Examining Textual Similarities Between TED Talk Speeches and Speaker Biographies

Final Project for INFO 3350: Texting Mining History and Literature with Matthew Wilkens

By: Eliza Salamon

Code and data provided: https://github.com/esalamon17/TEDTalks

---

1. Introduction and Hypothesis

TED Talks are a ubiquitous cultural phenomenon — almost every year of my education has featured at least one TED Talk in class. Both traditional celebrities, business people, and experts in their fields are featured in these videos, giving speeches under twenty minutes that describe and analyze a wide range of topics: the TED Talk tagline is "ideas worth spreading." TED Talks are useful for providing an engaging and succinct introduction to audiences, but have also been critiqued for their surface-level talking points:

> But plenty of observers have argued that some of the new channels for distributing information simplify and flatten the world of ideas, that they valorize in particular a quick-hit, name-branded, business-friendly kind of self-helpish insight—or they force truly important ideas into that kind of template. They favor the kind of idea that fits into our “life hacking” culture: providing pointers or data that can be translated into improved productivity or happiness (often assumed to be the same). (Shea, 2014)

On its website, TED Talks has a page titled "Making Sense of Too Much Data," featuring talks highlighting big data. There's a lot of data out there — and the descriptions and transcripts of TED Talks themselves are included! From a Natural Language Processing perspective, TED Talks provide rich data through their transcripts. Multiple studies have analyzed TED Talk speeches in different ways. Some projects analyze how views are affected by factors such as valence and sentiment (Fischer et al., 2021), the demographics of their speakers (Sugimoto et al., 2013), and the topics discussed in speeches (Johnson et al., 2023).

I am particularly interested in connecting the critiques of TED Talks which assert that it "dumbs down" academia and research into a text mining project. In one research project, Peter Wingrove, using the same baseline as my anecdotal observation of TED Talk prevalence as a part of modern education, found that TED Talks have significantly less academic vernacular than university lectures (Wingrove, 2017).

I was further inspired by the methods outlined in the paper “RELIC: Retrieving Evidence for Literary Claims,” which looked to rank semantically related text (Thai et al., 2022).

My dataset, scraped and organized by Katherine M. Kinnaird and John Laudun for a 2018 paper entitled "TED Talks as Data," contains 992 TED Talks given between 2006 and 2018, with their full transcripts and speaker biographies. I am interested in the relationship between an author’s biography and the speech they give. Presumably, speakers invited to TED Talks give speeches about topics they are interested in, which will align with their background, career, accomplishments, etc, which are generally outlined in a person's biography. Due to the criticism of TED talks as being surface-level and predictable, I want to explore whether models are successful at matching speech transcripts to the biographies of their speakers — and hypothesize that they will not.

Hypothesis 1: Models measuring similarity and performing similarity retrieval tasks will not be successful at matching TED Talk transcript to speaker biography. Due to the broad nature and concepts inherent to TED talks, speech transcripts will not show strong similarities to speaker biographies in vocabulary and contextual methods that models provide.

Hypothesis 2: I propose that TED Talk transcripts relate to their speaker biographies in varied ways that extend beyond observable semantics.

2. Methods

Cleaning

Before creating my matching models, I cleaned the dataset. All the cleaning, methods, and testing were done in Python in Jupyter Notebook and Google Colab. This consisted of removing unwanted characters, including audience reactions such as “(Laughter)” that were included in speech transcripts. Additionally, I removed the starts and ends of the speaker biographies, which were consistent across the set. Additionally, I created a custom set of stop words, including the common English ones, that came up most frequently in the texts such as “hey” and “actually.” After the data cleaning, I embarked on testing three models to vectorize and embed text and then calculate similarities:

1. TF-IDF vectorizer with cosine similarity
2. LDA vectorizer with cosine similarity
3. BERT tokenizer with FAISS similarity

TF-IDF Vectorizer

A TF-IDF vectorizer tokenizes text based on based on the frequency of a word in a document compared to the appearance of that word across the entire corpus. My custom TF-IDF vectorizer resulted in a feature matrix with 958 features. I iterated through the corpus, vectorizing each transcript and calculating a cosine similarity with each vectorized speaker biography. A cosine similarity measures the similarity between two vectors. I obtained the top five matches with the highest similarity calculations. If the correct match appeared in that top five, I counted it as accurate. The correct biography match appeared in the top five matches 33.2% of the time.

