Capstone Project 1: Project Proposal

Problem to Solve

TED (Technology, Entertainment, and Design) is a nonprofit organization whose core mission is to spread ideas in the form of short, powerful talks known as TED Talks. TED works to build a global community of thinkers and doers to make great ideas accessible and spark conversation in an effort to spread world-changing solutions. Living up to its organization's motto – *Ideas Worth Spreading* – TED Talks have become a popular format to share ideas on a range of topics and across a breadth of disciplines, including science, business, and the arts. TED's powerful approach to storytelling has acquired 60 million monthly unique viewers to the organization's library of TED Talks and YouTube channel.

With 35 people tuned in to a TED Talk every second, it's clear that there is a high demand for insightful lessons and innovative ideas that can lead to impactful change, whether it's in the classroom, boardroom, or courtroom. With such a broad, global reach, there is tremendous opportunity to amplify a company's ideas or brand's solutions by leveraging the power of the TED stage and applying TED's tried-and-true storytelling formula.

TED has created a special playlist of the most popular talks of all time. While it may be obvious why some talks made the list due to the attention-grabbing title (*Do Schools Kill Creativity?*), thought-provoking content (*How to Spot a Liar*), controversial subject matter (*Your Body Language May Shape Who You Are*), or speaker's celebrity status (David Blaine's *How I Held My Breath for 17 Minutes*), others are more surprising. Videos that have earned a spot on the popular playlist have received tens of millions of views and thousands of comments, generating high volumes of user engagement, a highly sought-after outcome. In fact, highly-engaged customers have been shown to buy more, promote more, and demonstrate more loyalty to a brand.

Companies that want to highlight how their ideas and solutions are creating a more user-friendly experience for their customers or how they are making a positive impact on the world would be interested in understanding how to harness the power of video content marketing while capitalizing on TED's storytelling approach. Specifically, knowing how to best utilize TED's verified format and platform to promote a product or service can increase user engagement on an organization's digital and social media channels, reach a larger audience through a shareable medium, and educate consumers on how a digital feature or tangible good can enhance their lives and allow them to better connect to the world around them, ultimately, building an emotional connection between a customer and a brand.

There are a number of factors to consider when predicting what makes a TED Talk generate high user engagement. With this in mind, this project will attempt to solve the following problem:

Which factors predict a high user engagement score for a TED Talk video?

Clients

TED Ideas Studio is an in-house brand studio that offers corporate partners like Delta, Ford, Google, Target, and Disney the opportunity to tap into TED's powerful storytelling expertise and utilize TED's signature format to create and distribute original custom content to directly impact business communication. Because of the breath of topics TED Talks cover, from education to healthcare, there is potential to apply this project's insights to multiple stakeholders. Clients from different industries will find value in understanding which topics or themes viewers are drawn to, and are more likely to engage with, based on the analysis of 2,464 transcripts of previous TED Talks.

The short films produced by TED's studio turn the spotlight on top business leaders across diverse industries who want to share a message with those within their organizations or those with a vested interest in the ideas being shared. Through stories, the leaders communicate how they are using their respective organizations to turn their ideas into workable solutions to create impactful change within their company's culture and amplify that impact within the larger community to solve global issues, to create opportunities for diverse voices to be heard, or to connect others through new technology. For example, the United States Postal Service (USPS) worked with TED to create a video showcasing how **our physical and digital worlds are intertwined**. The concept for the video is to illustrate how technology can be used to enhance our lives and help us connect with each other. The short film reveals the postal service's new platform called Informed Delivery, which takes images of the hard-copy mail pieces USPS processes for a customer and provides an email to show which pieces are scheduled to be delivered to the customer's doorstep.

The goal of this capstone project is to build three predictive models that would deliver valuable insights to potential key stakeholders who are looking for quantifiable ways to deliver their company's story to their target audience. Based on the findings of this project, clients would decide which factors (words, tags, titles) to either include or exclude from their videos to increase their likelihood of achieving higher user engagement (views, comments, ratings). This capstone could be expanded into a second project in which a recommendation system could be built that suggests which specific words, phrases, tags, or titles should be used to increase the user engagement score of a particular piece of content. Although video does include visual components that help to tell a story, TED Talks are more about the *content of what is said* than the creative techniques used to attract attention as is demonstrated in a standard television advertisement, which often employs techniques such as lighting, transitions, and music, among other devices, to get viewers to take notice of their product or service.

Data

The data that will be used for this project include information for all recordings of TED Talks uploaded to the official TED.com website until September 21, 2017. The data is contained within two datasets. The main dataset contains 17 columns, featuring metadata on the talks and speakers. The transcripts dataset contains 2 columns, featuring the official English translation of each talk and corresponding URL information. These open-source datasets were released on Kaggle, scraped from the official TED.com website and are available under the Creative Commons License. Click this **link** to access the datasets.

The main dataset variables include: the number of comments a talk received, a description of what the talk is about, the duration of the talk in seconds, the event where the talk took place, the timestamp of the filming, the number of languages the talk is available in, the first named speaker of the talk, the official name of the talk (includes the title and the speaker), the number of speakers in the talk, the timestamp for the talk's publication on TED.com, ratings given to the talk (a count of the positive and negative reviews), a list of related or recommended talks, the occupation of the main speaker, the tags (themes) associated with the talk, the title of the talk, the URL of the talk, and the number of views a talk received. The transcripts dataset variables include: the official English transcripts for 2,464 talks and the URL of each talk.

Approach to Solving the Problem

To solve the problem, **factors** and **user engagement score** will have to be defined. Based on the available variables collected in the two datasets, "tags", "title", and "transcript" will be **factors**. The "tags" variable represents single-word labels assigned by TED of the themes given to each talk. For the

"title" variable, NLPK will be used to extract the most frequently used parts of speech contained within the title of each talk. Instead of generating a list of the most frequently used words in the titles, this natural language processing method will show whether the titles contain words that belong to one of eight parts of speech (noun, pronoun, verb, adjective, adverb, preposition, conjunction, or interjection). This technique gives clients more of a deeper insight into which form of grammar is most engaging. For example, it may be that the most engaging talks use a lot of action words (verbs) in their titles. The "transcript" variable will be used to count the frequency of words within the text of each talk. If this project is extended, it would be more insightful to examine the frequency of entire phrases or complete sentences. A deep learning model would need to be applied to analyze this large-scale text dataset because the sentence lengths included in each transcript vary greatly.

"Comments", "ratings", and "views" will count towards generating a **user engagement score**, which is a combined measure of all three metrics. To achieve a high score, a TED Talk will need to have a high number of comments and views and receive a high count of positive ratings (adjectives used to describe a talk based on the viewers' feedback, including only those positive in sentiment). After the data is wrangled and the data exploration phase is complete, a threshold will be used to determine which talks are classified as high or low in user engagement on a scale from 1 to 0.

This project will build three separate models to predict user engagement:

> Model 1: Predict which **words in a TED Talk transcript** will have the **highest user engagement score** (a measure of 3 metrics, including highest # of *comments*, the most positive *ratings*, highest # of *views*).
>
> Model 2: Predict which **tags (themes) associated with a TED Talk** will have the **highest user engagement score** (a measure of 3 metrics, including highest # of *comments*, the most positive *ratings*, highest # of *views*).
>
> Model 3: Predict which **words (and their associated parts of speech) in TED Talk titles** will have the **highest user engagement score** (a measure of 3 metrics, including highest # of *comments*, the most positive *ratings*, highest # of *views*).

Deliverables

The deliverables for this project will include code, paper documentation, and a visual slide deck presentation – all of which will be posted on my **GitHub repository** and **LinkedIn profile**.