**Puri and Fischer HW5**

Representations of leadership ("images") are prominent in discourse about leadership and leadership education. Maxwell and Greenhalgh's 2019 study of images and essays of leadership submitted by Wharton undergrads made those images more real and had elements of natural language processing. However, the technology has advanced since then.

One surprising element of their paper was convergance: common images arose across individuals and class years. Martin Luther King Jr., is one popular and perennial example. What’s less clear is if the essays about these common images were semantically similar.

In this assignment, we ask if NLP can help us understand whether students who submit the same photos have converged on the same ideas in their essays. We hypothesize that while the essays of the same photos will be divergent because we suppose that what a student has to say about an image is less about the image, and more about their own individual experience.

Stated plainly, here are our hypotheses:

1. The same image can have different semantic meanings to different people.
2. Elements of identity (e.g. Gender) do not fully account for these semantic differences

And here is our data: ~300 images and essays responding to the topic "what is your image of leadership" from the 2023 Wharton freshman class

In moving through the data we learned a lot. First we learned the importance of pre-processing the data. We believe garbage in garbage out gold in gold out, so we worked to remove erroneous citations and hyperlinks that weren’t really about describing the students idea of leadership.

Next we moved onto text embeddings. We tried out three different methods of text embeddings, both to learn about the methods and also to see what they could tell us about the data. We had some priors that bert would be the best embedding because it’s the most recent technology, but LDA was built for the purpose of identifying topics, and modern doc2vec implementations use some attention-like technology so we were open to the results.

Since our study was unsupervised learning (we didn’t have class labels to start which said these essays are of \_\_ type and these are not) we thought we would diversify both the way we clustered the data and visualized while also checking the actual essays themselves to ground us.

We began by exploring the univariate distributions of the embeddings to see if they were fit for gaussian mixture models. Because BERT produced a very large feature space, we used ICA to reduce it. We picked ICA over PCA because we felt as though ICA fit the philosophy of the problem: we believe that there are different “signals” driving the essays: some essays have the directive leadership signal, and some might have the connection-signal and some may have both for example (see fig 1).

As it turned out, LDA and ICA-Bert embeddings were normally distributed and seemed like great candidates for GMM. Sadly, we thought of this late in the game, so we didn’t get to do all the test we would have liked on this model. However, we saved the results for comparisons at the end.

We pre-registered that we’d like to use tree based methods, and parallel line plots, to look for clusters. We thought these would be useful because they are so visually interpretable. Visual interpretability seemed to go hand in hand with the fact that this was unsupervised learning: we needed something to hold onto to assess the goodness of clustering.

It turned out that doing tree based clustering first was a good instinct. It gave us a set of cluster labels that we could apply to the parallel line plots which helped us color and interpret them as we were exploring our data. LDA is a good example of this. At first we didn’t see anything in the uncolored parallel line plot that would suggest LDA was doing a good job of clumping the data. But when we applied the tree based clustering of the LDA embeddings on the LDA parallel lines plot we saw it did a really good job noting clusters. That was great. (see fig 3)

The LDA embedding also gave us the most clear understanding of our data: some essays were about connection and some were about direction. While an old technology, having this easy handle on the data proved really nice in making interpretations.

On the other side of the spectrum, we were really impressed with the sophistication of the sentenceBERT. It seemed to us that it picked up on leadership’s scale and charity grouping together essays which had an international and global-transformational character to them.

Doc2Vec initially seemed like a great clustering approach. It seemed to clearly separate out essays into top-down directing, caring-connecting, and guiding-visioning. It’s not clear how much that would hold up if we had all the hand labeled data however. (see fig 2)

With these labels in mind we switched to UMAP. Now that we had a sense of what these categories were, what could we tell about the quality of the clustering by mapping the data into space and coloring it? We chose UMAP in part because it was state of the art, in part because it ran fast, but mostly because of the parameters and the underlying algorithm. When we started to play around with it we got over being intimidated by it and started to understand how it could help us explore the structure of our data. We started to see how there were consistent patterns, for example, in the doc2vec data which seemed to tell us about the separability or inter-relationship between topics. (see fig4)

One other benefit of this ensemble approach was the use of confusion matrices to evaluate our work. As we said above, we didn’t have the benefit of using labeled data to evaluate our clusters. To the extent that our embeddings and clusters found similar things we were able to match them against each other and see what lined up and what didn’t.

Since our goal was to exploit variance in the essays as they related to pictures, too see if we could show that the same image can carry distinct meanings for students, we feel like we succeeded. We generated many specific examples of this, and also demonstrated with some visualizations and tables how the essays can be binned either together (as in the case of the LDA MLK essays) or apart.

**Fig 1 ICA representation of bert encodings of essays**

A graph of a graph

Description automatically generated with medium confidence

**Fig 2 Cutting a doc2vec embedding dendogram for clusters**

A graph with colored lines and numbers

Description automatically generated

**Fig 3: coloring parallel lines of LDA embedding with tree-based clustering of LDA embeddings**

**A graph of a blue and yellow line

Description automatically generated with medium confidence**

**Fig 4 exploring data structure with umap**

A screenshot of a computer screen

Description automatically generated