LDA Vectorizer

Next, I vectorized the original speech transcripts using an LDA model. LDA, or Latent Dirichlet allocation, is a form of topic modeling that takes a vectorized corpus and organizes documents according to a predetermined number of topics. I chose twenty topics for my model, as this was a bit less than the number of topics provided in the TED dataset which appeared over fifty times. Similar to the above, I iterated through the corpus and calculated the cosine similarities. The correct biography match appeared in the top five matches 10.4% of the time.

BERT & FAISS

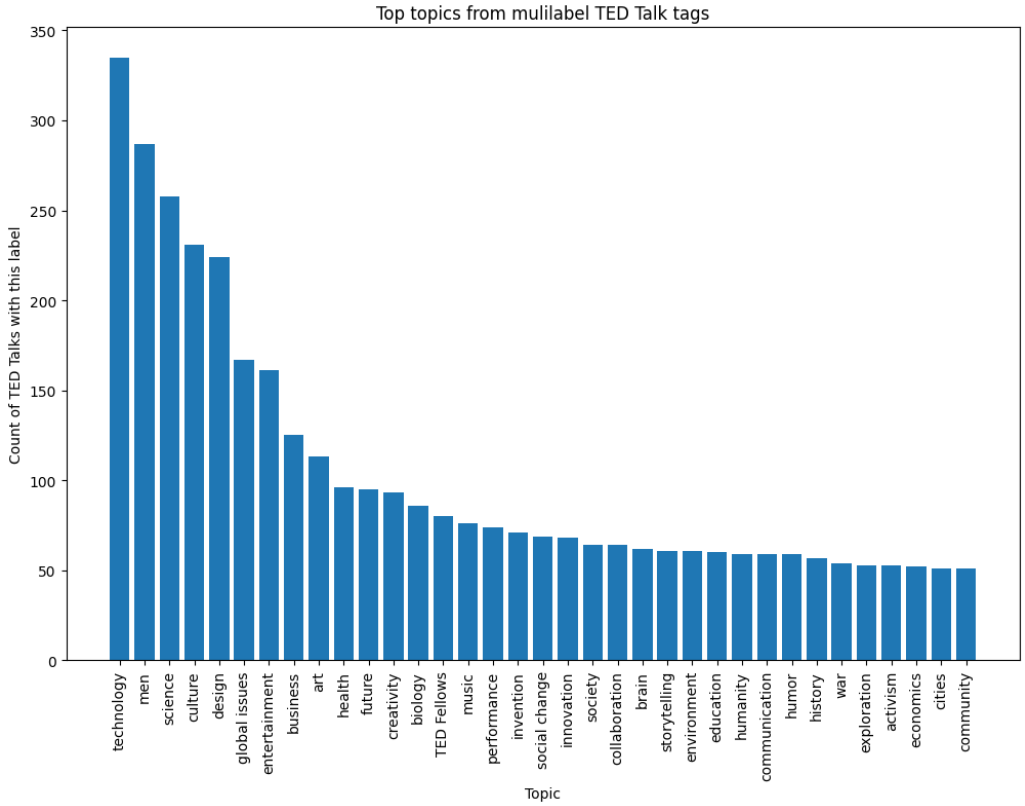
Finally, I encoded the speech transcripts using the DistilBERT uncased tokenizer and model from Hugging Face. DistilBERT is a pre-trained text embedder that retains contextual information in its tokenizers. After embedding the transcripts and biographies, I employed FAISS indexing. FAISS is an efficient similarity search algorithm developed by Facebook. Again, I found the top five similarity matches from the FAISS algorithm, which uses an L2 search metric. The correct biography match appeared in the top five matches 11.5% of the time.

