Executive Summary

Team Supermassive Blackhole developed a semantic search model which allows users to input a description of a movie they forgot the title of, along with some filters, and receive an NLP model's top ten guesses as to what that movie is.

While this model is a prototype, its uses are manifold. For instance, search queries on streaming services like Netflix, YouTube, or Hulu may be snippets of what a user can remember from a TV show or from a YouTube creator they enjoy. Creating a feature that allows users to re-discover their favorite media adds value to all of these services by improving user experience. Moreover, when users successfully re-engage with their favorite media, it increases overall platform engagement. Finally, since some services like YouTube already include auto-captioning software in some languages, the existing infrastructure complements the needs of semantic search.

Alternatively, this prototype would be useful to media librarians helping people find a desired movie or source. Librarians add value by being able to interpret an individual's query and helping them to find the correct source that they are attempting to describe. Therefore, allowing the librarian access to simple semantic search technology (which can be tailored to their kind of expertise) improves their capacity to help library patrons. For example, our University's catalog search cannot handle queries like:

"Fisherman discovers pearl and tries to sell it but has misfortune".

Our kind of semantic search model would likely be able to recognize this sentence as referring to *The Pearl* by John Steinbeck.

This semantic search model is also deployable in broader contexts. For example, a call center may want to show a trainee worker an example of a customer's question about a new piece of technology. This semantic search model could be adapted so that a manager could input, e.g., "new iPad update issue with Google Drive." The manager could also filter by the month the call was received in order to ensure the call pertained to the desired update version. In this scenario, the semantic search model would return, based on the calls' transcripts, the most relevant example call logs for the call center worker to review.

After building our embedding using the Kaggle "Wikipedia Movie Plots", Kaggle "The Movies Dataset" , and "CMU Movie Summary Corpus", training our model using the scraped user-summarized plots from IMDB, our model prototype achieves 84% accuracy on the test data, which is a big improvement over the 21% achieved by our baseline model of comparing substring similarity scores.