcends them. being critical while remaining generous. listening to and using body language--often performed in caricature or large gestures, if not outright interpretive dance. dry, dark, and raunchy humor. my large jaw and large glasses are the physical things people comment on the most. when sufficiently stimulated, i have an unmistakable cackle of a laugh. after that, it goes in more directions than i care to describe right now. maybe i'll come back to this. okay this is where the cultural matrix gets so specific, it's like being in the crosshairs. for what it's worth, i find myself reading more non-fiction than fiction. it's usually some kind of philosophy, art, or science text by silly authors such as ranciere, de certeau, bataille, baudrillard, butler, stein, arendt, nietzche, zizek, etc. i'll often throw in some weird new age or pop-psychology book in the mix as well. as for fiction, i enjoy what little i've read of eco, perec, wallace, bolao, dick, vonnegut, atwood, delilo, etc. when i was young, i was a rabid asimov reader. directors i find myself drawn to are makavejev, kuchar, jodorowsky, herzog, hara, klein, waters, verhoeven, ackerman, hitchcock, lang, gorin, goddard, miike, ohbayashi, tarkovsky, sokurov, warhol, etc. but i also like a good amount of "trashy" stuff. too much to name. i definitely enjoy the character development that happens in long form episodic television over the course of 10-100 episodes, which a 1-2hr movie usually can't compete with. some of my recent tv favorites are: breaking bad, the wire, dexter, true blood, the prisoner, lost, fringe. a smattered sampling of the vast field of music i like and deejay: art ensemble, sun ra, evan parker, lil wayne, dj funk, mr. fingers, maurizio, rob hood, dan bell, james blake, nonesuch recordings, omar souleyman, ethiopiques, fela kuti, john cage, meredith monk, robert ashley, terry riley, yoko ono, merzbow, tom tom club, jit, juke, bounce, hyphy, snap, crunk, b'more, kuduro, pop, noise, jazz, techno, house, acid, new/no wave, (post)punk, etc. a few of the famous art/dance/theater folk that might locate my sensibility: andy warhol, bruce nauman, yayoi kusama, louise bourgeois, tino sehgal, george kuchar, michel duchamp, marina abramovic, gelatin, carolee schneeman, gustav metzger, mike kelly, mike smith, andrea fraser, gordon matta-clark, jerzy grotowski, samuel beckett, antonin artaud, tadeusz kantor, anna halperin, merce cunningham, etc. i'm clearly leaving out a younger generation of contemporary artists, many of whom are friends. local food regulars: sushi zone, chow, ppq, pagolac, lers ros, burma superstar, minako, shalimar, delfina pizza, rosamunde, arinells, suppenkuche, cha-ya, blue plate, golden era, etc. movement conversation creation contemplation touch humor viewing. listening. dancing. talking. drinking. performing. when i was five years old, i was known as "the boogerman". you are bright, open, intense, silly, ironic, critical, caring, generous, looking for an exploration, rather than finding "a match" of some predetermined qualities. i'm currently in a fabulous and open relationship, so you should be comfortable with that.

2

i work in a library and go to school. . . reading things written by old dead people playing synthesizers and organizing books according to the library of congress classification system socially awkward but i do my best bataille, celine, beckett. . . lynch, jarmusch, r.w. fassbender. . . twin peaks & fishing

w/ john joy division, throbbing gristle, cabaret voltaire. . . vegetarian pho and coffee cats and german philosophy you feel so inclined.
3
hey how's it going? currently vague on the profile i know, more to come soon. looking to meet new folks outside of my circle of friends. i'm pretty responsive on the reply tip, feel free to drop a line. cheers. work work work work + play creating imagery to look at: http://bagsbrown.blogspot.com/ http://stayruly.blogspot.com/ i smile a lot and my inquisitive nature music: bands, rappers, musicians at the moment: thee oh sees. forever: wu-tang books: artbooks for days audiobooks: my collection, thick (thanks audible) shows: live ones food: with stellar friends whenever movies > tv podcast: radiolab, this american life, the moth, joe rogan, the champs
4
i'm an australian living in san francisco, but don't hold that against me. i spend most of my days trying to build cool stuff for my company. i speak mandarin and have been known to bust out chinese songs at karaoke. i'm pretty cheeky. someone asked me if that meant something about my arse, which i find really funny. i'm a little oddball. i have a wild imagination; i like to think of the most improbable reasons people are doing things just for fun. i love to laugh and look for reasons to do so. occasionally this gets me in trouble because people think i'm laughing at them. sometimes i am, but more often i'm only laughing at myself. i'm an entrepreneur (like everyone else in sf, it seems) and i love what i do. i enjoy parties and downtime in equal measure. intelligence really turns me on and i love people who can teach me new things. building awesome stuff. figuring out what's important. having adventures. looking for treasure. imagining random shit. laughing at aforementioned random shit. being goofy. articulating what i think and feel. convincing people i'm right. admitting when i'm wrong. i'm also pretty good at helping people think through problems; my friends say i give good advice. and when i don't have a clue how to help, i will say: i give pretty good hug. i have a big smile. i also get asked if i'm wearing blue-coloured contacts (no). books: to kill a mockingbird, lord of the rings, 1984, the farseer trilogy. music: the beatles, frank sinatra, john mayer, jason mraz, deadmau5, andrew bayer, everything on anjunadeep records, bach, satie. tv shows: how i met your mother, scrubs, the west wing, breaking bad. movies: star wars, the godfather pt ii, 500 days of summer, napoleon dynamite, american beauty, lotr food: thai, vietnamese, shanghai dumplings, pizza! like everyone else, i love my friends and family, and need hugs, human contact, water and sunshine. let's take that as given. 1. something to build 2. something to sing 3. something to play on (my guitar would be first choice) 4. something to write/draw on 5. a big goal worth dreaming about 6. something to laugh at what my contribution to the world is going to be and/or should be. and what's for breakfast. i love breakfast. out with my friends! i cried on my first day at school because a bird shat on my head. true story. you're awesome.

clean_word_tokens \
0   [would, love, think,

kind, intellectual, either, dumbest, smart, guy, smartest, dumb, guy, cant, say, tell, difference, love, talk, idea, concept, forge, odd, metaphor, instead, reciting, cliche, like, simularities, friend, mine, house, underwater, salt, mine, favorite, word, salt, way, weird, choice, know, thing, life, better, metaphor, seek, make, little, better, everyday, productively, lazy, way, got, tired, tying, shoe, considered, hiring, five, year, old, would, probably, tie, shoe, decided, wear, leather, shoe, dress, shoe, love, really, serious, really, deep, conversation, really, silly, stuff, willing, snap, light, hearted, rant, kiss, dont, funny, able, make, laugh, able, bend, spoon, mind, telepathically, make, smile, still, …]
1 [im, ashamed, much, writing, public, text, online, dating, site, make, pleasantly, uncomfortable, ill, try, earnest, possible, noble, endeavor, standing, naked, world, ive, lived, san, francisco, year, love, find, frustrated, deficit, lot, great, friend, acquaintance, increase, apprehension, put, anything, site, im, feeling, like, meeting, new, people, arent, friend, friend, okay, friend, friend, chance, make, complex, filtering, process, multiple, choice, question, lifestyle, statistic, photo, scanning, indulgent, blurb, text, without, moving, quickly, another, search, result, probably, already, cultural, peer, people, removed, first, thought, say, little, possible, avoid, seems, silly, far, culture, go, im, definitely, weird, side, spectrum, dont, exactly, wear, sleeve, get, talking, …]
2
[work, library, go, school, reading, thing, written, old, dead, people, playing, synthesizer, organizing, book, according, library, congress, classification, system, socially, awkward, best, bataille, celine, beckett, lynch, jarmusch, rw, fassbender, twin, peak, fishing, w, john, joy, division, throbbing, gristle, cabaret, voltaire, vegetarian, pho, coffee, cat, german, philosophy, feel, inclined]
3
[hey, hows, going, currently, vague, profile, know, come, soon, looking, meet, new, folk, outside, circle, friend, im, pretty, responsive, reply, tip, feel, free, drop, line, cheer, work, work, work, work, play, creating, imagery, look, httpbagsbrownblogspotcom, httpstayrulyblogspotcom, smile, lot, inquisitive, nature, music, band, rapper, musician, moment, thee, oh, see, forever, wutang, book, artbooks, day, audiobooks, collection, thick, thanks, audible, show, live, one, food, stellar, friend, whenever, movie, tv, podcast, radiolab, american, life, moth, joe, rogan, champ]
4 [im, australian, living, san, francisco, dont, hold, spend, day, trying, build, cool, stuff, company, speak, mandarin, known, bust, chinese, song, karaoke, im, pretty, cheeky, someone, asked, meant, something, arse, find, really, funny, im, little, oddball, wild, imagination, like, think, improbable, reason, people, thing, fun, love, laugh, look, reason, occasionally, get, trouble, people, think, im, laughing, sometimes, often, im, laughing, im, entrepreneur, like, everyone, else, sf, seems, love, enjoy, party, downtime, equal, measure, intelligence, really, turn, love, people, teach, new, thing, building, awesome, stuff, figuring, whats, important, adventure, looking, treasure, imagining, random, shit, laughing, aforementioned, random, shit,

goofy, articulating, think, feel, …]

clean_word_string

0
would love think kind intellectual either dumbest smart guy smartest dumb guy cant say tell difference love talk idea concept forge odd metaphor instead reciting cliche like simularities friend mine house underwater salt mine favorite word salt way weird choice know thing life better metaphor seek make little better everyday productively lazy way got tired tying shoe considered hiring five year old would probably tie shoe decided wear leather shoe dress shoe love really serious really deep conversation really silly stuff willing snap light hearted rant kiss dont funny able make laugh able bend spoon mind telepathically make smile still work love life cool letting wind blow extra point reading guessing favorite video game hint given yet lastly good attention span currently working international agent freight forwarding company import export domestic know work online class trying better free time perhaps hour worth good book video game lazy sunday making people laugh ranting good salting finding simplicity complexity complexity simplicity way look six foot half asian half caucasian mutt make tough notice blend book absurdistan republic mouse men book made want cry catcher rye prince movie gladiator operation valkyrie producer periscope show borgia arrested development game throne monty python music aesop rock hail mary mallon george thorogood delaware destroyer felt food im anything food water cell phone shelter duality humorous thing trying find someone hang anything except club new california looking someone wisper secret want swept foot tired norm want catch coffee bite want talk philosophy
1 im ashamed much writing public text online dating site make pleasantly uncomfortable ill try earnest possible noble endeavor standing naked world ive lived san francisco year love find frustrated deficit lot great friend acquaintance increase apprehension put anything site im feeling like meeting new people arent friend friend okay friend friend chance make complex filtering process multiple choice question lifestyle statistic photo scanning indulgent blurb text without moving quickly another search result probably already cultural peer people removed first thought say little possible avoid seems silly far culture go im definitely weird side spectrum dont exactly wear sleeve get talking probably become increasingly apparent id like think like everybody else definition certainly people dont see way thats fine people find gravitating towards pretty weird probably make nerdy software musician artist experimenter indulge weirdness like spend time away computer working artwork typically concerned group dynamic communication visual form object technology also record deejay dance noise pop experimental music electronic least studio based besides relatively ego driven activity ive enjoying thing like meditation tai chi try gently flirt ego death improvising different context alternating present decidedly outside moment trying hold rambling intellectual conversation hold said conversation contempt seeking find something transcends critical remaining generous listening using body languageoften performed caricature large gesture outright interpretive dance dry dark raunchy humor large jaw large glass physical thing people comment sufficiently stimulated unmistakable cackle laugh

go direction care describe right maybe ill come back okay cultural matrix get specific like crosshairs worth find reading nonfiction fiction usually kind philosophy art science text silly author ranciere de certeau bataille baudrillard butler stein arendt nietzche zizek etc ill often throw weird new age poppsychology book mix well fiction enjoy little ive read eco perec wallace bolao dick vonnegut atwood delilo etc young rabid asimov reader director find drawn makavejev kuchar jodorowsky herzog hara klein water verhoeven ackerman hitchcock lang gorin goddard miike ohbayashi tarkovsky sokurov warhol etc also like good amount trashy stuff much name definitely enjoy character development happens long form episodic television course episode movie usually cant compete recent tv favorite breaking bad wire dexter true blood prisoner lost fringe smattered sampling vast field music like deejay art ensemble sun ra evan parker lil wayne dj funk mr finger maurizio rob hood dan bell james blake nonesuch recording omar souleyman ethiopiques fela kuti john cage meredith monk robert ashley terry riley yoko ono merzbow tom tom club jit juke bounce hyphy snap crunk bmore kuduro pop noise jazz techno house acid newno wave postpunk etc famous artdancetheater folk might locate sensibility andy warhol bruce nauman yayoi kusama louise bourgeois tino sehgal george kuchar michel duchamp marina abramovic gelatin carolee schneeman gustav metzger mike kelly mike smith andrea fraser gordon mattaclark jerzy grotowski samuel beckett antonin artaud tadeusz kantor anna halperin merce cunningham etc im clearly leaving younger generation contemporary artist many friend local food regular sushi zone chow ppq pagolac ler ro burma superstar minako shalimar delfina pizza rosamunde arinells suppenkuche chaya blue plate golden era etc movement conversation creation contemplation touch humor viewing listening dancing talking drinking performing five year old known boogerman bright open intense silly ironic critical caring generous looking exploration rather finding match predetermined quality im currently fabulous open relationship comfortable

2

work library go school reading thing written old dead people playing synthesizer organizing book according library congress classification system socially awkward best bataille celine beckett lynch jarmusch rw fassbender twin peak fishing w john joy division throbbing gristle cabaret voltaire vegetarian pho coffee cat german philosophy feel inclined

3

hey hows going currently vague profile know come soon looking meet new folk outside circle friend im pretty responsive reply tip feel free drop line cheer work work work work play creating imagery look httpbagsbrownblogspotcom httpstayrulyblogspotcom smile lot inquisitive nature music band rapper musician moment thee oh see forever wutang book artbooks day audiobooks collection thick thanks audible show live one food stellar friend whenever movie tv podcast radiolab american life moth joe rogan champ

4

im australian living san francisco dont hold spend day trying build cool stuff company speak mandarin known bust chinese song karaoke im pretty cheeky someone asked meant something arse find really funny im little oddball wild imagination like think improbable reason people thing fun love laugh look reason

```
occasionally get trouble people think im laughing sometimes often im laughing im
entrepreneur like everyone else sf seems love enjoy party downtime equal measure
intelligence really turn love people teach new thing building awesome stuff
figuring whats important adventure looking treasure imagining random shit
laughing aforementioned random shit goofy articulating think feel convincing
people im right admitting im wrong im also pretty good helping people think
problem friend say give good advice dont clue help say give pretty good hug big
smile also get asked im wearing bluecoloured contact book kill mockingbird lord
ring farseer trilogy music beatles frank sinatra john mayer jason mraz andrew
bayer everything anjunadeep record bach satie tv show met mother scrub west wing
breaking bad movie star war godfather pt ii day summer napoleon dynamite
american beauty lotr food thai vietnamese shanghai dumpling pizza like everyone
else love friend family need hug human contact water sunshine let take given
something build something sing something play guitar would first choice
something writedraw big goal worth dreaming something laugh contribution world
going andor whats breakfast love breakfast friend cried first day school bird
shat head true story youre awesome

[5 rows x 27 columns]
```

The dataset structure confirms that preprocessing was successful: all generated text, grouping, and readability variables are fully populated for all 50,834 profiles, and derived measures align with the essay content.