Qualitative Manual Examination

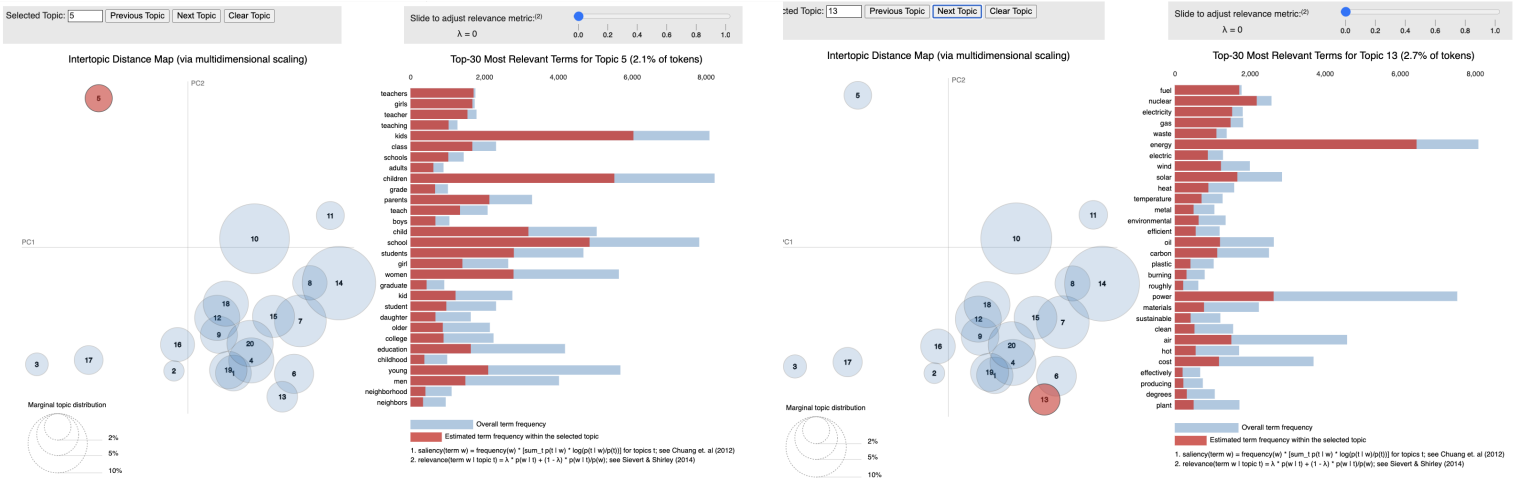
If speech transcripts and speaker biographies are not semantically or contextually similar, how else do they relate? I manually examined speeches and transcripts to gain a qualitative understanding of the dataset and the ins and outs of the three models. This fell into three parts:

