Captions Exploration

- Each dot represents a word or hashtag used in Instagram posts from our original dataset.
- Words that are further to the right are used more frequently by posts that have a negative sentiment. Words that are further to the top are used more frequently by posts that have a positive sentiment.
- Words that are frequently used by both posts that positive-sentiment and negative-sentiment posts are on the top right corner. Conversely, words that are infrequenlty used by both kinds of posts would be on the lower left hand corner.
- Terms are colored by their association so that words that are more associated with negative sentiment Instagram posts are red while words that are more associated with positive sentiment Instagram posts are blue.

Characteristic Terms:

```
- covid_posts_sentiment_key_words.html
- emoji_covid_posts.html
```

- How are the terms displayed chosen?
  - The terms displayed are called "characteristic terms" and they aim to represent words that have a high **recall** and **precision** in our posts.
- Terms with high **recall** have a high frequency within our category. Thus, they aim to maximize the probability of a term appering, given a category.

$$P(term|category)$$

- However, many stop words, such as "a, the, and" also appear frequently in each category and they aren't particularily important. So, to address this issue, we also want the characteristic terms to have high **precision**. Terms with high precision tend to appear more often in our category than in other categories. -

$$P(category|term)$$

- The final Fscore for a category j and word *i* is:

$$Fscore_{i,j} = HarmonicMean(P(term|category), P(category|term))$$

  The harmonic mean is computed as follows:

$$H_\beta(i,j) = (1+\beta^2) * \frac{precision(i,j) * frequency(i,j)}{\beta^2 * precision(i,j) + freq(i,j)}$$

  β is a scaling parameter between 0 and 1.
- We are using positive sentiment as our positive category and negative sentiment as our negative category. So to compute the final score for a single word i across both our categories, we compute the Fscore for the positive-sentiment category which we will call $FScore_{i,pos}$, and the Fscore for the negative-sentiment category, which we will call $FScore_{i,neg}$. The final score Si will be calculated as follows:

$$S_i = 2 * (-0.5 + \begin{cases} FScore_{i,pos} & \text{if } FScore_{i,pos} > FScore_{i,neg} \\ 1 - FScore_{i,neg} & \text{if } FScore_{i,pos} < FScore_{i,neg} \\ 0 \end{cases})$$

- The range of S_i is between -1 and 1.
- That means that words with negative values are more associated with negative-sentiment posts while words with a high positive scores are associated with positive-sentiment posts.

Useful Links:

Analyzing Yelp Dataset with Scattertext spaCy - Towards Data Science (https://towardsdatascience.com/analyzing-yelp-dataset-with-scattertext-spacy-82ea8bb7a60e)

GitHub - JasonKessler/scattertext: Beautiful visualizations of how language differs among document types. (https://github.com/JasonKessler/scattertext#understanding-scaled-f-score)

Categories Visualization

```
- covid_posts_empath_topics.html
```

The categories are extracted with Empath: a text analysis tool that constructs categories based on a set of predefined seed words. It covers a pre-validated set of 200 emotional and topical categories. It uses a combination of deep learning and microtask crowdsources. Empath can generate categories that are very similar to categories that have been hand-tuned and psychometrically validated by humans.

We can analyze this chart in the same way as before. Each dot represents a category. Dots that are further to the right represent categories that are more frequently used on negative-sentiment posts while dots further to the top represent categories that are more often seen on positive-sentiment posts. The scores displayed are the same Fscores as before, but instead of calculating the precision and recall of specific terms, we are calculating the precision and recall of topics in our posts. So, more negative scores correspond to categories that are associated more strongly with negative-sentiment posts while larger positive socres correspond to categories that are associated more strongly with positive-sentiment posts.

Useful link: https://hci.stanford.edu/publications/2016/ethan/empath-chi-2016.pdf (https://hci.stanford.edu/publications/2016/ethan/empath-chi-2016.pdf)

Citation

Jason S. Kessler. Scattertext: a Browser-Based Tool for Visualizing how Corpora Differ. ACL System Demonstrations. 2017.