知能情報実験 III（データマイニング班）YouTube 急上昇に乗る動画のタグから、現在の流行を掴み、顧客をアシストする

グループの学籍番号列挙 215708A, 215758G, 215752H,215764B

提出日：2023 年 8 月 3 日

概要

本文書は、知能情報実験班 II(I データマイニング) 班 GP5 班の報告書である。GP5 班は、「YouTube 急上昇に載る動画のタグから、現在の流行を掴み、顧客をアシストする」ことを目的とし、現在の 流行となる動画を、急上昇に乗っている動画のタグから掴むことができる機械学習モデルの開発を 行っている。データセットは、YoutubeAPI を用いて情報を取得し.csv ファイル形式で構築して いる。学習モデルはクラスタリングを使用した。 しかし、結論として、このプログラミングは成功しなかった。原因はクラスタリングの使用はあまり適切でないことと、データ量が不足していたことである。本来、機械学習によってクラスタリングする時には単語を数値化させる必要があるが、今回の我々はタグの出現回数を単語の数値として扱ったためであると考察する。

目次

1　実験の目的と達成目標

私達 GP5 班の実験目的は「YouTube 急上昇に乗る動画のタグから、現在の流行を掴み、顧客をアシストする」ということである。YouTube の API から急上昇に乗っている動画のタグの情報を取得し、機械学習を使って現在の流行であるキーワードを抽出する。このキーワードは、今現在注目されている、もしくはされ始めているものであると考えられる為、これを新人動画クリエーターや企業などに提供することで、現在の流行のキーワードを共有し、アシストすることを目標とする。

2　テーマ：急上昇とは

YouTube でバズると言われると思い浮かぶ事として急上昇ランキングであろうと思われる。急上昇とは [1] によると現在話題になっていることであり、視聴者が動画をコメント、評価することで視聴回数の伸びに寄与する。本グループでは急上昇における動画のタグから、今トレンドとなっているキーワードを摘出できるのではと考え、対象問題として設定した。

3　実験方法

今回の実験では、トレンドとなっているキーワードは急上昇動画の中のタグでも頻出しているだろうと予測し、1 週間ごとに急上昇に乗った動画のタグを全てまとめて特徴量を抽出することで、その週に流行したタグを抽出した。また、トレンドとなっている内容であれば、何週間かにわたって急上昇に乗り続ける可能性が高いと予測し、過去の 1 週間の急上昇データと現在の急上昇データを比較することで、より大きな流行となっているキーワードの抽出を行いたいと考えた。今回の実験では下記の三段階の手順を踏む。

1. 自分の Google アカウントから「Youtube Data API」を生成しそれを用いて、現在の急上昇に乗っている動画の情報を取得する。
2. 取得した動画から頻度の多いタグ順に並べて csv ファイルに保存する。
3. 先ほどの csv ファイルは tag の数が多く出るため frequency が 2 以上であるものだけにフィルターをかけて新たに csv ファイルとして保存する。
4. 2 と 3 を数回繰り返し 1 週間ごとのデータを抽出する。
5. 過去と現在のデータを読み込み機械学習 (クラスタリング) にかける

3.1　データセット構築

今回は、既存のデータセットを用いるのではなく、[2] による YoutubeAPI を用いて情報を取得することで、新たにデータセットを作成することにした。取得する情報としては、現在の YouTube の急上昇 50 位までのチャンネル名、動画のタイトル、動画につけられている tag、投稿日、チャン

ネル ID、視聴回数、高評価である。取得した情報は csv ファイルに変換して保存している。この csv ファイルから tag の頻出度を数え、頻出度が 2 以上であるものをトレンドであると定義し新しく csv ファイルで保存している。

3.2 モデル選定

tag それぞれの重みをつけてジャンル分けをするために k-means のクラスタリングを使用することにした。なぜなら、私たちが今回行う実験では YouTube の急上昇から tag の情報を抽出するだけだと統計計算だけで済むから機械学習を取り入れるためクラスタリングを取り入れた。クラスタリングを行うことで視覚化しユーザにあったジャンルをアシストできると考えたからである。

3.3 パラメータ調整

クラスタ数を 4、初期セントロイドを 42 とし、その他はデフォルト値のままににしてある。

4 実験結果

今回の実験では YouTube API を用いて YouTube においての急上昇動画 50 本の情報を抜き出した (図 1)。そこからタグとして使用されている単語を抜き出し、出現回数と共に csv ファイルに記録した (図 2)。また、その際に比較的出現回数の多いタグを厳選したファイルも作成した (図 3)。その後クラスタリングによる機械学習を行ったが、実行した可視化するためのグラフが予想したものと異なっていた。