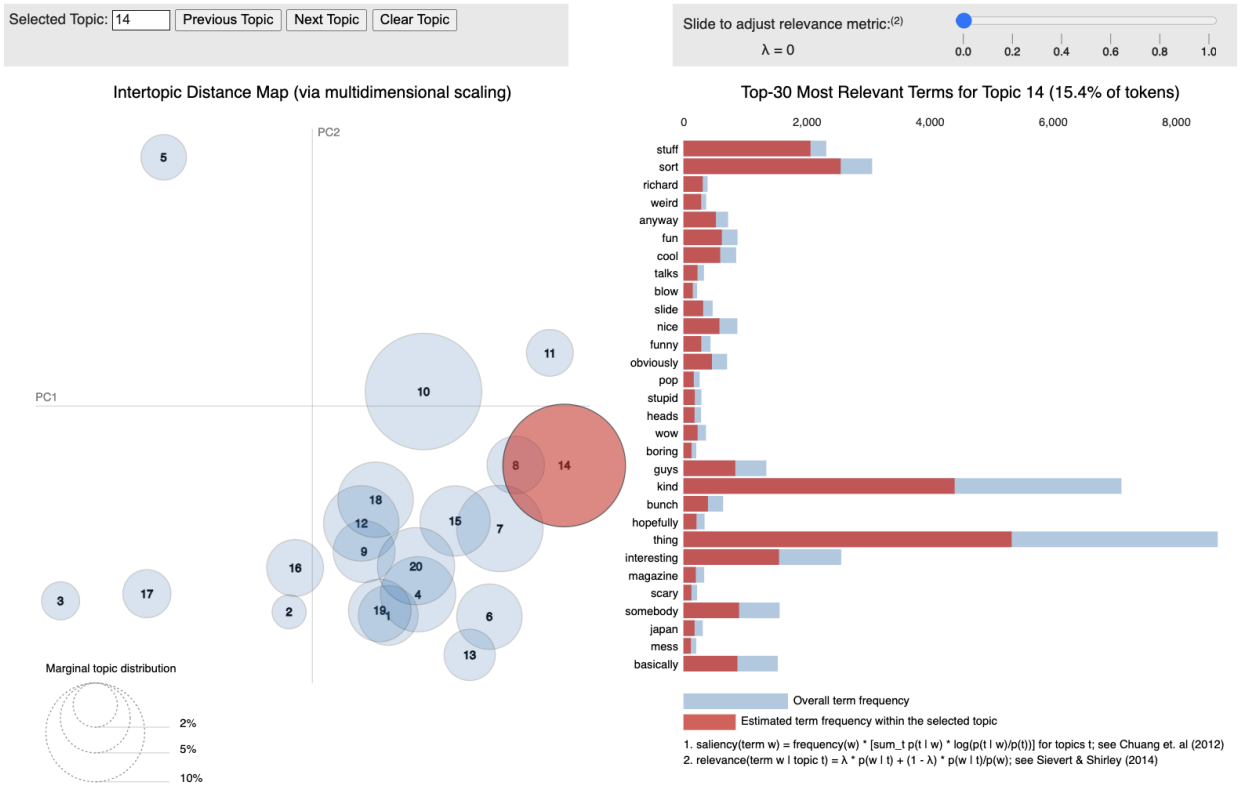
1. Examine prelabeled topics as well as topic modeling to examine groups of speech types.
2. Examine a sample of transcript/biography combinations to see if their connection is clear to a human reader (me).
3. Examine misclassified biographies to try to determine where the models went wrong.

The dataset came with multiple topic labels assigned to each speech on the TED website. The topics that appear in over fifty videos are shown below. 974 out of 992 of the speeches were categorized into at least one of these topics.



Next, I visualized the 20 topics I used for my earlier model to examine how the LDA topics differed from the TED-labeled ones.





