Springboard – Capstone 2 Project Updated Proposal - 2019-09-13
Tom Preston

In April 2019, my Springboard mentor, Srdjan Santic, agreed that for my NLP track, the Youtube and NLP based analysis would be my Capstone 2 proposal. Below is the original summary proposal aligned to the Springboard capstone 2 rubric.

NLP Track Proposal: Youtube – Finding the most applicable content for a specific topic

Capstone 2 Proposal Outline

1. What is the problem you want to solve?

Millions of people use Youtube daily for learning how to do things. Diverse topics from how to replace a watch battery to how to learn Python is available however you often have to sift through lots of videos to find the "really good ones". I personally use Youtube daily for finding training materials on Python and Data Science. My challenge is that Youtube's search is helpful but the search options are basically "relevance " (determined by Youtube), view count, and Ratings. I need a better way to assess and prioritize the content I am looking for.

My goal for this Capstone would be use the Youtube API to read in the video title, summary description, comments and likes / dislikes to find relevant content more quickly. I would use NLP techniques to review the document titles, summaries and comments to rank these videos to determine their relevance to my search criteria. I would test this on a couple subjects like python and golf instruction (subjects I personally spend a lot of time in Youtube looking at) to see if this approach improves upon my manual searches and then try it on some additional topic areas.

2. Who is your client and why do they care about this problem? In other words, what will

your client do or decide based on your analysis that they wouldn't have done otherwise?

Any user of youtube would be a potential client of this new feature. A user of this feature about fine tune their search parameters to better find the "diamonds in the rough" videos. For example, youtube's proprietary algorithms user the title, description, user watch times and user watch history to provide specific video search responses. My hypothesis allowing a user to pick videos by using items like the positive or negative sentiment of the comments (to also include the number of comments versus questions to clarify the video content) is a key leading indicator for a user to find relevant videos.

3. What data are you using? How will you acquire the data?

I will define a set of queries (to be coordinated with my Springboard mentor) around two specific areas, python and golf. Below is the initial set of draft queries.

Python Queries	Golf Queries
Using Itertools	Hitting the Driver
Reading CSV files	Bunker Shots
Creating DataFrames	Fairway Bunkers
Using Lamba functions	Putting tips
Using the OS module	Stop hooking your driver
Understanding Numpy	Chipping

Python sort functions	Pitch shots
For Loops	Topping fairway woods
Python dictionaries	Golf grip
Python Lists	Flop shots

I will read in the top 100 search responses from youtube (using youtube's standard "sort by relevance"). I will extract the video title, description, likes & dislikes, and the video comments (replies will be ignored).

4. Briefly outline how you'll solve this problem.

- Gather the video search data (title, description, likes, etc, number of comments) and order in the search results
- Gather the video comments
    - Use NLP techniques to analyze the comments
        - Separate comments from questions
        - Determine the number of positive vs negative questions
- Develop a scoring system to score and rerank the results
- Perform an analysis comparing youtube's "most revelance" results versus my revised scoring system to see how my search ranking differs from Youtubes
- Finally – I will watch a sample of the videos – from both the youtube search results and my search results to perform a qualitative comparison (ie – what videos did I find more relevant)

5. What are your deliverables? Typically, this includes code, a paper, or a slide deck.

As per the Springboard Capstone 2 rubric, the deliverables will include

- Data wrangling notebook
- Storytelling notebook
- Milestone 1 Report
- Baseline Modeling and analysis of results
- Final Report
- Presentation Slide Deck