バズりやすい動画の予測分析

G3
大城 龍太郎
新城 巧也

目次

1. 実験の目的
2. 目標までのアプローチ方法
3. データセットの詳細
4. 学習の進め方
5. 実験
6. 今後の課題

実験の目的

- 近年、様々なYouTuberが膨大な数の動画を投稿している。そのたくさんの動画の中で莫大な再生回数を稼いでいるものや、あまり再生されていない動画がある。
- 私たちの実験では、その再生回数に注目し、YouTuberが投稿している動画のどういった要素が再生回数と関係があるのかを分析する。

目標までのアプローチ方法

- 有名YouTuberであるHIKAKIN(HikakinTV)の動画に着目する
- APIを使いYouTuberごとの動画のデータセットがあるのでそれを利用する
- データセットの中には数値情報などがあるので様々な機械学習手法を用いる

データセットの詳細

- 動画のタイトル
- 視聴回数
- 高評価
- 低評価
- タグ
- etc..

21個ものculumがある！！

id	title	description	liveBroadcastContent	tags	publishedAt	thumbnails	viewCount	likeC
;Q	【大食い】超高級寿司店で３人で食べ放題したらいくらかかるの!?【大トロ1カン2,000円】	提供：ポコロンダンジョンズ \r\r\r\niOS：https://bit.ly/2sGg...	none	['ヒカキン', 'ヒカキンtv', 'hikakintv', 'hikakin', 'ひか...	2018-06-30T04:00:01.000Z	https://i.ytimg.com/vi/R7V5d94XkGQ/default.jpg	2244205.0	27
d4	【女王集結】女性YouTuberたちと飲みながら本音トークしてみたら爆笑www	しばなんチャンネルの動画 \r\r\r\nhttps://www.youtube.com/...	none	['ヒカキン', 'ヒカキンtv', 'hikakintv', 'hikakin', 'ひか...	2018-06-29T08:00:01.000Z	https://i.ytimg.com/vi/2R9_bkcWNd4/default.jpg	1869268.0	30
IPI	【悪質】偽物ヒカキン許さねぇ…注意してください！【なりすまし】	◆チャンネル登録はこちら↓\r\r\r\nhttp://www.youtube.com/...	none	['ヒカキン', 'ヒカキンtv', 'hikakintv', 'hikakin', 'ひか...	2018-06-27T08:38:55.000Z	https://i.ytimg.com/vi/EU8S-zxS9PI/default.jpg	1724625.0	33

学習の進め方

方法

- モデル
  - DecisionTreeClassifier
  - LinearRegression
  - Kmeans
- 形態素解析
  - janome.tokenizer

パラメータ

- DecisionTreeClassifier
  - DecisionTreeClassifier(max_depth=8)
- LinearRegression
  - デフォルトの値
- Kmeans
  - KMeans(n_clusters=15, random_state=0)