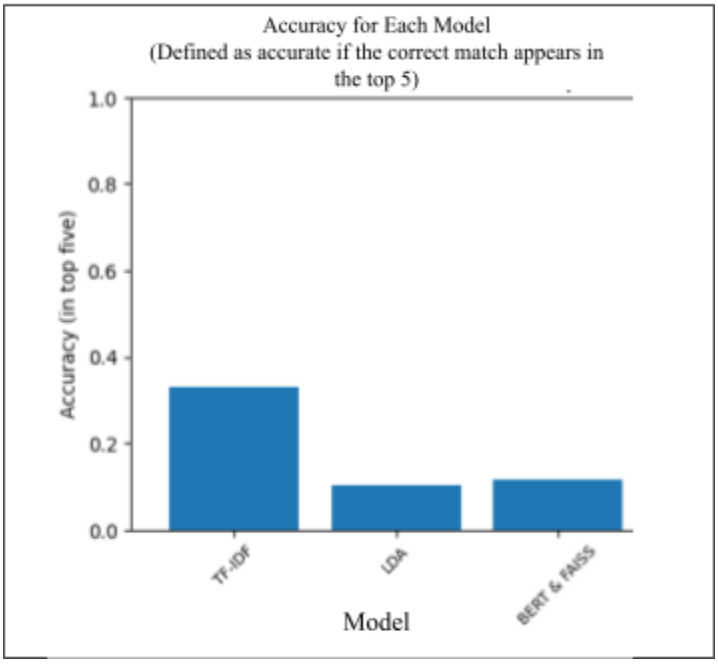
The LDA topic modeling reveals some interesting groupings within the speeches. Some topics are clear: 5 is speeches related to education and children, 13 relates to energy and efficiency, etc. The topic with the highest distribution of speeches, 14, does not have any clear patterns, and their most identifiable words appear pretty random and generic.

For the remaining three manual examinations, I randomly sampled about twenty speeches. I analyzed the relationships between speeches and their true biographies and speeches and their predicted similar biographies. The takeaways are outlined in the results below.

3. Results

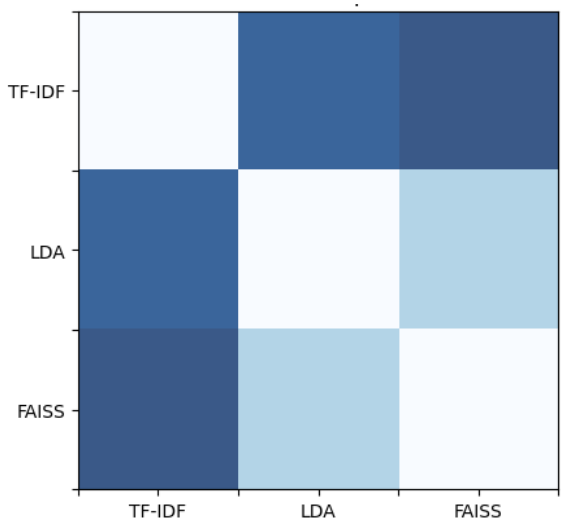
My results include the accuracy metric of my three models, a comparison of their labeling, and findings from manual examination.

The first three accuracies are the main models assessed for similarity: both TF-IDF and LDA vectorizers were used alongside similarity,



and a BERT tokenizer with FAISS similarity match was used for the third. We see that the TF-IDF vectorizer performed the best, though accuracy was still very low, around 33%. There was built-in leniency in this accuracy metric, as a match was considered accurate if it appeared in the top five most similar biographies, it did not have to be the most similar

When running the three models, I saved the correctly classified speech titles. I examined the overlap between what the models got right.



This color map shows the number of overlapped speeches matched correctly. Of the three models, TF-IDF got 326 "accurate" matches, as I defined accuracy. The darker the square, the more correct matches the models had in common. TF-IDF and FAISS had the most overlap, with 65 of the same speech titles. LDA and FAISS had the least overlap, with 20 of the same speech titles. Overall, there were only 13 speeches classified correctly by all three models. After examination, I have determined that there does not seem to be any relation in topic across these speeches.

Finally, my manual examination of speech/biography relationship and model performance highlighted some key patterns and takeaways. First, I wanted to ascertain if similarities were obvious to me, a human reader, between the transcripts of the speeches and the biographies of their speakers. I randomly sampled and manually examined twenty speech/biography pairs and found some major patterns in how they related to each other:

- **Speeches and their biographies *were* topically related to each other**

  One example of this is the speech *My Daughter Malala* by Ziauddin Yousafzai. His speech was about his daughter Malala. The speech and the biography are exactly aligned, as his notoriety is tied to his personal relationships and resulting work.

  Another example of this is the speech *Treat design as art* by Paola Antonelli. Her speech is about exhibiting and understanding design and her biography is about her work in museums and with design. Again — their speech is directly about their life's work.