3.0.1 Education Distribution

[8]:
```
# Distribution of Education Levels
# Since our research question focuses on how writing varies
# across education levels, we first examine how many users
# fall into each education category.

df_cupid['education'].value_counts()
df_cupid['education'].value_counts(normalize=True)
```

[8]:
```
education
graduated from college/university    0.471318
graduated from masters program       0.176280
working on college/university        0.112366
working on masters program           0.033108
graduated from two-year college      0.030118
graduated from high school           0.028091
graduated from ph.d program          0.025023
graduated from law school            0.022072
working on two-year college          0.021128
dropped out of college/university    0.019574
working on ph.d program              0.019337
graduated from med school            0.008774
working on law school                0.005292
```

```
two-year college                    0.004367
working on med school               0.004170
dropped out of two-year college     0.003757
dropped out of masters program      0.002754
masters program                     0.002675
dropped out of ph.d program         0.002498
dropped out of high school          0.002007
high school                         0.001888
working on high school              0.001711
ph.d program                        0.000511
law school                          0.000374
dropped out of law school           0.000354
dropped out of med school           0.000236
med school                          0.000216
Name: proportion, dtype: float64
```

3.0.2 Catagorize Educational Levels

[9]:
```
df_cupid['education'].unique()
print(df_cupid['education'].value_counts(dropna=False))
```

```
education
graduated from college/university    23959
graduated from masters program        8961
working on college/university         5712
working on masters program            1683
graduated from two-year college       1531
graduated from high school            1428
graduated from ph.d program           1272
graduated from law school             1122
working on two-year college           1074
dropped out of college/university      995
working on ph.d program                983
graduated from med school              446
working on law school                  269
two-year college                       222
working on med school                  212
dropped out of two-year college        191
dropped out of masters program         140
masters program                        136
dropped out of ph.d program            127
dropped out of high school             102
high school                             96
working on high school                  87
ph.d program                            26
law school                              19
dropped out of law school               18
dropped out of med school               12
```

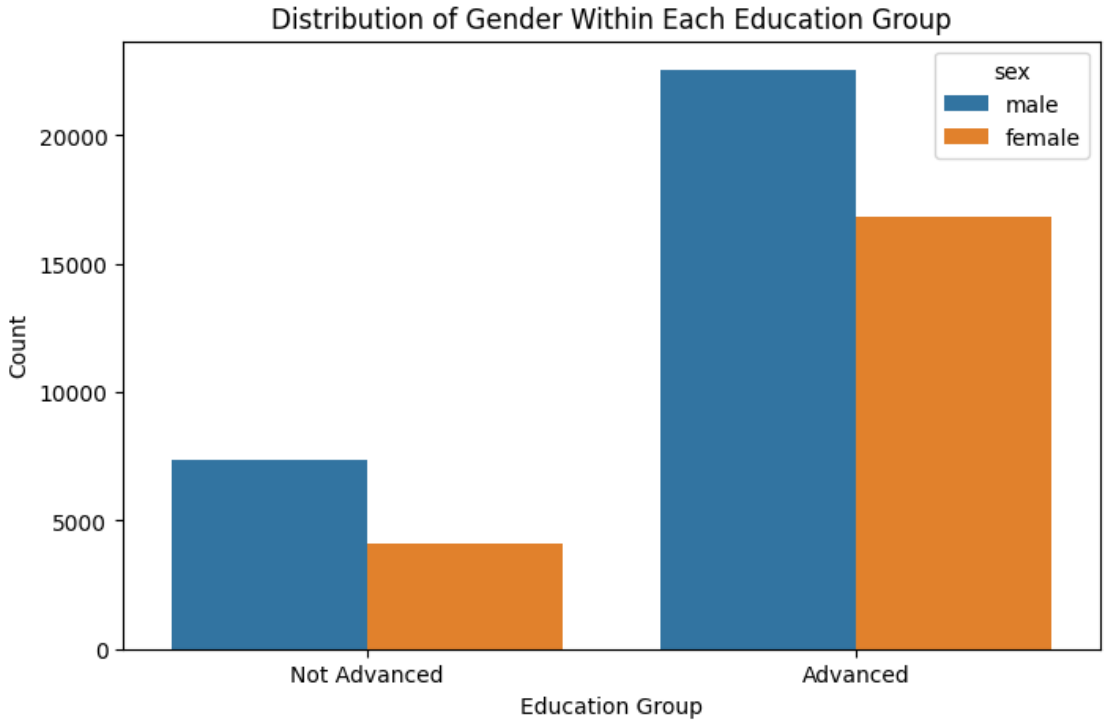
```
med school                                11
Name: count, dtype: int64
```

Interpretation: This ranks entries that fall under different academic categories (e.g., "graduated from college/university," "working on masters program", etc). However, some categories are ambiguous (like "college/university") and some are clearly non-academic (e.g., "graduated from space camp"). We decided to remove those subgroups before we start the analysis.

3.0.3 Distribution of Gender Within Each Education Group

[10]:
```
# Create a new data frame:
grouped = df_cupid.groupby(['education_group', 'sex'])
```

[11]:
```
plt.figure(figsize=(8,5))
sns.countplot(data=df_cupid, x='education_group', hue='sex')
plt.title("Distribution of Gender Within Each Education Group")
plt.xlabel("Education Group")
plt.ylabel("Count")
plt.show()
```



Interpretation: After removing education categories that were unrelated to academic degrees (e.g., "graduated from space camp") or too ambiguous to classify (e.g., "college/university"), we grouped the remaining responses into two categories: Advanced and Not Advanced. After cleaning the data, most users fell into the Advanced category, with 16,843 female and 22,542 male individuals classified as Advanced. In contrast, the Not Advanced category included 4,119 female and 7,330

male users. Men appeared more often in both groups simply because there are more male profiles in the dataset overall.

3.0.4 Distribution of Education Level Within Each Gender

[12]:
```
df_cupid['essay_length'] = df_cupid['all_essays'].apply(lambda x: len(str(x).
 ↪split()))

print(df_cupid.groupby(['education_group', 'sex'])['essay_length'].describe())
```

```
                          count        mean         std  min    25%    50%  \
education_group sex
Advanced        female  16843.0  384.012409  273.898832  0.0  199.0  334.0
                male    22542.0  374.797667  297.222310  0.0  180.0  317.0
Not Advanced    female   4119.0  348.316824  306.986860  0.0  149.0  279.0
                male     7330.0  330.307913  314.299045  0.0  127.0  265.0

                           75%      max
education_group sex
Advanced        female   510.0   3382.0
                male     502.0  10602.0
Not Advanced    female   457.0   4446.0
                male     446.0   7446.0
```

Interpretation: Across all groups, users with advanced education write slightly longer essays, and women write longer essays than men. However, essay lengths vary widely, with a few extremely long entries creating very large maximum values.

3.0.5 Validation for Essay Length Statistics

[13]:
```
df_cupid['essay_length'] = df_cupid['all_essays'].apply(lambda x: len(str(x).
 ↪split()))



groups = [
    ('Advanced', 'male'),
    ('Advanced', 'female'),
    ('Not Advanced', 'male'),
    ('Not Advanced', 'female')
]

sample_frames = []

for ed, sex in groups:
    subset = df_cupid[(df_cupid['education_group'] == ed) &
                      (df_cupid['sex'] == sex)]
    sample = subset.sample(10)
    sample_frames.append(sample)
```

```
df_sample = pd.concat(sample_frames)

print(df_sample.groupby(['education_group', 'sex'])['essay_length'].describe())
```

```
                        count   mean         std    min     25%    50%  \
education_group sex
Advanced        female   10.0  307.7  194.479962   80.0  153.25  267.0
                male     10.0  331.4  127.062714  149.0  219.25  344.0
Not Advanced    female   10.0  344.6  211.181965   21.0  210.75  336.0
                male     10.0  250.1  317.144709    0.0   30.00  167.0

                           75%     max
education_group sex
Advanced        female  440.75   688.0
                male    429.50   510.0
Not Advanced    female  530.50   648.0
                male    292.50  1006.0
```

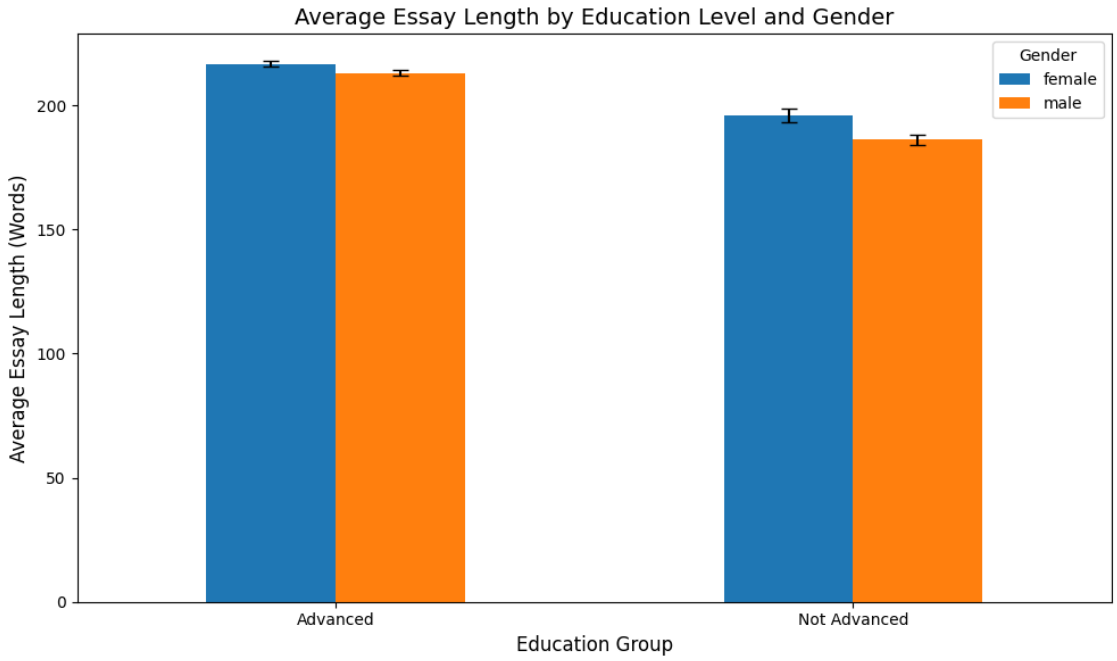
Interpretation: Using random samples from each subgroup, we find that the mean, median, and range of essay lengths remain consistent with the full dataset. The same directional patterns appear. Advanced users write longer essays, women write longer than men, and there is substantial variability. This validation supports that our earlier summary statistics are reliable.

3.0.6 Average Essay Length by Education Level and Gender

[14]:
```
# 1. Compute essay length
df_cupid['essay_length'] = df_cupid['clean_word_tokens'].apply(len)

# 2. Group statistics
group_stats = df_cupid.groupby(['education_group', 'sex'])['essay_length'].
 ↪agg(['mean','std','count'])

# Standard error of the mean
group_stats['sem'] = group_stats['std'] / np.sqrt(group_stats['count'])

# Pivot tables for plotting
means = group_stats['mean'].unstack()
errors = group_stats['sem'].unstack()

# 3. Plot with error bars
plt.figure(figsize=(10,6))
means.plot(kind='bar', yerr=errors, capsize=5, figsize=(10,6))

plt.title("Average Essay Length by Education Level and Gender", fontsize=14)
plt.xlabel("Education Group", fontsize=12)
plt.ylabel("Average Essay Length (Words)", fontsize=12)
plt.xticks(rotation=0)
```

```
plt.legend(title="Gender")
plt.tight_layout()
plt.show()
```

```
<Figure size 1000x600 with 0 Axes>
```



Interpretation: Across all groups, users with advanced education levels wrote slightly longer essays than those without advanced education. Women also wrote somewhat longer essays than men within the same education category. Although the differences are not large, the pattern is consistent in both groups. The error bars in the figure are extremely small because each subgroup contains thousands of users, which makes the average estimates very stable. Overall, both education level and gender show small but reliable associations with essay length.

