YouTopic

Youtube Topic Analysis

Motivation

In case you aren't aware...

A lot of information is exchanged on youtube

Thanks to auto generated captions, nlp can be used to analyze it

Business use cases:

Check if video content is appropriate for product being marketed

Sentiment analysis on product reviews

Tools

Pytube: Extract captions and transcripts

Gensim: Text rank and sentence summarizer

Glove: Word embeddings

(Pattern): Parts of speech recognition (to be implemented)

Starting off..

Gensim is good.

Example Video: *3Blue1Brown -So why do colliding blocks compute pi?*

Keywords: ['block', 'theta', 'collision', 'energy', 'momentum']

Summary: “If that first block has a mass which is some power of 100 times the mass of the second, for example 1,000,000 times as much, an insanely surprising fact popped out: The total number of collisions, including those between the second mass and the wall, has the same starting digits as pi.”

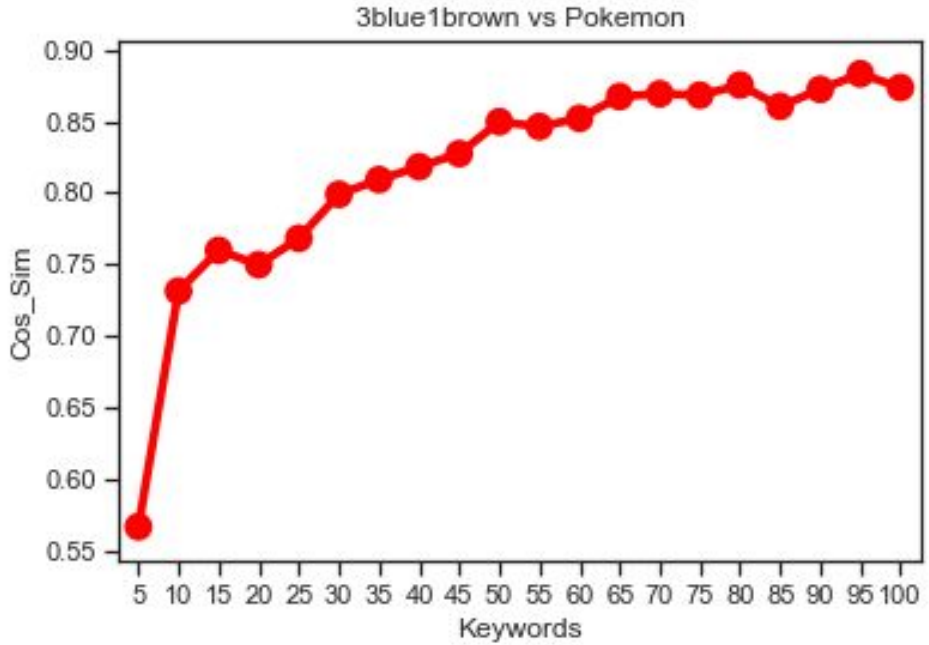
Word Space…. ?

Idea: Extract keywords from videos on same topic + Attach word embeddings +sum keywords + average over examples = A topic vector.

Goal: Perform topic analysis by comparing keyword vectors to topic vector.

Problem: Cosine similarity increases for all topics as the number of averaged vectors increases, in terms of both number of examples and keywords extracted

Cosine Convergence



Result

PART 1 of Michael Cohen Testimony Taking On President Trump

Keywords:['hes', 'people', 'ive', 'thats', 'happening']

Summary: “Putting up silly things like this all right so it really unbecoming of Congress.
I protected Mr trump for ten years.
Unlike my calling for trump that has a thousand followers hes got over sixty million people.
Have you ever seen Mr trump personally threaten people with the physical harm No. One he would use others.”

Topic Analysis: {Stormy Dan': 0.62, 'Border Wall': 0.60,'Mueller': 0.48, 'NBA': 0.44, 'Pokemon': 0.28}

Future Work

Use parts of speech filter for better keyword extraction(Pattern +python2.7)

FastText can handle OOV by taking roots

Manually generate topic keywords

Try word word mover distance between summaries

fin