- **Speakers had general biographies that didn't provide any context for the speech content**

One example of this is the speech *Should you live for your résumé ... or your eulogy?* by David Brooks. The speech is about legacy and human nature. David Brooks has a long resume: his biography lists being a columnist, commentator, author, professor, studying history and foreign affairs, and editing a book review. The topic of his speech doesn't have a clear tie to any of these things. I would not be able to blindly match biography to speech in this case.

Next, I randomly sampled and examined twenty speeches at random along with the top five most similar biographies. I did this using the TF-IDF vectorizer with cosine similarity model, as that had the highest overall accuracy. Again, I found some broad categories of patterns in how similarities were related.

- **Speeches were found to be most similar as speeches with authors of similar occupations/lines of work**

  One example of this is the speech *The Mathematics of History* by Jean-Baptiste Michel. Michel's biography outlines a general academic career using data to study human culture and his speech is about the mathematics of historical grammar. Though his biography was not matched in the top five, the other five biographies were all people who worked in history, language, and culture fields, all topically related to the speech.

  Another example of this is the speech *A Skateboard, with a Boost* by Sanjay Dastoor. Dastoor's biography is about his work as an engineer creating an electric longboard and working in robotics and the speech is about the electric skateboard, although *never mentions the word skateboard*. His biography was not matched in the top five, the other five biographies were people who worked in renewable energy research and development.

- **The speech was clearly matched to the correct biography**

  One example of the model performing well is the speech *The Illusion of Consciousness* by Dan Dennett. The themes of human consciousness in this speech were also clearly outlined in his biography.

  Another example of this is the speech *My Wish: Help Me Stop Pandemics* by Larry Brilliant outlining a need for pandemic preparedness. Brilliant's biography about his experience working as a doctor during health epidemics was the top result in the similarity model.

- **The model is picking up on a similar, false positive theme**

An example of this is the speech *Try Something New for 30 Days* by Matt Cutts about setting new goals. He repeatedly (and understandably) uses the phrase '30 days' in his speech. This concept is not related to his biography about his work as a search optimizer at Google. The biography that was found to be the most similar to his speech was that of Morgan Spurlock whose biography speaks extensively about his TV show '30 days'. The similarities between the transcript and biography, although not a match, are clear, and I feel I would probably tie them together if I labeled them manually.

- **The matches between speech and biographies seemed completely unrelated**

  An example of this total mismatch is the speech *17 Words of Architectural Inspiration* by Daniel Libeskind, which is about emotional responses to architecture. The top four most similar speeches were from people in the space exploration industry. Something about the transcript found false similarities in another topic genre, likely just the frequent use of the word 'space' which Libeskind used in his speech.

4. Discussion and Conclusion

My first hypothesis posited that models would not be successful at matching speaker biography to their speech, and that hypothesis was validated. Out of the TF-IDF/cosine similarity model, the LDA/cosine similarity model, and the BERT/FAISS model, the best one was the TF-IDf/cosine similarity, and the correct biography match only appeared in the top five most similar matches a third of the time.

There are a few reasons why I think these models did not perform well

1. The basis of the vectorizers for TF-IDF and LDA was the transcript data, and the biographies were vectorized off of those. I did this because the corpus for transcripts was much longer and varied, and because I wanted the topics outlined in the speech to be what connected it to the biography rather than biography-specific information such as educational background and place of employment.
2. From manual examination, there often was not a clear topical or even semantic relationship between a speech and its speaker's biography. A human performing this task likely would have performed mediocrely.

My second hypothesis was analyzed qualitatively, and this is where the takeaways are most interesting.

1. Biographies are inconsistent with each other. I am not sure if speaker biographies are created by the speaker or by the TED organization, but they differ greatly in their form,

length, and level of detail. Some biographies are very traditional outlines of a person's educational background, and a list of all their achievements, publications, etc. Others are short, quippy, and humorous. These distinctions make it difficult to analyze a person through their biography.