3.0.7 Frequency Distribution

[15]:
```
# 1. Split Data into Groups

adv_male = df_cupid[(df_cupid['education_group'] == 'Advanced') &␣
 ↪(df_cupid['sex'] == 'male')]
adv_female = df_cupid[(df_cupid['education_group'] == 'Advanced') &␣
 ↪(df_cupid['sex'] == 'female')]
notadv_male = df_cupid[(df_cupid['education_group'] == 'Not Advanced') &␣
 ↪(df_cupid['sex'] == 'male')]
notadv_female = df_cupid[(df_cupid['education_group'] == 'Not Advanced') &␣
 ↪(df_cupid['sex'] == 'female')]
```

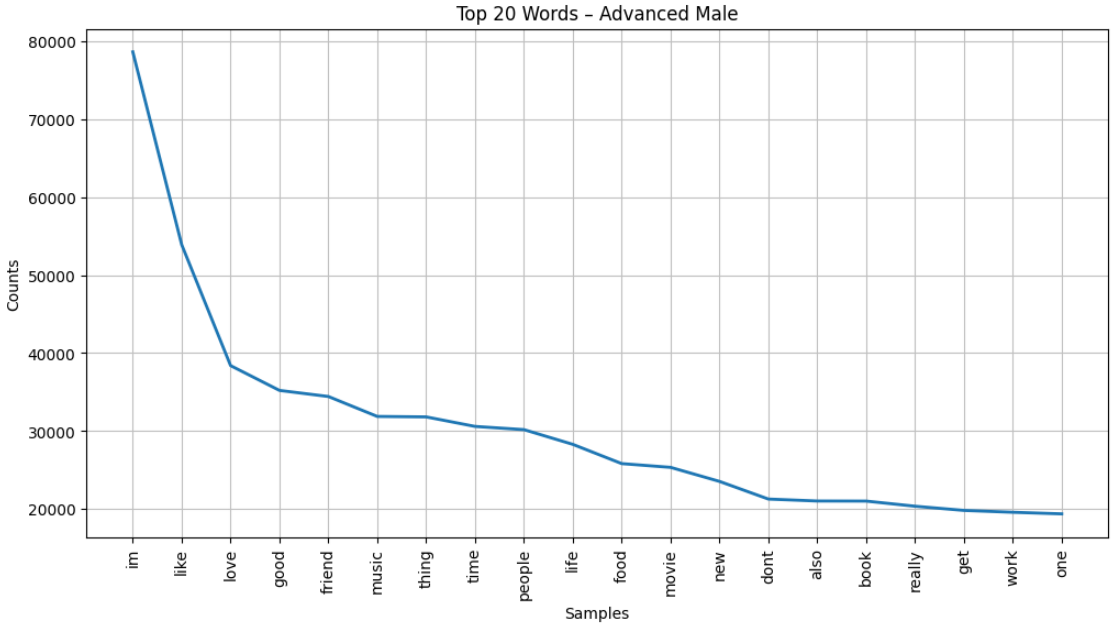
```python
# 2. Function to extract cleaned tokens for a group
def extract_tokens(group):
    return [
        word
        for token_list in group['clean_word_tokens']   # <<- USE CLEAN TOKENS
        for word in token_list
    ]

# 3. Extract tokens for each group

tokens_adv_male = extract_tokens(adv_male)
tokens_adv_female = extract_tokens(adv_female)
tokens_notadv_male = extract_tokens(notadv_male)
tokens_notadv_female = extract_tokens(notadv_female)

# 4. Frequency distributions

fdist_adv_male = FreqDist(tokens_adv_male)
fdist_adv_female = FreqDist(tokens_adv_female)
fdist_notadv_male = FreqDist(tokens_notadv_male)
fdist_notadv_female = FreqDist(tokens_notadv_female)

# 5. Print Top 20 Words in Each Group

print("Advanced - Male:", fdist_adv_male.most_common(20))
print("\nAdvanced - Female:", fdist_adv_female.most_common(20))
print("\nNot Advanced - Male:", fdist_notadv_male.most_common(20))
print("\nNot Advanced - Female:", fdist_notadv_female.most_common(20))
```

```
Advanced - Male: [('im', 78612), ('like', 53893), ('love', 38391), ('good',
35212), ('friend', 34431), ('music', 31876), ('thing', 31816), ('time', 30596),
('people', 30183), ('life', 28296), ('food', 25827), ('movie', 25347), ('new',
23547), ('dont', 21284), ('also', 21034), ('book', 21024), ('really', 20365),
('get', 19828), ('work', 19590), ('one', 19387)]

Advanced - Female: [('im', 59107), ('love', 45893), ('like', 39493), ('friend',
32032), ('good', 27274), ('life', 24247), ('thing', 24066), ('people', 23983),
('time', 23814), ('music', 23658), ('food', 20733), ('movie', 19618), ('new',
19428), ('book', 18037), ('also', 16752), ('dont', 16361), ('really', 15992),
('work', 14304), ('get', 14267), ('one', 14229)]

Not Advanced - Male: [('im', 27559), ('like', 17269), ('love', 12459), ('music',
10478), ('good', 10349), ('friend', 9792), ('time', 9514), ('people', 9435),
('thing', 9354), ('movie', 8675), ('life', 8143), ('dont', 7622), ('food',
7488), ('really', 6974), ('know', 6670), ('get', 6588), ('want', 6454), ('also',
6163), ('book', 6067), ('one', 5721)]

Not Advanced - Female: [('im', 16656), ('love', 11042), ('like', 9978),
```

```
('friend', 6181), ('people', 5955), ('good', 5562), ('music', 5492), ('time',
5485), ('thing', 5470), ('movie', 5066), ('dont', 5050), ('life', 4912),
('food', 4544), ('really', 4301), ('want', 4021), ('book', 3973), ('know',
3959), ('get', 3799), ('also', 3758), ('one', 3360)]
```

3.0.8 Frequency Distribution for Advanced Male

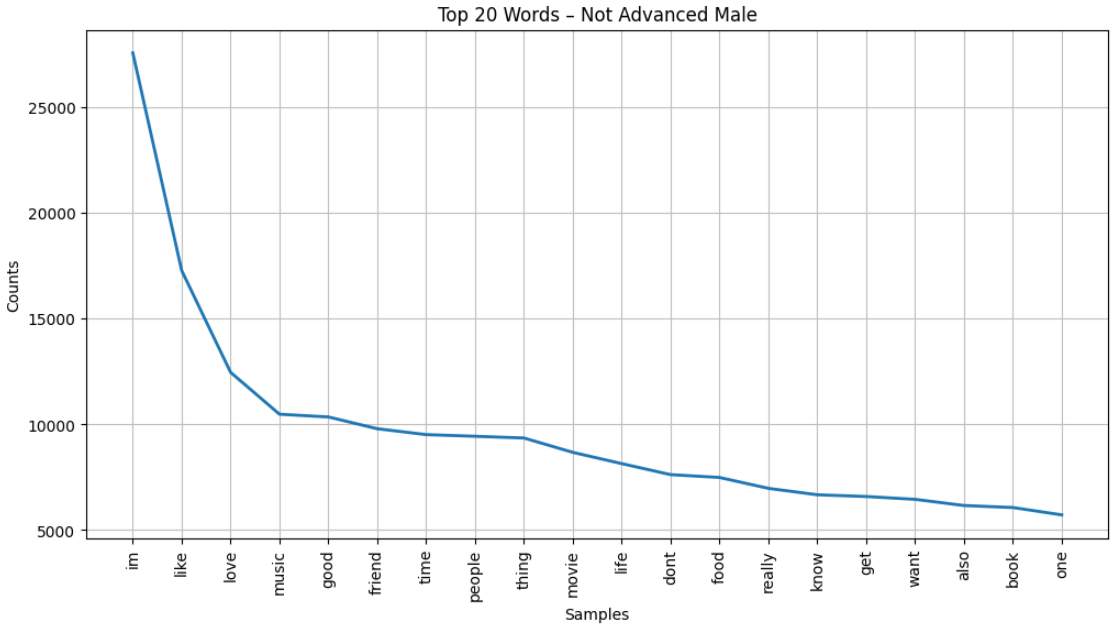
[16]:
```
plt.figure(figsize=(12,6))
fdist_adv_male.plot(20, title="Top 20 Words - Advanced Male")
plt.show()
```



Interpretation: For advanced male users, the most common words include “im,” “like,” “love,” “good,” “friend,” “music,” “time,” “people,” and “life.” Their writing centers on describing themselves, their interests, and general life themes. The distribution shows a steep drop after the most frequent word (“im”), which is typical in frequency plots. Overall, their language is casual and focused on personality and hobbies.

3.0.9 Frequency Distribution for Not Advance Male

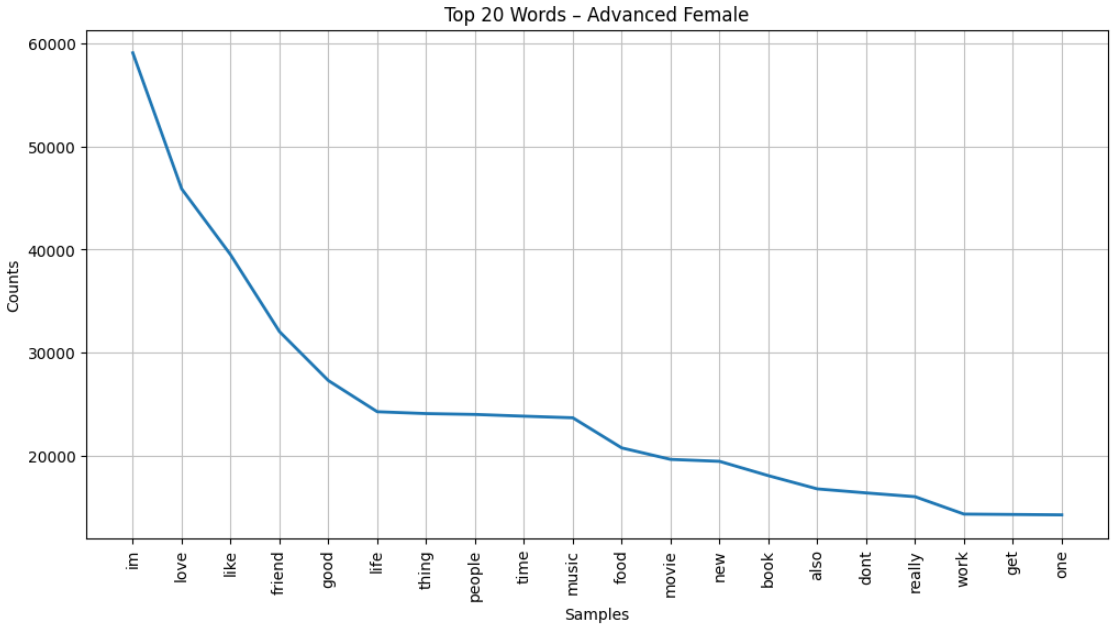
[17]:
```
plt.figure(figsize=(12,6))
fdist_notadv_male.plot(20, title="Top 20 Words - Not Advanced Male")
plt.show()
```



Interpretation: For not advanced male users, frequent words include “im,” “like,” “love,” “music,” “good,” “friend,” “time,” “people,” “thing,” and “movie.” Their essays also focus on everyday interests, activities, and self-description. Similar to the advanced male group, the pattern shows a sharp decline after the top word. The presence of words like “movie,” “thing,” and “food” suggests a slightly more concrete, day-to-day style.

3.0.10 Frequency Distribution for Advance Female

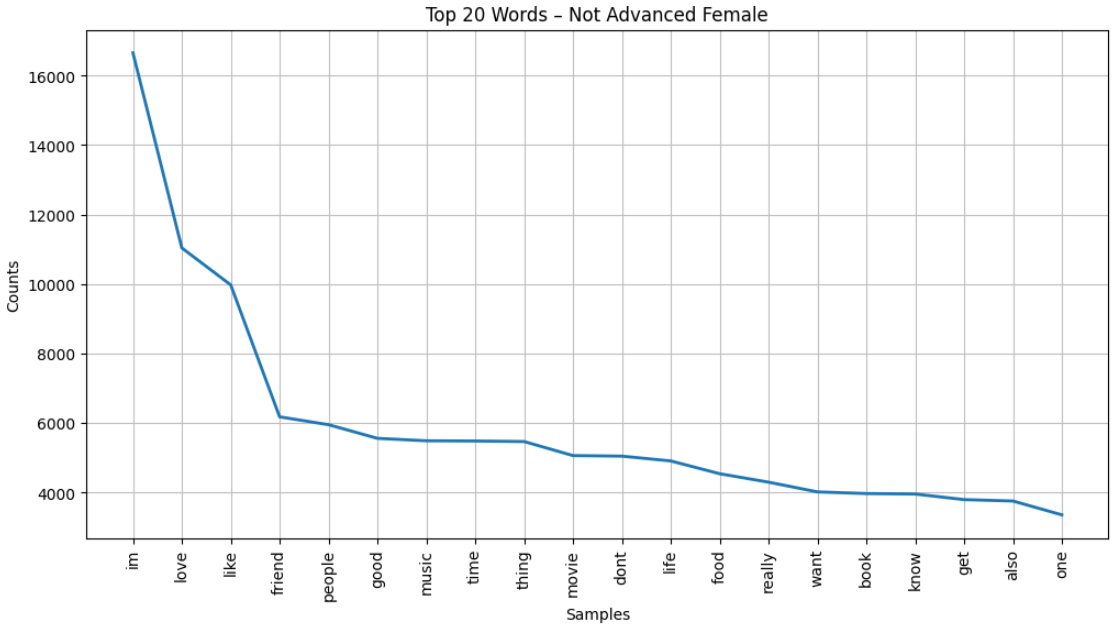
[18]:
```
plt.figure(figsize=(12,6))
fdist_adv_female.plot(20, title="Top 20 Words - Advanced Female")
plt.show()
```



Interpretation: For advanced female users, common words are "im," "love," "like," "friend," "good," "life," "thing," "people," "music," and "food." Their writing frequently highlights emotions, relationships, and personal interests. The distribution again shows the typical steep drop-off. Compared to males, the advanced female group uses words related to feelings and relationships ("love," "friend") more prominently.

