Similar Staff Picks

Metrics included in model

There are a number of metrics available to compare video clips and assign similarity scores. I chose to use the following:

- Parent categories (any category that is a parent of another category)
- Secondary categories (any category that has a parent category)
- Caption
- Title
- Time Created

Metrics excluded from model

- I deliberately excluded total comments, total plays and total likes because they do not convey information about the similarity of two videos. For instance: if we have a video of Justin Bieber with one trillion likes I would consider another video of Justin Bieber with one hundred likes to be more similar to it than a video about fractal geometry with one trillion likes.
- I also excluded the thumbnails of the clips. Detecting signal from the thumbnail (and ignoring noise such as background color) would require a sophisticated neural network, which I am not yet confident building.

Architecture of model

Given a clip id, there were two ways to combine the similarity scores based on the metrics chosen to find the ten most similar clips:

1. Hierarchical model
    - Algorithm: Order metrics by importance. Given a clip id, determine the n most similar clips based on the most important metric. Then from that set of n clips, choose the m most similar clips ($m < n$) based on the next most important metric. Iterate for each metric (primary category, secondary category, caption, title, creation date.)
    - Strengths:
        - Makes intuitive sense – traversing down a tree where topmost level is broadest metric and bottom level is most specialized metric is how people naturally think about similarity.
    - Weaknesses:
        - Sizes of subsets (n, m in algorithm) are arbitrary.
        - Hierarchical order of metrics will work better for some clips than others – if we start with parent categories for a clip that only shares a parent category with one other clip we are in trouble.
2. **Aggregate Model – this is what I chose to use**
    - Algorithm: Normalize similarity scores for each metric (the reason for which is demonstrated in part VIII of my jupyter notebook.) Combine similarity scores for all metrics. This is demonstrated in part X of my jupyter notebook.
    - Strengths:
        - Easy and efficient.
        - Consistent for all clips.
    - Weaknesses:
        - Not natural / low interpretability.