2. The backgrounds and works of speakers were not clearly related to their speech. Speakers who are household names, celebrities, or academic celebrities often gave speeches about topics unrelated to why they were well known, or only tangentially related to their field of expertise. Some people (ahem, Bill Gates), are so universally known and respected that they seem to be given leeway to speak on a range of topics.
3. The topics defined by TED itself are broad and intangible, which translates into the content of the speeches. Having read more TED Talks than I was ever expecting to, I found myself repeatedly asking what the point was of many of these speeches, or who their audience was. Though some were important calls to action on climate change and political advocacy, others were more thought experiments and personal anecdotes.

Overall, the results of this project imply that it is difficult to compare what somebody talks about and how they talk about it to their biography. In examining many TED Talks, I found that their focuses were often very broad — the climate change crisis, refugee crises, public health management, space exploration — or very narrow — their experience in a cult or the engineering feat of a camel. TED Talks are rightfully criticized for their oversimplification and the effects that they have on our ability to learn and think critically. In “Should TED Talks Be Teaching Us Something?,” the authors write, “TED gives learners a false sense of simplicity of the real world and reinforces a convenient approach to learning that one rarely encounters in everyday life” (Romanelli et al., 2014). The fact that speakers are difficult to match between speech and biography may not be a facet of their personality or expertise, but rather the way they must fit into the TED Talk mold with its prescribed timing and neat takeaways.

TED Talks platform certain voices over others, and this affects who is given leeway to speak about what. These speeches certainly have positives and negatives in terms of education, outreach, impact, and representation. At the end of the day, the big topics of culture, policy, war, and technology that are so often discussed in TED Talks (and can't forget about that 'men' category!), are not understood or solved through an 18-minute talk. A person's life cannot be summed up in their biography, nor can their interests and work be summed up in a speech. The results of this project demonstrate the intricacies of self-identification and expertise when trying to move or influence an audience and the constraints set by an organization which ostensibly must limit nuance to inform at all.

References

Fischer, O., Jeitziner, L., & Wulff, D. U. (2021, December 23). Affect in science communication: A data-driven analysis of TED talks on YouTube. https://doi.org/10.31234/osf.io/28yc5

Johnson, C. N., Khakhariya, J., Leung, C. K., Pazdor, A. G. M., Peters, S. J., & Salo, A. M. (2023). Mining Popular Trends from TED Talk Data. *2023 IEEE International Conference on Industrial Technology (ICIT), Industrial Technology (ICIT), 2023 IEEE International Conference On*, 1–6. https://doi.org/10.1109/ICIT58465.2023.10143092

Kinnaird, Katherine M. and John Laudun. 2018. TED Talks Data Set. https://github.com/johnlaudun/tedtalks/tree/master/data/Release_v0.

Romanelli, F., Cain, J., & McNamara, P. J. (2014). Should TED talks be teaching us something?. *American journal of pharmaceutical education*, *78*(6), 113. https://doi.org/10.5688/ajpe786113

Shea, C. (2014, April 14). *The New Academic Celebrity.* The Chronicle of Higher Education. https://www.chronicle.com/article/the-new-academic-celebrity/

Sugimoto, C. R., Thelwall, M., Larivière, V., Tsou, A., Mongeon, P., & Macaluso, B. (2013). Scientists popularizing science: characteristics and impact of TED talk presenters. *PloS one*, *8*(4), e62403. https://doi.org/10.1371/journal.pone.0062403

Thai, K., Chang, Y., Krishna, K., & Iyyer, M. (2022). *RELIC: Retrieving Evidence for Literary Claims*.

Wingrove, P. (2017). How suitable are TED talks for academic listening? *Journal of English for Academic Purposes*, *30*, 79–95. https://doi.org/10.1016/j.jeap.2017.10.010