3.0.11 Frequency Distribution for Not Advance Female

[19]:
```
plt.figure(figsize=(12,6))
fdist_notadv_female.plot(20, title="Top 20 Words - Not Advanced Female")
plt.show()
```



Interpretation: For not advanced female users, the top words include "im," "love," "like," "friend," "people," "good," "music," "time," "thing," and "movie." Their essays tend to focus on friendships, emotions, and hobbies. The pattern is similar to the other groups, with the most frequent words appearing much more than the rest. This group also uses several everyday-life words such as "movie," "food," and "thing."

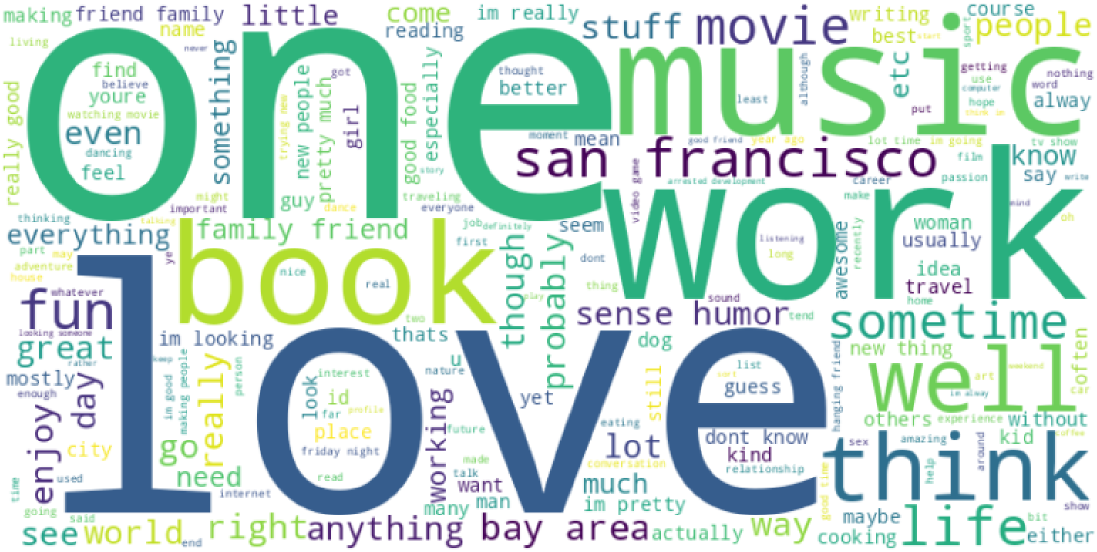
Interpretation (all four groups): Across all four groups, the overall writing style is very similar: users mainly talk about themselves, their interests, friendships, and everyday life. However, there are a few subtle differences. Advanced users tend to include slightly more abstract or identity-oriented words (like "life," "book," and "work"), while not advanced users use more concrete, daily-activity words (like "movie," "food," and "thing"). Women, regardless of education level, use more relationship- or emotion-related words ("love," "friend," "really"), whereas men use slightly more activity or interest-based words ("music," "time," "good"). These differences are small, but they show some variation in how different groups describe themselves in their dating profiles.

3.0.12 Word Cloud

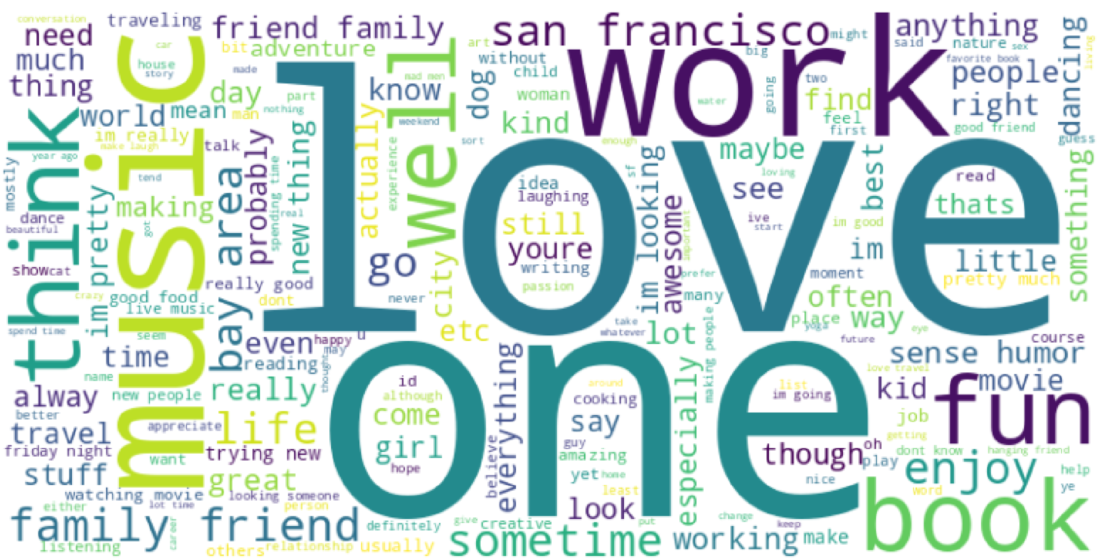
[20]:
```
# Helper function to extract CLEAN tokens for a group
def extract_tokens(group):
    return [
        word
        for token_list in group['clean_word_tokens']    # <<- FIXED: use cleaned␣
 ↪tokens
        for word in token_list
    ]

# Split groups
```

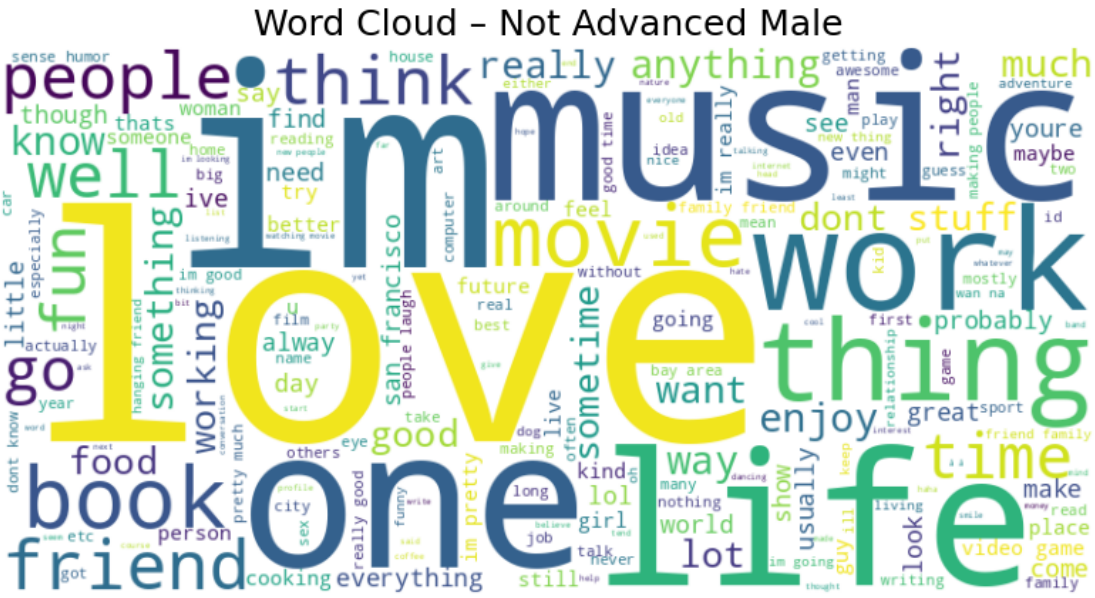
```
adv_male = df_cupid[(df_cupid['education_group'] == 'Advanced') &␣
 ↪(df_cupid['sex'] == 'male')]
adv_female = df_cupid[(df_cupid['education_group'] == 'Advanced') &␣
 ↪(df_cupid['sex'] == 'female')]
notadv_male = df_cupid[(df_cupid['education_group'] == 'Not Advanced') &␣
 ↪(df_cupid['sex'] == 'male')]
notadv_female = df_cupid[(df_cupid['education_group'] == 'Not Advanced') &␣
 ↪(df_cupid['sex'] == 'female')]

# Extract tokens for each group (clean!)
tokens_adv_male = extract_tokens(adv_male)
tokens_adv_female = extract_tokens(adv_female)
tokens_notadv_male = extract_tokens(notadv_male)
tokens_notadv_female = extract_tokens(notadv_female)

# Convert lists to clean strings
text_adv_male = " ".join(tokens_adv_male)
text_adv_female = " ".join(tokens_adv_female)
text_notadv_male = " ".join(tokens_notadv_male)
text_notadv_female = " ".join(tokens_notadv_female)

# Function to plot a word cloud
def plot_wc(text, title):
    wc = WordCloud(width=800, height=400, background_color='white').
 ↪generate(text)
    plt.figure(figsize=(10,5))
    plt.imshow(wc, interpolation='bilinear')
    plt.axis('off')
    plt.title(title, fontsize=18)
    plt.show()

plot_wc(text_adv_male, "Word Cloud - Advanced Male")
plot_wc(text_adv_female, "Word Cloud - Advanced Female")
plot_wc(text_notadv_male, "Word Cloud - Not Advanced Male")
plot_wc(text_notadv_female, "Word Cloud - Not Advanced Female")
```

Word Cloud – Advanced Male



Word Cloud – Advanced Female





Interpretation: The word clouds provide another way to visualize word frequency, with larger words representing more commonly used terms in each subgroup. Across all groups, shared high-frequency words such as “im,” “love,” “like,” “music,” and “life” appear prominently, which matches the earlier frequency plots. Even though the overall patterns are similar, the word clouds highlight some subtle stylistic differences: advanced users tend to show more words related to interests or identity, such as “work,” “book,” and “think,” while not advanced users include slightly more everyday or concrete terms like “movie” and “friend.” Female users across education levels feature

more emotionally oriented words such as "love" and "life," whereas male users show more hobby- or activity-related words like "music" and "work." These distinctions are minor, but the visualization helps make small differences in emphasis more noticeable.

3.0.13 Descriptive Statistics Grouped by Gender and Education

[21]:
```
df_cupid.groupby(['education_group', 'sex']).describe()
```

[21]:
```
                             age                                             \
                           count       mean        std   min   25%   50%   75%
education_group sex
Advanced        female   16843.0  33.959983   9.566085  18.0  27.0  31.0  38.0
                male     22542.0  33.034824   8.753483  18.0  27.0  30.0  37.0
Not Advanced    female    4119.0  28.108521  10.099343  18.0  21.0  24.0  31.0
                male      7330.0  28.592769   9.026087  18.0  22.0  26.0  32.0

                                height              …   income             \
                           max    count       mean  …      75%        max
education_group sex                                 …
Advanced        female    69.0  16843.0  65.154901  …     -1.0  1000000.0
                male     109.0  22542.0  70.473072  …     -1.0  1000000.0
Not Advanced    female    69.0   4119.0  64.983006  …     -1.0  1000000.0
                male      69.0   7330.0  70.316235  …  20000.0  1000000.0

                        essay_length                                       \
                               count        mean         std  min    25%
education_group sex
Advanced        female       16843.0  216.753963  150.504534  0.0  115.0
                male         22542.0  213.073818  164.332177  0.0  104.0
Not Advanced    female        4119.0  196.088614  169.446136  0.0   85.0
                male          7330.0  186.231924  174.965334  0.0   73.0


                           50%    75%     max
education_group sex
Advanced        female   190.0  287.0  1748.0
                male     183.0  286.0  5724.0
Not Advanced    female   157.0  257.0  2260.0
                male     150.0  253.0  3867.0

[4 rows x 32 columns]
```

Reasoning: Summarizing the data by gender and education provides a clearer picture of how the main variables in the dataset, such as age, essay length, and the readability scores, are distributed across the groups included in the analysis. Looking at these statistics helps establish basic patterns, including the typical ranges and averages within each gender and education category, and gives context for how the writing measures appear within these subgroups. This grouped summary serves as an initial overview of the dataset and helps guide the later stages of analysis.

4 Readability/Analysis

4.0.1 Test Readability on a sample essay

[22]:
```
# Select a sample row to inspect
sample_row = 6

# Print basic info about the selected profile
print("Sex:", df_cupid.at[sample_row, 'sex'])
print("Education Group:", df_cupid.at[sample_row, 'education_group'])

# Get the cleaned essay text for readability testing
test_data = df_cupid.at[sample_row, 'clean_readability_text']

# Show the essay so we can see what kind of text the readability functions will␣
 ↪analyze
print("\nESSAY TEXT:\n", test_data)
```

```
Sex: female
Education Group: Advanced

ESSAY TEXT:
 writing. meeting new people, spending time with friends, seeing films, going to
literary events and lectures, sifting through bookstores and thrift stores,
exploring the city. i also work full time at an interactive agency. remembering
people's birthdays, sending cards, being thoughtful, arm wrestling i'm rather
approachable (a byproduct of being from a small town in the midwest). i like:
alphabetized lists, aquariums, autobiographies, beer on tap, ben folds, biking,
brunch, citrus, cocktails, color, comfort food, craft projects, dancing, design,
diy, essays, fabric stores, field trips, flea markets, foreign films, glee,
good, hammond organs, helping lost tourists, indie rock, ice cream, languages,
lectures, letterpress, libraries, literary fiction, live shows, mad men, martha
stewart living, memoir, mix tapes, non-fiction, npr, plants, puns, sewing, short
stories, siestas, singer-songwriters, spicy food, stationery, storytelling,
sufjan stevens, talking to strangers, tea, tegan and sara, the office, 30 rock,
travel, quilts, quirky movies, wes anderson, wine, writing, yoga. friends,
family, notebook/pen, books, music, travel things that amuse and inspire me out
and about or relaxing at home with a good book or netflix
```

Reasoning: The sample essay is printed to verify the text cleaning and to see the style and structure before running readability metrics. It looks like a valid sample since it does have an essay column.