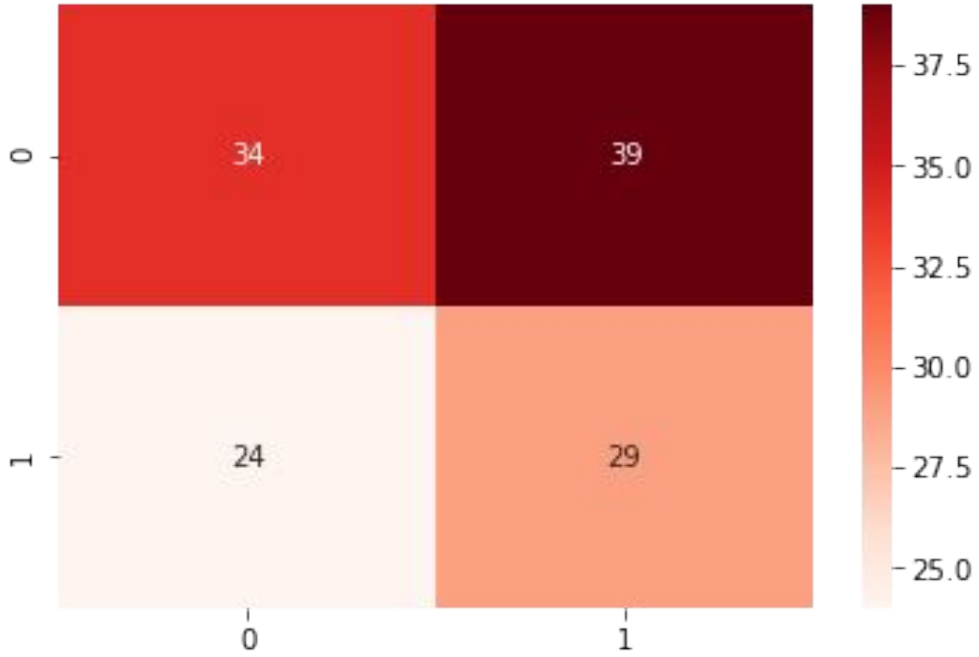
実験

実験 その1

1. 概要
   a. 1回目の実験では、タイトルが動画の再生数に影響をするのかを調査する。
2. 実験方法
   a. 形態素解析でよく使われる単語を使い、決定木で分類し可視化する。
   b. 教師データを再生回数 (300万以上を1)

実験結果 その1

- 精度
  - Train 60.3%
  - Test 50%



混同行列

実験 その2

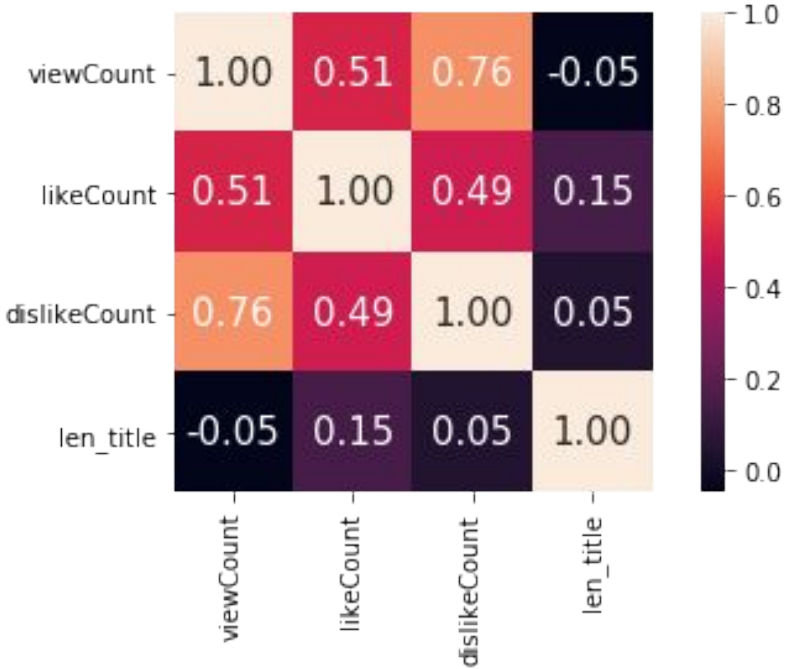
1. 概要
   a. データセットにあるタイトル以外の項目を説明変数として再生回数を目的変数とした線形回帰分析を行う
2. 実験方法
   a. 後退法
   b. 散布図行列と相関行列の作成

