**Project 2: Equality in Hollywood**

**Word Cloud Analysis and Summary**

<https://python-graph-gallery.com/wordcloud/>

**Analysis**:

The Bechdel Test has three simple guidelines. To pass, it must be a movie or piece of fiction that:

1. Has a conversation
2. Between two women
3. That is not about a man.

The third guideline prompted us to ask . . .

“If a conversation between two women was not about a man, what was it about?”

To answer this, we needed several points of data that would first assess a movie to determine whether it had passed the Bechdel test, and if so, it had to provide information on what the conversation was about. Our dataset resource, [www.bechdeltest.com](http://www.bechdeltest.com) had no structured dataset to download that would address this issue. It did, however, contain several viewer comments where, once webscraped, gave us an interesting set of contextual datapoints to consider.

First, movies with more than 20 reviewer comments were analyzed. Having at least this many comments or more created a natural frequency in viewer sentiment on the accuracy of the movies that were listed as passing the Bechdel test. Common topics or discussions were cited repeatedly as viewers agreed that our datapoints were accurate.

Next, two years of data for movies that passed were reviewed. This was done to analyze 2009 and 2019 review comments to determine evolution of topics. The process for ETL of the data can be summarized as follows:

1. Extract the data from bechdeltest.com via webscrape.
2. Transform the data using Python in Jupyter Lab.
   1. import requests
   2. import urllib.request
   3. import time
   4. from bs4 import BeautifulSoup
   5. from wordcloud, import WordCloud
   6. import matplotlib.pyplot as plt
   7. Clean the data by using the tokenization method, and extracting articles, punctuation, and non-descriptive words from the tokens.
   8. Assign the wordcloud comments to several variables for easier code handling.
   9. Generate several test word clouds and repeat to reduce remaining anomaly words or phrases.
   10. Use the frequency distribution method to get a count of the recurring topics.
3. Load the data by generating wordclouds and a line plot, then exporting them to the file folder.

**Summary**: The approach taken to analyze what conversations not about men using webscrape was effective. In a line plot, the top 25 reviewed responses were listed to reveal that movies passing the Bechdel Test have women that converse about meatloaf, life, death, and their childhoods. Sometimes, the conversations were about saving the world or their careers. The most overwhelming evidence from the sample set, however, proved that nail polish and perfume led the topic list!