[23]:
```
# Show the text being analyzed
print("Text used for the redability")
print(test_data[:500])

# Compute readability metrics for the sample essay
print("Flesch Reading Ease:", textstat.flesch_reading_ease(test_data))
```

```
print("Flesch-Kincaid Grade:", textstat.flesch_kincaid_grade(test_data))
print("Gunning Fog:", textstat.gunning_fog(test_data))
print("SMOG:", textstat.smog_index(test_data))
print("Dale-Chall:", textstat.dale_chall_readability_score(test_data))

# Basic text structure counts
print("Syllables:", textstat.syllable_count(test_data))
print("Lexicon (words):", textstat.lexicon_count(test_data, removepunct=True))
print("Sentences:", textstat.sentence_count(test_data))
print("Polysyllables:", textstat.polysyllabcount(test_data))


print("End of the redability check")
```

```
Text used for the redability
writing. meeting new people, spending time with friends, seeing films, going to
literary events and lectures, sifting through bookstores and thrift stores,
exploring the city. i also work full time at an interactive agency. remembering
people's birthdays, sending cards, being thoughtful, arm wrestling i'm rather
approachable (a byproduct of being from a small town in the midwest). i like:
alphabetized lists, aquariums, autobiographies, beer on tap, ben folds, biking,
brunch, citrus, cocktails, c
Flesch Reading Ease: 27.383136363636368
Flesch-Kincaid Grade: 18.184590909090918
Gunning Fog: 18.852727272727275
SMOG: 15.112257680678326
Dale-Chall: 13.815715454545455
Syllables: 299
Lexicon (words): 176
Sentences: 5
Polysyllables: 22
End of the redability check
```

Interpretation: The readability test seems to be working on the sample. The text we checked has long sentences and many complex words, and the scores show that it is difficult to read. The results indicate that the writing is at a college or graduate-school level, which matches what we see in the sample text. Since the previewed text matches the values produced by the readability formulas, we can confidently say that the readability code is functioning as expected.

4.0.2 Add readability columns to okcupid data set

[24]:
```
df_cupid['syllable_count'] = df_cupid['clean_readability_text'].apply(textstat.
 ↪syllable_count)
df_cupid['lexicon_count'] = df_cupid['clean_readability_text'].apply(lambda x:␣
 ↪textstat.lexicon_count(x, removepunct=True))
df_cupid['sentence_count'] = df_cupid['clean_readability_text'].apply(textstat.
 ↪sentence_count)
```

```
df_cupid['polysyllab_count'] = df_cupid['clean_readability_text'].
 ↪apply(textstat.polysyllabcount)

df_cupid['flesch_reading_ease'] = df_cupid['clean_readability_text'].
 ↪apply(textstat.flesch_reading_ease)
df_cupid['flesch_kincaid_grade'] = df_cupid['clean_readability_text'].
 ↪apply(textstat.flesch_kincaid_grade)
df_cupid['gunning_fog'] = df_cupid['clean_readability_text'].apply(textstat.
 ↪gunning_fog)
df_cupid['smog_index'] = df_cupid['clean_readability_text'].apply(textstat.
 ↪smog_index)
df_cupid['dale_chall'] = df_cupid['clean_readability_text'].apply(textstat.
 ↪dale_chall_readability_score)
```

Reasoning: We compute these readability metrics for the entire dataset so we can analyze patterns in writing complexity across all profiles, not just the sample.

4.0.3 Calculate proportion measures

[25]:
```
# Proportion of polysyllabic words
df_cupid['prop_polysyll'] = df_cupid['polysyllab_count'] /␣
 ↪df_cupid['lexicon_count']

# Average number of words per sentence
df_cupid['words_per_sentence'] = df_cupid['lexicon_count'] /␣
 ↪df_cupid['sentence_count']
```

4.0.4 Describe readability variables

[26]:
```
df_cupid.describe()[[
    'syllable_count', 'lexicon_count', 'sentence_count',
    'polysyllab_count', 'flesch_reading_ease', 'flesch_kincaid_grade',
    'gunning_fog', 'smog_index', 'dale_chall',
    'prop_polysyll', 'words_per_sentence'
]]
```

[26]:
```
       syllable_count  lexicon_count  sentence_count  polysyllab_count  \
count    50834.000000   50834.000000    50834.000000      50834.000000
mean       537.966479     366.200043       25.049868         38.561829
std        430.377875     291.164158       19.607763         33.649425
min          0.000000       0.000000        0.000000          0.000000
25%        255.000000     173.000000       12.000000         17.000000
50%        455.000000     310.000000       21.000000         32.000000
75%        720.000000     490.000000       34.000000         52.000000
max      17501.000000   10486.000000      584.000000       1781.000000

       flesch_reading_ease  flesch_kincaid_grade   gunning_fog    smog_index  \
```

count	50834.000000	50834.000000	50834.000000	50834.000000
mean	64.336595	7.682970	9.504202	9.917154
std	19.318410	4.272319	4.369075	2.643864
min	-1486.180000	-3.400000	0.000000	0.000000
25%	60.162346	5.975972	7.794787	8.990254
50%	67.536095	7.279113	9.175186	9.984396
75%	73.541040	8.831486	10.813664	11.088447
max	121.220000	246.648730	253.714286	41.874830

	dale_chall	prop_polysyll	words_per_sentence
count	50834.000000	49318.000000	49325.000000
mean	8.770255	0.105460	15.733999
std	2.041674	0.038693	9.466284
min	0.000000	0.000000	0.000000
25%	8.165586	0.083799	11.833333
50%	8.852864	0.102804	14.230769
75%	9.631562	0.123635	17.400000
max	37.090087	1.000000	630.000000

Interpretation: Most of the writing in the dataset appears normal and fairly easy to read. A small number of profiles, however, show negative readability scores. A negative Flesch Reading Ease score is unusual but still possible when the text is very irregular or highly simplified. A negative Flesch–Kincaid Grade score usually happens when the writing is extremely short, such as one-word sentences, very short lines, or emoji-only profiles. This style is common in dating apps, so some negative scores may simply reflect the way people write in that context.

At the same time, negative values can also appear when something in the preprocessing does not work correctly. Examples include missing punctuation, lines that are not recognized as sentences, or profiles that contain almost no text at all. For this reason, it is important to examine the profiles that produced negative scores. This will help us determine whether the values are caused by the natural simplicity of dating-app writing or by possible issues in cleaning or processing the text.

4.0.5 Investingating negative values for Flesch Reading Ease and Flesch–Kincaid Grade

[27]:
```
# extract the sample essay that has -1486.180000 for flesch_reading_ease and -3.
↪400000 for flesch_kincaid_grade
# so we can figure out why it is negative
suspect_1 = df_cupid.loc[df_cupid['flesch_reading_ease'] == -1486.18]
df_cupid.iloc[12194]
```

[27]:
```
age
31
status
single
sex
male
orientation
```

straight
body_type
athletic
diet
NaN
drinks
socially
drugs
sometimes
education graduated
from college/university
ethnicity
white
height
68.0
income
-1
job
other
last_online
2011-12-20-00-49
location
san francisco, california
offspring
NaN
pets
likes dogs
religion
agnosticism
sign taurus but
it doesn’t matter
smokes
NaN
speaks
english
all_essays
aaaaaaaaaajdkjdjdjsjjsjsjdjsndndjdjsnsnmsmsmddmmsmsmdmdmdmdmdnxnxnckcjffkf
education_group
Advanced
edu_gender_group
Advanced_male
clean_readability_text
aaaaaaaaaajdkjdjdjsjjsjsjdjsndndjdjsnsnmsmsmddmmsmsmdmdmdmdmdnxnxnckcjffkf
clean_word_tokens
[aaaaaaaaaajdkjdjdjsjjsjsjdjsndndjdjsnsnmsmsmddmmsmsmdmdmdmdmdnxnxnckcjffkf]
clean_word_string
aaaaaaaaaajdkjdjdjsjjsjsjdjsndndjdjsnsnmsmsmddmmsmsmdmdmdmdmdnxnxnckcjffkf

```
essay_length
1
syllable_count
20
lexicon_count
1
sentence_count
1
polysyllab_count
1
flesch_reading_ease
-1486.18
flesch_kincaid_grade
220.8
gunning_fog
40.4
smog_index
8.841846
dale_chall
19.4761
prop_polysyll
1.0
words_per_sentence
1.0
Name: 12194, dtype: object
```

Interpretation: The essay the user wrote, "aaaaaaaaaajdkjdjdjsjjsjsjdjsndndjdjsnsnmsmsmsmddmmsmsmdmdmdmdmdmdnxnxnckcjffkf," gets a very negative Flesch Reading Ease score because it's just a long string of random letters instead of real words. The formula treats it as one giant word with a huge number of syllables, and that makes the score drop far below zero. That's why the reading-ease score for their essay ends up being –1486.18.

[28]:
```
suspect_2 = df_cupid.loc[df_cupid['flesch_kincaid_grade'] < -3]
df_cupid.iloc[7611]
```

[28]:
```
age                                          22
status                                   single
sex                                        male
orientation                            straight
body_type                                skinny
diet                           mostly vegetarian
drinks                                 socially
drugs                                 sometimes
education          working on college/university
ethnicity                                 white
height                                     73.0
income                                       -1
job                                        other
```

```
last_online                              2012-06-20-19-36
location                               fairfax, california
offspring                                             NaN
pets                             likes dogs and likes cats
religion                       other and laughing about it
sign                                          sagittarius
smokes                                                 no
speaks                           english, spanish (okay)
all_essays                                           bwah
education_group                               Not Advanced
edu_gender_group                         Not Advanced_male
clean_readability_text                               bwah
clean_word_tokens                                  [bwah]
clean_word_string                                    bwah
essay_length                                            1
syllable_count                                          1
lexicon_count                                           1
sentence_count                                          1
polysyllab_count                                        0
flesch_reading_ease                                121.22
flesch_kincaid_grade                                 -3.4
gunning_fog                                           0.4
smog_index                                         3.1291
dale_chall                                        19.4761
prop_polysyll                                         0.0
words_per_sentence                                    1.0
Name: 7611, dtype: object
```

Interpretation: The essay this user wrote is just the word "bwah," which is not a real sentence and barely even a word. The Flesch-Kincaid Grade formula expects normal language with real words and sentences, so a short nonsense string like this confuses the calculation. Because the formula penalizes unusual or non-standard text, the result drops below –3. This means the writing is considered extremely simple and not meaningful according to the formula, which is why this entry appears in the suspect_2 group.

4.0.6 Display and inspect all essays with their Flesch Reading Ease and Flesch-Kincaid Grade scores.

[29]:
```
# Select essays with Flesch Reading Ease Grade below 0
neg_fre_ease = df_cupid[df_cupid['flesch_reading_ease'] < 0]
neg_fre_ease
```

[29]:
```
       age     status     sex orientation body_type              diet  \
122     24  available    male    bisexual    skinny               NaN
430     25     single    male    straight       fit               NaN
495     45     single  female    straight     curvy    mostly anything
720     28     single    male    straight    skinny    mostly anything
843     47     single    male    straight  athletic               NaN
```

```
…    …        …     …          …        …                 …
50361   58     single    male    straight   average    mostly anything
50645   28     single  female    straight       NaN  mostly vegetarian
50684   21     single  female    straight      thin  strictly anything
50741   66     single  female    straight     curvy           anything
50744   38     single    male    straight   average           anything

         drinks      drugs                           education  \
122    socially  sometimes    graduated from two-year college
430    socially      never         graduated from high school
495    socially      never     graduated from masters program
720    socially        NaN  graduated from college/university
843    socially  sometimes  dropped out of college/university
…         …        …                                   …
50361  socially      never      working on college/university
50645  socially      never  graduated from college/university
50684  socially  sometimes    graduated from two-year college
50741     often        NaN  graduated from college/university
50744    rarely      never  graduated from college/university

                    ethnicity  …  lexicon_count  sentence_count  \
122                       NaN  …             99               1
430    hispanic / latin, other  …             90               1
495                     white  …             90               1
720                     white  …            122               1
843                     white  …            471               2
…                       …  …           …             …
50361                   white  …            103               1
50645                   white  …            590               5
50684                   asian  …             14               1
50741                   white  …            100               1
50744                     NaN  …             78               1

      polysyllab_count flesch_reading_ease flesch_kincaid_grade gunning_fog  \
122                  7           -2.177273            38.157374   40.408081
430                  8           -7.655000            36.685556   38.222222
495                 14          -21.755000            38.652222   41.333333
720                 25          -64.004836            52.494918   56.996721
843                 52         -160.803869            94.193004   98.191507
…                 …                   …                   …         …
50361               19          -29.127476            42.910097   47.413592
50645               98          -53.887203            50.090000   52.827119
50684                4         -103.475000            31.170000   17.028571
50741               10          -18.181000            40.638000   43.600000
50744                6           -0.319615            32.681282   33.764103

      smog_index dale_chall prop_polysyll words_per_sentence
```

```
122    18.243606  10.301344      0.070707               99.0
430    19.287187  11.609389      0.088889               90.0
495    24.504239  14.065611      0.155556               90.0
720    31.692831  19.135815      0.204918              122.0
843    32.258505  20.949402      0.110403              235.5
…           …          …             …                   …
50361  28.030384  12.577824      0.184466              103.0
50645  28.420506  15.162995      0.166102              118.0
50684  14.554593  11.098043      0.285714               14.0
50741  21.194390  13.175600      0.100000              100.0
50744  17.122413  11.554018      0.076923               78.0

[196 rows x 39 columns]
```

[30]:
```
%%capture
# Take a random sample of 20 essays with negative Flesch Reading Ease
neg_fre_ease_sample20 = neg_fre_ease.sample(20, random_state=42)
neg_fre_ease_sample20
```

[31]:
```
%%capture
for i, essay in enumerate(neg_fre_ease_sample20['all_essays'], 1):
    print(f"\n--- Essay {i} ---\n")
    print(essay)
```

[32]:
```
%%capture
# Select essays with Flesch-Kincaid Grade below 0
neg_fk_grade = df_cupid[df_cupid['flesch_kincaid_grade'] < 0]
neg_fk_grade
```

[33]:
```
%%capture
# Take a random sample of 20 essays with negative Flesch-Kincaid Grade
neg_fk_grade_sample20 = neg_fk_grade.sample(20, random_state=42)
neg_fk_grade_sample20
```

[34]:
```
%%capture
for i, essay in enumerate(neg_fk_grade_sample20['all_essays'], 1):
    print(f"\n--- Essay {i} ---\n")
    print(essay)
```

Interpretation:
When we looked at the sample, we saw that many essays cannot be used with Flesch Reading Ease or Flesch–Kincaid Grade Level because both formulas require clear, complete sentences. Each measure depends on counting the number of sentences, the number of words per sentence, and the number of syllables in those words. Many essays in our sample do not use any punctuation, are written as one long stream of text, hyperlinks, or are simply lists of movies, music, or hobbies. When a text has no periods, question marks, or real sentence structure, the formulas cannot tell where one sentence ends and the next begins. This makes both the Flesch Reading Ease score and the Flesch–Kincaid Grade Level extremely inaccurate or even impossible to calculate. Some

essays are also far too short, which produces unstable scores because one long word or one short sentence can change the result too much. For this reason, our data cleaning keeps only essays that meet three simple conditions: they must have a valid Flesch score, they must contain at least two sentences or at least one sentence with proper punctuation, and they must include at least 100 words. These steps remove essays that are too short, unstructured, or missing punctuation, so the readability scores for both Flesch Reading Ease and Flesch–Kincaid Grade Level are meaningful and reliable for comparing writing across education levels and gender.