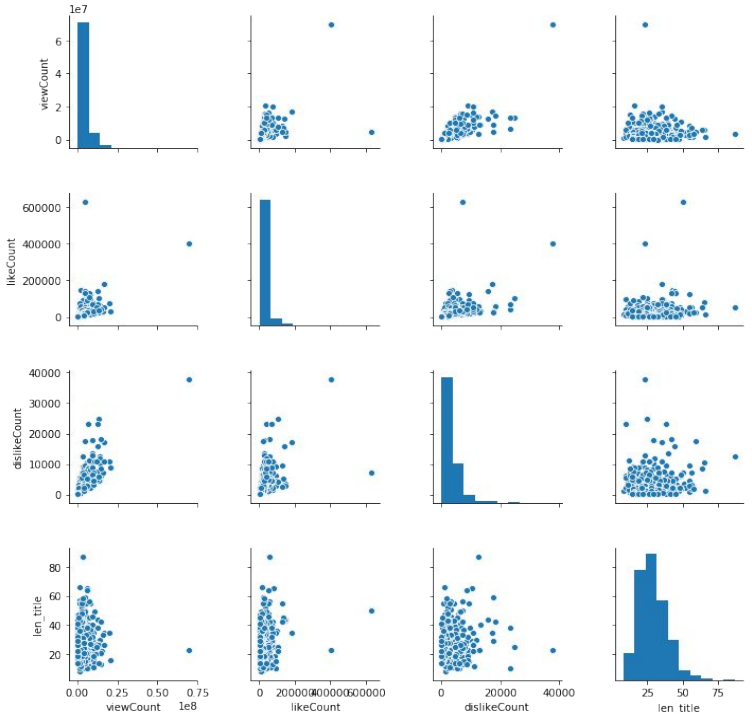
実験 その2

後退法

全ての特徴量を説明変数として、学習させてそこからもっとも精度がよくなる組み合わせを見つける手法

実験 その2


viewCount
likeCount
dislikeCount
len_title
1.00
0.51
0.76
-0.05
0.51
1.00
0.49
0.15
0.76
0.49
1.00
0.05
-0.05
0.15
0.05
1.00
1.0
0.8
0.6
0.4
0.2
0.0


1e7
viewCount
likeCount
dislikeCount
len_title
1e8

実験結果 その2

- 精度
  - Train 約62%, Test 約60%
- 説明変数
  - 高評価, 低評価, タイトルの長さ
    - ➔ 低評価がもっとも影響が大きかった。

実験 その3

1. 概要
   a. 動画のタイトルやタグを元にクラスタリングをしてどのような動画が人気なのか調査する
   b. 視聴回数と低評価の数を特徴量として用いる
2. 実験方法
   a. K-means法

実験結果 その3

1. 色が濃くなっているところは相対的に大きな値である

2. 低評価の数が多ければ視聴回数も増加する傾向にある

得られたクラスター

cluster: 0

379 アブラたりてます!? EDGE驚き&やりすぎ鬼背脂とんこつ醤油ラーメン食べてみた!
403 カップ麺にも激辛スパイス入れすぎてみた - EDGE 鬼辛とんこつ醤油ラーメン
408 【激辛】蒙古タンメン中本の『北極ラーメン』のカップ麺食べてみた!
411 激辛焼そばJACK食べてみた!
415 EDGE - カップ焼そばに激辛スパイス入れすぎた件 食べてみた!

cluster: 1

159 YouTubeテーマソング -Tetsuya Komuro Rearrange-
160 【1つだけ超激辛】フィッシャーズ&ヒカキンでロシアンチョコレートやったら爆笑しすぎたw
169 ゲーセンのクレーンゲームでお菓子とりまくってやんよ!
178 ヒカキン怒る!ピカチュウバッテリーの転売価格にブチ切れ【ポケモン】
181 復刻版ポケモンカード箱買いしたら大当たり!?

cluster: 2

0 【大食い】超高級寿司店で3人で食べ放題したらいくらかかるの!?【大トロ1カン2,000円】
1 【女王集結】女性YouTuberたちと飲みながら本音トークしてみたら爆笑www
2 【悪質】偽物ヒカキン許さねぇ…注意してください!【なりすまし】
4 【放送事故】酒飲みながら東海オンエア×ヒカキンで質問コーナーやったらヤバかったwww

	viewCount	dislikeCount
0	322835	607.667
1	3.49125e+06	3528.82
2	6.09851e+06	5107.6
3	3.87435e+06	3546.03
4	2.83249e+06	1517.69
5	1.02066e+07	5837
6	3.80434e+06	3327.07
7	4.2345e+06	3132.2
8	7.25843e+06	8631.25
9	3.72836e+06	3627.44
10	3.13781e+06	1735.33
11	3.38965e+06	2557
12	4.77750e+06	7023
13	4.4509e+06	3393
14	6.1163e+06	4591.75

実験結果 その3

- **お金を使う系の動画**
- **その時の流行に乗った動画**
- **高級なものを分解するなどの”もったいない”動画**

これらのジャンルの動画が低評価も多くもらいまた、視聴回数も多く稼いでいる

まとめ

1. 動画のバズる要素にタイトルは影響は大きくなかった。
2. 低評価が多いと動画のバズり度が高かった。
3. ある特定の分野の動画は低評価が多く比較的にバズっていた。

今後の課題

1. HIKAKINの動画でしか実験していないので他のYouTuberでも試す
2. 低評価以外の重要な要素をデータセットの中から抽出する
3. 重要な要素の調整が十分に行えていなかったので今後調整する
4. 今回は数字データのみを利用しているが、サムネイル等で画像処理などをすれば違った結果が得られる可能性があるので実践する
5. 他にもYouTubeのデータセットがあったので、それでも同様に実験したい

ご清聴
ありがとうございました