4.1 Part 2

The second part of our preprocessing was necessary because we noticed that some profiles produced negative readability values, which showed that the text was too short or not structured well enough for the formulas to work properly. Readability measures such as Flesch Reading Ease, Flesch–Kincaid Grade Level, and similar formulas can become unstable when the writing is extremely short, unstructured, or missing punctuation, so we removed profiles that produced negative scores. To make sure the readability analysis was based on meaningful writing samples, we kept only entries with positive readability values, at least two sentences (or one properly punctuated sentence), and a minimum of 100 words so the formulas had enough material to work with. These steps helped us filter out overly short or poorly structured writing that could distort our results. By applying these criteria, we created a cleaner and more reliable dataset for readability analysis while still respecting the original writing style preserved through our minimal preprocessing in the first stage.

4.1.1 Data Cleaning For Readability

[35]:
```
# Create a cleaned version of df_cupid by filtering out texts that are too␣
 ↪short
# or invalid for readability analysis.
clean_df_cupid = df_cupid[
    # Keep only essays with positive readability scores to exclude gibberish or␣
 ↪extremely unreadable text
    (df_cupid['flesch_reading_ease'] > 0)
    &
    (df_cupid['flesch_kincaid_grade'] > 0)
    &
    # Readability formulas (Flesch, FKGL, Fog, SMOG, Dale-Chall) rely on stable␣
 ↪estimates of sentence length. One-sentence texts produce
    # extremely unstable readability scores, especially in short social media␣
 ↪style writing.
    #This filter ensures the text has at least some structure.
    ((df_cupid['sentence_count'] >= 2) |
    # Some users write only one sentence but with proper punctuation (e.g. "Hi,␣
 ↪are you reading this?")
    # So those are still valid, grammatical sentences and should be kept.
    # Texts with NO punctuation (e.g. "hi imma Natali i hella love TeXt aS␣
 ↪$Data$ dog tennis idk ^_^ $_$ TAT QAQ")
    # are NOT real sentences and cannot be validly analyzed using readability␣
 ↪formulas.
```

```
    (df_cupid['clean_readability_text'].str.contains(r'[.!?]')))
    &
    # Because short texts produce unstable readability estimates (Klare 1974;␣
 ↪Gunning 1952),
    # we restrict our analysis to profiles containing at least 100 words.
    (df_cupid['lexicon_count'] >= 100)
]

clean_df_cupid.shape
```

[35]: (44044, 39)

Reasoning: After applying these filters, we still have 44044 profiles, which is still a large sample size. This gives us enough data to analyze readability patterns by gender and education level.

[36]:
```
sample20 = clean_df_cupid.sample(20, random_state=42)
```

[37]:
```
%%capture
for i, essay in enumerate(sample20['all_essays'], 1):
    print(f"\n--- Essay {i} ---\n")
    print(essay)
```

Interpretation: Based on a random sample of profiles, the remaining texts appear valid for readability analysis. They contain meaningful sentences and sustained writing rather than extremely short or fragmented text. This gives us confidence that applying readability measures to the cleaned dataset is appropriate.

4.1.2 Add readability columns to the cleaned okcupid data set

[38]:
```
# Component counts on cleaned text
clean_df_cupid['clean_syllable_count'] =␣
 ↪clean_df_cupid['clean_readability_text'].apply(textstat.syllable_count)
clean_df_cupid['clean_lexicon_count'] =␣
 ↪clean_df_cupid['clean_readability_text'].apply(
    lambda x: textstat.lexicon_count(x, removepunct=True)
)
clean_df_cupid['clean_sentence_count'] =␣
 ↪clean_df_cupid['clean_readability_text'].apply(textstat.sentence_count)
clean_df_cupid['clean_polysyllab_count'] =␣
 ↪clean_df_cupid['clean_readability_text'].apply(textstat.polysyllabcount)

# Readability scores on cleaned text
clean_df_cupid['clean_flesch_reading_ease'] =␣
 ↪clean_df_cupid['clean_readability_text'].apply(textstat.flesch_reading_ease)
clean_df_cupid['clean_flesch_kincaid_grade'] =␣
 ↪clean_df_cupid['clean_readability_text'].apply(textstat.flesch_kincaid_grade)
clean_df_cupid['clean_gunning_fog'] = clean_df_cupid['clean_readability_text'].
 ↪apply(textstat.gunning_fog)
```

```
clean_df_cupid['clean_smog_index'] = clean_df_cupid['clean_readability_text'].
 ↪apply(textstat.smog_index)
clean_df_cupid['clean_dale_chall'] = clean_df_cupid['clean_readability_text'].
 ↪apply(textstat.dale_chall_readability_score)
```

```
/tmp/ipykernel_1191794/2072709035.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  clean_df_cupid['clean_syllable_count'] =
clean_df_cupid['clean_readability_text'].apply(textstat.syllable_count)
/tmp/ipykernel_1191794/2072709035.py:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  clean_df_cupid['clean_lexicon_count'] =
clean_df_cupid['clean_readability_text'].apply(
/tmp/ipykernel_1191794/2072709035.py:6: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  clean_df_cupid['clean_sentence_count'] =
clean_df_cupid['clean_readability_text'].apply(textstat.sentence_count)
/tmp/ipykernel_1191794/2072709035.py:7: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  clean_df_cupid['clean_polysyllab_count'] =
clean_df_cupid['clean_readability_text'].apply(textstat.polysyllabcount)
/tmp/ipykernel_1191794/2072709035.py:10: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  clean_df_cupid['clean_flesch_reading_ease'] =
clean_df_cupid['clean_readability_text'].apply(textstat.flesch_reading_ease)
/tmp/ipykernel_1191794/2072709035.py:11: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead
```

```
See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  clean_df_cupid['clean_flesch_kincaid_grade'] =
clean_df_cupid['clean_readability_text'].apply(textstat.flesch_kincaid_grade)
/tmp/ipykernel_1191794/2072709035.py:12: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  clean_df_cupid['clean_gunning_fog'] =
clean_df_cupid['clean_readability_text'].apply(textstat.gunning_fog)
/tmp/ipykernel_1191794/2072709035.py:13: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  clean_df_cupid['clean_smog_index'] =
clean_df_cupid['clean_readability_text'].apply(textstat.smog_index)
/tmp/ipykernel_1191794/2072709035.py:14: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  clean_df_cupid['clean_dale_chall'] = clean_df_cupid['clean_readability_text'].
apply(textstat.dale_chall_readability_score)
```

[39]:
```
clean_df_cupid.columns
```

[39]:
```
Index(['age', 'status', 'sex', 'orientation', 'body_type', 'diet', 'drinks',
       'drugs', 'education', 'ethnicity', 'height', 'income', 'job',
       'last_online', 'location', 'offspring', 'pets', 'religion', 'sign',
       'smokes', 'speaks', 'all_essays', 'education_group', 'edu_gender_group',
       'clean_readability_text', 'clean_word_tokens', 'clean_word_string',
       'essay_length', 'syllable_count', 'lexicon_count', 'sentence_count',
       'polysyllab_count', 'flesch_reading_ease', 'flesch_kincaid_grade',
       'gunning_fog', 'smog_index', 'dale_chall', 'prop_polysyll',
       'words_per_sentence', 'clean_syllable_count', 'clean_lexicon_count',
       'clean_sentence_count', 'clean_polysyllab_count',
       'clean_flesch_reading_ease', 'clean_flesch_kincaid_grade',
       'clean_gunning_fog', 'clean_smog_index', 'clean_dale_chall'],
      dtype='object')
```

Reasoning: We compute these readability metrics for the entire dataset so we can analyze patterns in writing complexity across all profiles, not just the sample.

4.1.3 Calculate proportion measures for the cleaned data set

[40]:
```
# Proportion of polysyllabic words (using cleaned counts)
clean_df_cupid['clean_prop_polysyll'] = (
    clean_df_cupid['clean_polysyllab_count'] /␣
 ↪clean_df_cupid['clean_lexicon_count']
)

# Average number of words per sentence (using cleaned counts)
clean_df_cupid['clean_words_per_sentence'] = (
    clean_df_cupid['clean_lexicon_count'] /␣
 ↪clean_df_cupid['clean_sentence_count']
)
```

```
/tmp/ipykernel_1191794/900086139.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  clean_df_cupid['clean_prop_polysyll'] = (
/tmp/ipykernel_1191794/900086139.py:7: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  clean_df_cupid['clean_words_per_sentence'] = (
```

4.1.4 Describe readability variables for the cleaned data set

[41]:
```
clean_df_cupid.describe()[[
    'clean_syllable_count',
    'clean_lexicon_count',
    'clean_sentence_count',
    'clean_polysyllab_count',
    'clean_flesch_reading_ease',
    'clean_flesch_kincaid_grade',
    'clean_gunning_fog',
    'clean_smog_index',
    'clean_dale_chall',
    'clean_prop_polysyll',
    'clean_words_per_sentence'
]]
```

[41]:

	clean_syllable_count	clean_lexicon_count	clean_sentence_count	\
count	44044.000000	44044.000000	44044.000000	
mean	610.199505	415.453796	28.347970	

```
std              416.965452          281.659905             18.988453
min              122.000000          100.000000              2.000000
25%              335.000000          228.000000             16.000000
50%              513.000000          349.000000             24.000000
75%              773.000000          526.000000             36.000000
max            17501.000000        10486.000000            584.000000

       clean_polysyllab_count  clean_flesch_reading_ease  \
count            44044.000000               44044.000000
mean                43.728930                  66.667193
std                 33.146456                  10.122025
min                  0.000000                   0.636154
25%                 23.000000                  61.364152
50%                 36.000000                  67.793851
75%                 56.000000                  73.327581
max               1781.000000                 101.949963

       clean_flesch_kincaid_grade  clean_gunning_fog  clean_smog_index  \
count                44044.000000       44044.000000      44044.000000
mean                     7.821840           9.721974         10.284212
std                      2.553893           2.625055          1.627418
min                      0.305596           2.660000          3.129100
25%                      6.224719           8.072923          9.221296
50%                      7.401443           9.312199         10.094169
75%                      8.868621          10.858289         11.119845
max                     36.133830          38.275269         22.641843

       clean_dale_chall  clean_prop_polysyll  clean_words_per_sentence
count      44044.000000         44044.000000              44044.000000
mean           9.001506             0.105367                 15.542729
std            1.108294             0.030004                  5.658992
min            0.376960             0.000000                  4.911765
25%            8.256345             0.085366                 12.166667
50%            8.876221             0.103139                 14.428571
75%            9.600497             0.122644                 17.428571
max           19.496060             0.344961                 94.000000
```

Reasoning: Because readability metrics are scaled differently, we focus our main interpretation on Flesch Reading Ease and Flesch–Kincaid Grade. A lower Flesch Reading Ease score and a higher Flesch–Kincaid Grade score both indicate more complex writing, even though the numbers move in opposite directions. Looking at these two measures together allows us to interpret writing complexity more clearly. Other readability metrics, such as Gunning Fog and SMOG, are included as supporting checks and show similar overall patterns.

4.1.5 Group-by Readability: Education x Sex

Flesch Reading Ease by Gender × Education

[42]:
```
clean_df_cupid.groupby(['education_group', 'sex'])[
    'clean_flesch_reading_ease'
].describe()
```

[42]:

		count	mean	std	min	25%	50%	75%	max
education_group	sex								
Advanced	female	15119.0	66.265138	9.583561	1.942011	61.194563	67.320000	72.677805	93.619559
	male	19635.0	65.898168	10.079250	0.636154	60.724884	67.174653	72.498919	101.949963
Not Advanced	female	3461.0	69.266115	10.391925	1.807069	63.927501	70.348010	76.279970	97.482584
	male	5829.0	68.757364	10.902489	0.953429	63.176838	70.147993	75.959578	98.371374

Interpretation: Flesch Reading Ease measures how easy a text is to read, with higher scores indicating simpler, more conversational writing. Across all groups, scores are relatively similar and fall within a moderate range, suggesting that most dating profiles are written to be broadly accessible. This is consistent with the idea that dating profiles are meant to be readable and approachable rather than formal or academic.

That said, some small patterns appear by education level. Profiles in the Advanced education group tend to have slightly lower Reading Ease scores than those in the Not Advanced group for both men and women, indicating marginally more complex writing. Gender differences within the same education group are minimal, with male and female users showing very similar average scores

Flesch–Kincaid Grade by Gender × Education

[43]:
```
clean_df_cupid.groupby(['education_group', 'sex'])[
    'clean_flesch_kincaid_grade'
].describe()
```

[43]:

		count	mean	std	min	25%	50%	75%	max
education_group	sex								
Advanced	female	15119.0	7.873760	2.370335	1.918354	6.349915	7.487196	8.919076	29.508910
	male	19635.0	7.922448	2.499720	0.305596	6.360042	7.507879	8.959994	32.976293
Not Advanced	female	3461.0	7.353122	2.721131	1.532047	5.677978	6.915895	8.355813	32.430596
	male	5829.0	7.626577	3.009862	1.300000	5.785742	7.084202	8.651366	36.133830

Interpretation: Flesch–Kincaid Grade Level estimates the U.S. school grade level needed to understand a text, with higher values indicating more complex sentence structure and vocabulary. Across gender and education groups, scores cluster around grades 7–8, again suggesting that dating profiles tend to use relatively simple, conversational language.

However, profiles in the Advanced education group generally have slightly higher Flesch–Kincaid Grade levels than those in the Not Advanced group. This pattern holds for both men and women, indicating that people with higher educational attainment write in somewhat more complex ways on average. As with Reading Ease, gender differences within the same education group are small compared to differences between education levels.

[44]:
```
clean_df_cupid.groupby(['education_group', 'sex'])[[
    'clean_gunning_fog',
    'clean_smog_index',
    'clean_dale_chall',
    'clean_prop_polysyll',
    'clean_words_per_sentence'
]].describe()
```

[44]:
```
                        clean_gunning_fog                                  \
                                    count      mean       std       min
education_group sex
Advanced        female            15119.0  9.762467  2.442163  3.785982
                male              19635.0  9.849551  2.566724  3.040000
Not Advanced    female             3461.0  9.178869  2.785669  3.299145
                male               5829.0  9.509667  3.086011  2.660000

                                                                    \
                                  25%       50%       75%        max
education_group sex
Advanced        female  8.192582  9.381085  10.899075  31.235772
                male    8.240226  9.447001  10.964877  36.017424
Not Advanced    female  7.472281  8.744756  10.242731  34.818462
                male    7.593706  8.964572  10.619165  38.275269

                        clean_smog_index             … clean_prop_polysyll  \
                                    count       mean  …                 75%
education_group sex                                  …
Advanced        female          15119.0  10.358400  …            0.122951
                male            19635.0  10.404577  …            0.125688
Not Advanced    female           3461.0   9.833012  …            0.114035
                male             5829.0   9.954239  …            0.114754

                                    clean_words_per_sentence             \
                                 max                    count       mean
education_group sex
Advanced        female  0.285714                  15119.0  15.525990
                male    0.344961                  19635.0  15.515937
```

```
Not Advanced    female  0.284483                  3461.0  15.115160
                male    0.280000                  5829.0  15.930262

                                                                    \
                              std       min        25%        50%        75%
education_group sex
Advanced        female  5.112209  4.911765  12.311400  14.538462  17.514333
                male    5.440316  5.080645  12.227273  14.458333  17.410256
Not Advanced    female  6.244284  5.263158  11.600000  13.892857  16.750000
                male    7.148217  5.880000  11.857143  14.315789  17.629630


                               max
education_group sex
Advanced        female  68.333333
                male    83.000000
Not Advanced    female  83.200000
                male    94.000000

[4 rows x 40 columns]
```

Reasoning: In addition to these two measures, we computed several other standard readability metrics, including Gunning Fog, SMOG, and Dale–Chall, as well as sentence length and the proportion of polysyllabic words. These metrics produce similar overall patterns by education level and gender, suggesting that the results from Flesch Reading Ease and Flesch–Kincaid Grade are not driven by a single formula.

4.1.6 Validation for Flesch Reading Ease and Flesch–Kincaid Grade scores by education and gender

4.1.7 Min/Max Flesch Ease Scores

[45]:
```
%%capture
# 1. Get min and max Flesch Reading Ease values per group
flesch_minmax = (
    clean_df_cupid
    .groupby(['education_group', 'sex'])['clean_flesch_reading_ease']
    .agg(['min', 'max'])
    .reset_index()
)

print("=== Min/Max Flesch Reading Ease by Education Group and Gender ===")
print(flesch_minmax)


# 2. Loop through each group and extract the corresponding essays
print("\n\n=== PRINTING ESSAYS FOR MIN/MAX READABILITY ===")
```

```
for _, row in flesch_minmax.iterrows():
    edu = row['education_group']
    gender = row['sex']
    min_val = row['min']
    max_val = row['max']

    print("\n" + "#" * 80)
    print(f"Education Group: {edu} | Gender: {gender}")
    print(f"Minimum Flesch Reading Ease: {min_val}")
    print(f"Maximum Flesch Reading Ease: {max_val}")
    print("#" * 80)

    # ---- Essays with MIN readability ----
    min_essays = clean_df_cupid[
        (clean_df_cupid['education_group'] == edu) &
        (clean_df_cupid['sex'] == gender) &
        (clean_df_cupid['clean_flesch_reading_ease'] == min_val)
    ]['all_essays']

    print("\n--- ESSAY(S) WITH MIN READABILITY ---")
    for idx, text in min_essays.items():
        print(f"\n[ID {idx}]")
        print(text)

    # ---- Essays with MAX readability ----
    max_essays = clean_df_cupid[
        (clean_df_cupid['education_group'] == edu) &
        (clean_df_cupid['sex'] == gender) &
        (clean_df_cupid['clean_flesch_reading_ease'] == max_val)
    ]['all_essays']

    print("\n--- ESSAY(S) WITH MAX READABILITY ---")
    for idx, text in max_essays.items():
        print(f"\n[ID {idx}]")
        print(text)

    print("\n" + "#" * 80 + "\n")
```

Interpretation : The minimum and maximum Flesch Reading Ease scores in each education and gender group help us check whether the readability results behave in a realistic way. The essays with the lowest Flesch scores are long, dense, and packed with complex ideas, unusual vocabulary, or very long sentences. This makes them harder to read, which is exactly why their Flesch scores fall close to zero. The essays with the highest Flesch scores, on the other hand, use short sentences, simple wording, and a casual conversational tone, which makes them much easier to read and produces scores around 90–100. Seeing this clear difference between the hardest and easiest essays shows that the Flesch measure is responding correctly to real writing features. Also, the fact that both low-score essays and high-score essays appear in every education × gender group suggests that the cleaning process worked: we kept only structurally valid text, but we did not accidentally

remove natural variation in how people write. Overall, these min/max results confirm that the Flesch Reading Ease scores are meaningful and can be used to compare writing style differences across education levels and gender.

4.1.8 Min/Max Flesch–Kincaid Grade

[46]:
```
%%capture
# 1. Get min/max Flesch-Kincaid Grade per group
fk_minmax = (
    clean_df_cupid
    .groupby(['education_group', 'sex'])['clean_flesch_kincaid_grade']
    .agg(['min', 'max'])
    .reset_index()
)

print("=== Min/Max Flesch-Kincaid Grade by Education Group and Gender ===")
print(fk_minmax)


print("\n\n=== PRINTING ESSAYS FOR MIN/MAX FLESCH-KINCAID GRADE ===")

for _, row in fk_minmax.iterrows():
    edu = row['education_group']
    gender = row['sex']
    min_val = row['min']
    max_val = row['max']

    print("\n" + "#" * 80)
    print(f"Education Group: {edu} | Gender: {gender}")
    print(f"Minimum FK Grade: {min_val}")
    print(f"Maximum FK Grade: {max_val}")
    print("#" * 80)

    # Essay(s) with minimum FK Grade
    min_essays = clean_df_cupid[
        (clean_df_cupid['education_group'] == edu) &
        (clean_df_cupid['sex'] == gender) &
        (clean_df_cupid['clean_flesch_kincaid_grade'] == min_val)
    ]['all_essays']

    print("\n--- ESSAY(S) WITH MIN FLESCH-KINCAID GRADE ---")
    for idx, text in min_essays.items():
        print(f"\n[ID {idx}]")
        print(text)

    # Essay(s) with maximum FK Grade
    max_essays = clean_df_cupid[
```

```
        (clean_df_cupid['education_group'] == edu) &
        (clean_df_cupid['sex'] == gender) &
        (clean_df_cupid['clean_flesch_kincaid_grade'] == max_val)
    ]['all_essays']

    print("\n--- ESSAY(S) WITH MAX FLESCH-KINCAID GRADE ---")
    for idx, text in max_essays.items():
        print(f"\n[ID {idx}]")
        print(text)

    print("\n" + "#" * 80 + "\n")
```

Interpretation: The minimum and maximum Flesch–Kincaid Grade Level scores in each education and gender group help us check whether the grade-level results behave in a meaningful way. The essays with the lowest FK Grade scores use short sentences, simple words, and a casual writing style, which results in grade levels around 0–2. These essays read more like everyday conversation, which explains their low grade scores. In contrast, the essays with the highest FK Grade scores are much more complex: they contain very long sentences, advanced vocabulary, abstract ideas, and dense writing with few breaks. These features naturally push the FK Grade Level into the high 20s and 30s, which represents college-level or even graduate-level writing. This pattern appears across all education and gender groups, showing that the FK Grade formula is responding correctly to real differences in sentence length and word complexity. The presence of both very low and very high grade levels also shows that our data-cleaning steps worked: we removed only texts that lacked punctuation or were too short, but we kept the full range of natural writing styles. Overall, these min/max results confirm that the Flesch–Kincaid Grade Level scores are valid and useful for comparing writing complexity across education groups and genders.

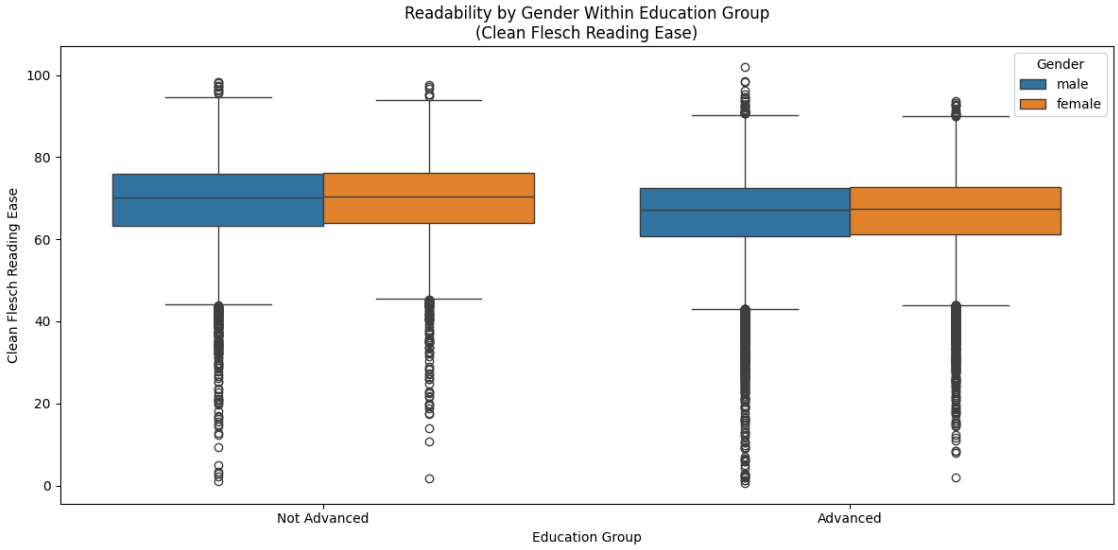
Reasoning: We chose Flesch Reading Ease and Flesch–Kincaid Grade Level because these two scores work best for short, casual writing like dating profiles. They look at things like how long the sentences are and how difficult the words are, which is important for understanding how people describe themselves and how complex their writing is. This fits our hypothesis because we think people with different education levels may write in different ways, such as using longer sentences or more advanced vocabulary. Flesch and FK Grade can capture these differences even when the profile only has a few sentences. Other readability formulas do not fit our dataset well. For example, the SMOG index needs around thirty sentences to be accurate, and most dating profiles are much shorter than that. Formulas like Gunning Fog or Dale–Chall can also give confusing results when the writing is short or very informal. For these reasons, Flesch Reading Ease and Flesch–Kincaid Grade Level are the best choice for understanding writing style differences across education levels and gender in our dataset of short and personal profile essays.

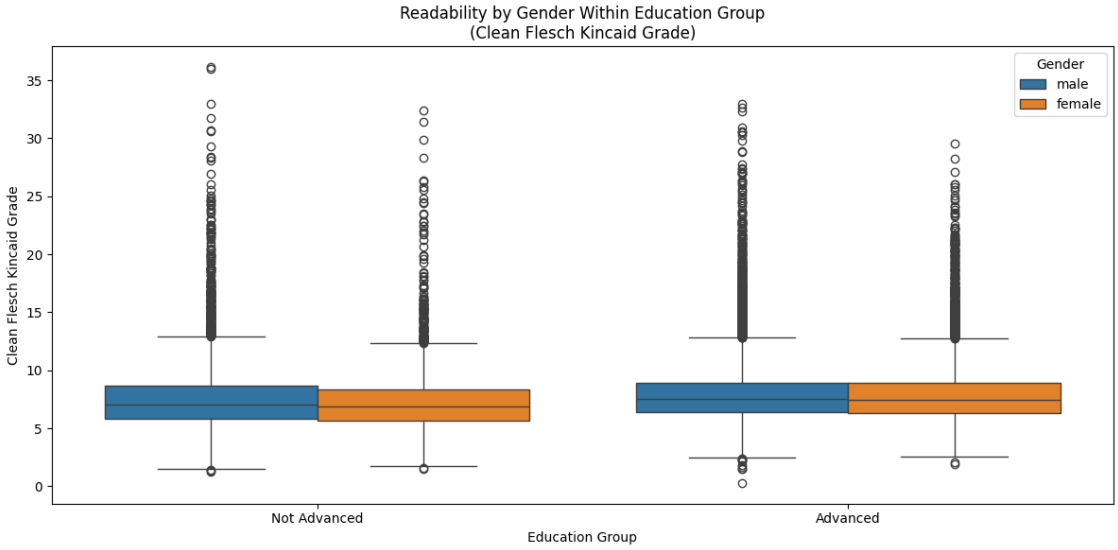
4.1.9 Visualizing Readability Scores

[47]:
```
%%capture
df_cupid['all_essays'].sample(20).tolist()
```

Reasoning: We sample 20 essays to get a quick sense of what the texts look like, so we can better understand the readability results we computed earlier

[48]:
```
# All readability variables we want to visualize
readability_vars = [
    'clean_flesch_reading_ease',
    'clean_flesch_kincaid_grade',
]

# Loop through and create boxplots for each
for var in readability_vars:
    plt.figure(figsize=(12, 6))
    sns.boxplot(
        data=clean_df_cupid,
        x='education_group',
        y=var,
        hue='sex'
    )

    plt.title(f"Readability by Gender Within Education Group\n({var.
 ↪replace('_', ' ').title()})")
    plt.xlabel("Education Group")
    plt.ylabel(var.replace('_', ' ').title())
    plt.legend(title="Gender")
    plt.tight_layout()
    plt.show()
```





Interpretation: Looking at the boxplots, the median readability scores are very close across all groups, and the interquartile ranges overlap a lot. This suggests that neither gender nor education level has a strong effect on the typical readability of dating profiles. Men and women within the same education group tend to write at similar readability levels, and the advanced and not advanced education groups also look quite similar overall.

What stands out much more is the amount of variation within each group. Each boxplot shows a wide spread and many outliers, especially on the lower end of readability. Some people write very simple, easy-to-read profiles, while others write very long or complicated ones, which creates many extreme outliers. This shows that individual writing style varies a lot from person to person, and this variation is much larger than any difference between gender or education.

5 Results

Overall, our results suggest that gender and education level have only a small influence on writing style in dating profiles, especially when compared to the large amount of variation across individuals. When examining readability measures such as Flesch Reading Ease and Flesch–Kincaid Grade Level, average scores were very similar across gender and education groups. Men and women within the same education level showed nearly identical readability scores, indicating that gender alone does not strongly shape how people write their profiles.

Education level shows a slightly clearer pattern, but the differences are still modest. Users in the Advanced education group tend to write marginally more complex profiles, reflected in slightly lower Reading Ease scores and slightly higher Flesch–Kincaid Grade Levels. This pattern is consistent across genders, suggesting that education is more strongly associated with writing style than gender, though the overall effect remains small. These findings partially align with our hypothesis that people with higher educational attainment may share similar writing habits, but the similarity appears subtle rather than pronounced.

It is also important to note that our sample size is very large. With so many profiles in each group,

there is substantial individual-level noise in writing style, which can make small between-group differences harder to visually detect in boxplots. As a result, the large within-group variation may obscure more subtle effects that would require formal statistical testing to identify.

What stands out most across all analyses is the large within-group variation. Each gender–education group contains profiles that are extremely short and simple as well as profiles that are long and complex, leading to many outliers in the readability distributions. This suggests that individual writing style and self-expression choices matter more than demographic characteristics when people describe themselves and signal preferences to potential partners.

The frequency distributions and word clouds support this interpretation. Across all groups, users rely on very similar high-frequency words related to self-description, interests, relationships, and everyday life. Small differences appear, with advanced users using slightly more abstract or identity-related terms and non-advanced users using more concrete, everyday words. Gender-based differences in word choice are also minor, with women using slightly more relationship-oriented words and men using slightly more activity-oriented terms.

5.1 Discussion and Conclusion

- *What has your analysis shown, and what new questions have arisen?*

Our analysis shows that gender and education level are only weakly related to how people write dating profiles, especially in terms of readability and overall writing complexity. While users with higher education tend to write slightly longer and more complex profiles on average, the differences are small compared to the large variation within each group. Many profiles are very short, casual, or even unconventional, which suggests that individual choices and styles matter more than demographic characteristics. As a result, readability does not appear to be a strong or consistent way that users signal education or compatibility on dating profiles. These findings raise new questions about how seriously users treat dating apps and how that affects the way they write. Not everyone approaches a dating profile with the same level of investment. Some users may be actively looking for a long-term partner, while others may be browsing casually or not taking the platform very seriously. This could explain why many profiles use informal language, incomplete sentences, or even gibberish, rather than carefully written text. If users are not committed to presenting themselves in a detailed or polished way, readability measures may not capture meaningful differences across groups. Another important consideration is methodological. Much of the existing literature using readability measures focuses on formal writing contexts, such as academic publications, corporate reports, or medical materials. In contrast, dating profiles are overwhelmingly informal, conversational, and expressive. This mismatch suggests that traditional readability metrics may be limited in this setting and may miss other important aspects of how people signal identity or preferences. Future analyses could focus on different features of dating profiles, such as tone, humor, emoji use, narrative structure, or specific self-presentation strategies, rather than relying mainly on sentence complexity or grade-level measures.

- *Did you find what you expected to find? Why or why not?*

Generally speaking, our analysis shows that gender and education level do not strongly influence writing style on dating profiles, at least when writing style is measured using readability metrics and basic text features. Across our exploratory data analysis, the average readability scores are very similar across gender and education groups. What stands out more than group differences is the large amount of variation within each group. Some users write extremely short and simple

profiles, while others write much longer and more detailed ones. This suggests that individual writing choices matter more than demographic characteristics when people describe themselves on dating profiles.

These results only weakly support our original hypothesis. While users with higher education levels tend to write slightly longer profiles and show slightly higher Flesch–Kincaid Grade Level scores on average, these differences are small. We do not observe strong clustering of writing style within education groups, and the overlap between groups is large. As a result, our findings do not provide strong evidence that readability is an important signal of assortative mating on dating profiles. This may be partly because dating profiles involve very informal writing. Much of the existing literature that uses readability measures focuses on formal contexts such as academic publishing, corporate reports, or medical materials, where readability standards are clearer and more relevant. Dating profiles, on the other hand, are casual, conversational, and often uneven in structure.

Our exploratory analysis of word frequency and word clouds further reinforces this point. Across all groups, users consistently focused on similar topics, including everyday interests and self-descriptions. Common words such as "life," "music," "movie," "work," "love," and "friend" dominated the text regardless of gender or education level. For most people, a dating profile should always only contain brief descriptions of hobbies, interests, and personality rather than deeper reflections on values or life experiences. More personal or complex aspects of identity are likely revealed later through one-on-one conversations rather than through the initial profile text, which helps explain why we see strong similarities across groups in both readability and word usage.

In conclusion, we did not find the strong relationships we initially expected between gender, education level, and writing style. This does not mean that assortative mating is unimportant in dating, but rather that readability measures alone may not be the best way to capture how people signal background or compatibility in dating profiles. We expected that people with higher education, whether they are male or female, will write more complex sentences, while people with lower education write in a simpler way. These expected findings would present themselves in the 2 by 2 table, where the advanced education groups appear similar to each other, and the not advanced education groups look similar to each other as well. This would represent assortative mating because it suggests that people who communicate in similar ways are likely to match up. Our exploratory analysis did not fully yield the expected results, and future work might benefit from using methods more suited to informal writing, such as analyzing tone, level of self-disclosure, or indicators of effort and seriousness. Future research may also yield more results if data on which users match and/or message each other is obtainable and used. The study "Matching and Sorting in Online Dating", conducted by Günter J. Hitsch, Ali Hortaçsu, and Dan Ariely, was able to incorporate this data, but unfortunately the dataset from OKCupid did not contain information on matches. The study involved creating predictions based on human sorting patterns, and then comparing these predictions to actual matches made on the site (Hitsch, Hortaçsu, and Ariely, 2010). Additionally, there are countless websites and apps for online dating, some of which are targeted to specific groups, so there could be different results depending on which one is studied. Overall, the limitations of our research led to minimal results, but this does not discredit the phenomenon of assortative mating, or the idea that people of similar educational backgrounds communicate alike.

- *What are the limitations of this study?*

1. We do not have information on how seriously users view the dating app. Some people may treat it casually, while others may be actively looking for a partner, which likely affects how much effort they put into writing their profiles.

2. We did not conduct interviews or surveys, so we cannot ask users about their intentions, how they think about the app, or what they are trying to signal through their writing.
3. We do not observe actual matching or relationship outcomes, so we cannot directly test assortative mating or whether similarities in writing style are related to partner selection.
4. We cannot validate whether the information users provide about themselves is accurate, exaggerated, or incomplete.
5. Readability measures were originally designed for more formal writing, such as academic papers, medical materials, or corporate reports, so they may not fully capture differences in informal and conversational writing like dating profiles.
6. Many users write in a very informal way, using short sentences, slang, jokes, or filler text, which makes readability scores less sensitive to subtle differences by gender or education.
7. Users with higher education levels may write shorter profiles due to time constraints from work or other responsibilities, rather than differences in writing ability or style.
8. The sample may be biased if people with higher education levels are less likely to use dating apps or to write long profiles on them.
9. Writing style in dating profiles may reflect personality or mood more than education or gender, which makes it harder to isolate demographic effects using text alone.

- *What might future researchers do to overcome the limitations of this study?*

1. Collect information on how seriously users view dating apps, such as whether they are looking for long term relationships, casual dating, or just exploring. This could help explain why some profiles are detailed while others are very short or informal.
2. Combine text analysis with surveys or interviews asking users what they are trying to communicate in their profiles. This would give more context to the writing styles seen in the data.
3. Include matching or messaging data to study assortative mating more directly, rather than relying only on profile text.
4. Use methods better suited for informal writing, such as topic modeling, sentiment analysis, or theme based analysis, which may capture differences that readability scores do not.

- *How reliable and valid are your measures?*

Our measures are reliable because both readability formulas use fixed rules for counting words, sentences, and syllables, and they give consistent results when the text is structured in a similar way. Our preprocessing helped this by removing negative scores and very short or unstructured writing that would make the formulas behave unpredictably. To check validity, we did a simple manual test. We looked at the essays with the highest and lowest scores in each gender and education group and read them to see whether the scores matched what a person would expect. For Flesch Reading Ease, the essays with higher scores were easier to read, and the ones with lower scores were harder. For Flesch–Kincaid, the essays with higher scores required a higher grade level, and the ones with lower scores were simpler. This showed that both formulas were working in the direction they are supposed to and were capturing real differences in text complexity. The Flesch formulas were developed in the 1940s, and are still utilized in modern studies involving readability, such as Hengel's study on readability of male and female academic writing. At the same time, these formulas only measure basic structure and do not capture the tone of the writing, so their validity is limited to text complexity alone. George R. Klare's work, "Assessing Readability" examines the many formulas that have been developed throughout the years, claiming that there are "almost too many" to choose from. He concludes that with the abstract concept of readability, chosen formulas are "likely to be either too simple to be accurate or too complex to be convenient", and

it is important to be mindful of the context and purpose in which these formulas are being used (Klare, 1975). Future research on our topic would benefit from a dedicated period of research on text analysis formulas in order to find ones that are the most suitable for the informal and personal writing found on online dating profiles.

5.2 Works Cited

1. Abbott, L. J., S. Parker, and T. J. Presley. "Female Board Presence and the Likelihood of Financial Restatement." Accounting Horizons, vol. 26, no. 4, 2012, pp. 607–629. Aldridge, Michael D. "Writing and Designing Readable Patient Education Materials." Nephrology Nursing Journal, vol. 31, no. 4, July–Aug. 2004, pp. 373–377.

2. Anderson, Monica. "1. Americans' Personal Experiences with Online Dating." Pew Research Center, Pew Research Center, 6 Feb. 2020, www.pewresearch.org/internet/2020/02/06/americans-personal-experiences-with-online-dating/.

3. Buss, David M. "Human Mate Selection: Opposites Are Sometimes Said to Attract, but in Fact We Are Likely to Marry Someone Who Is Similar to Us in Almost Every Variable." American Scientist, vol. 73, no. 1, 1985, pp. 47–51. JSTOR, http://www.jstor.org/stable/27853061 .

4. Hengel, Erin. Publishing While Female: Are Women Held to Higher Standards? Evidence from Peer Review. Cambridge Working Papers in Economics, no. 1753, Faculty of Economics, University of Cambridge, 2017.

5. Hitsch, G. J., Hortaçsu, A., & Ariely, D. (2010). Matching and Sorting in Online Dating. The American Economic Review, 100(1), 130–163. http://www.jstor.org/stable/27804924

6. Kim, Albert, and Adriana Escobedo-Land. "OkCupid Data for Introductory Statistics and Data Science Courses." Journal of Statistics Education, vol. 23, no. 2, 2015, https://doi.org/10.1080/10691898.2015.11889737. Klare, George R. "Assessing Readability." Reading Research Quarterly, vol. 10, no. 1, 1974, pp. 62–102. JSTOR, https://doi.org/10.2307/747086.

7. Stephen Whyte, Benno Torgler, Things change with age: Educational assortment in online dating, Personality and Individual Differences, Volume 109, 2017, Pages 5-11, ISSN 0191-8869, https://doi.org/10.1016/j.paid.2016.12.031.