youtubetop_50

	0	channelTitle	publishedAt	channelId	title
0	E98gYs7hUJU	ジャにのちゃんねる	2023-08-02T10:00:19Z	UC2aIHD2WkakOiTxCxF-uMAg	#259 【ふっか博士召喚】僕山田の同期。
1	-BnK_7jGm3Y	HikakinTV	2023-08-01T10:00:13Z	UCZf__ehlCEBPop-_sldpBUQ	20億円のヒカキン新居ハウスツアー！超巨大室内温水プール&庭付きの家
2	JGDyoFr7a3U	LOVECARSITV!	2023-08-02T10:15:01Z	UCtLo4nwb3ObCDZ4m8b8u7fA	トヨタ ランドクルーザー250 実車詳細レビュー！ コレは欲しい！魅力的1台が登場！
3	OqWQDSTcMAE	スカイピース	2023-08-02T11:00:09Z	UC8_wmm5DX9mb4jrLiw8ZYzw	【カルマ降臨】牛角で人気ランキング1位～10位当てるまで帰れま10！！！
4	2xKZhora-og	JFATV	2023-08-02T11:32:16Z	UCgIeUSV91-FfmCayG4ISBcw	【LIVE】第103回天皇杯 ＦＣ町田ゼルビア vs.アルビレックス新潟｜ラウンド16（4回戦）
5	FHxsHL2QaTw	舞元啓介	2023-08-01T16:56:34Z	UCJubINhCcFXlsBwnHp0wl_g	【#にじ甲2023】熱狂！にじさんじ甲子園2023 vol.9【にじさんじ/舞元啓介】
6	6sKKvf5p450	カラフルピーチ	2023-08-01T08:00:07Z	UCh7bThzNArch6TgoHB1HVhA	殺人ピエロがまた現れた！？【ミニゲーム】
7	1xYXGRiYrOo	トヨタ YouTubeショールーム	2023-08-02T01:17:43Z	UCxvxlxnS9cjPyeebbiOd4Ig	トヨタ自動車 新型ランドクルーザー　ワールドプレミア
8	S0itP92FO3I	IKITERU【イキテル】	2023-08-01T11:01:28Z	UCBBLybG7OxGtnaGuJAr1xjw	【人生初】韓国の高校生が初めて日本に来て衝撃!!! 教科書に書いてあった国と全く違って1日目の24時間が驚きの連続！
9	x-eEFx3QvLU	FNNプライムオンライン	2023-08-01T11:30:31Z	UCoQBJMzcwmXrRSHBFAITsIw	長男・宏一前副社長に大声あいさつ　ビッグモーター幹部ら"環境整備点検"映像入手
10	ukFrEn0GGMI	jun channel	2023-08-01T15:04:06Z	UCx1nAvtVDlsaGmCMSe8ofsQ	夏のホラゲー配信１【Shadow Corridor dlc】
11	Zi1EP0Hh2Es	不破 湊 / Fuwa Minato【にじさんじ】	2023-08-01T13:02:36Z	UC6wvdADTJ88OfIbJYIpAaDA	【重大発表アリ】FuwaHiba Session LIVE【不破湊/渡会雲雀/にじさんじ】
12	SkyQR3Tdl_E	TREASURE (트레저)	2023-08-01T15:00:05Z	UCx9hXYOCvUYwrprEqe4ZQHA	TREASURE - 'BONA BONA' DANCE PRACTICE VIDEO
13	P77PjzKDMWU	【東海テレビ公式】ドラHOTpress	2023-08-01T12:58:30Z	UC2G0uq2KGFkonnGByWeNZ_Q	8/1 中日×阪神 ハイライト
14	AW43iGJms1M	バンカラジオ / Bankaradio	2023-08-01T09:05:27Z	UCT5BVYrrhS7gD5xzloZ8FhA	ボケカの転売で荒稼ぎする小学生
15	twj2h5qXXtQ	SAWAYAN CHANNEL / サワヤン チャンネル	2023-08-01T11:00:41Z	UChdm3NwkRpuyCAxVRTzPQ9g	【超過酷】減量中のヤンと１日全く同じ食生活したらまさかの結果に、、、
16	lyF8-YdAtkk	ぼへチャンネル	2023-08-01T11:00:38Z	UCu0HaR1ul4RBgbdHnd5GWAQ	ポケモン界の歌姫メロエッタを救いたい【ポケモンSV】【ゆっくり実況】
17	-tL8GzT3CK0	アニメ『おでかけ子ザメ』チャンネル	2023-08-01T11:00:09Z	UCj5n8eLbvEGyK_bSYu51KCQ	アニメ『おでかけ子ザメ』第1話「映画」
18	8LGZDFbfkIE	日テレNEWS	2023-08-01T00:26:10Z	UCuTAXTexrhetbOe3zgskJBQ	【ビッグモーター不正】“新たな疑惑” 現役従業員「整備せず納車」証言　会社側の回答は…
19	0bIRwBpBcZQ	JYP Entertainment	2023-07-31T08:58:09Z	UCaO6TYtlC8U5ttz62hTrZgg	ITZY "CAKE" M/V @ITZY
20	nVPy7HpkuZw	MBS NEWS	2023-07-31T09:33:16Z	UCkKJhK073xF1pK5h9R82ZGQ	【ビッグモーター】副社長のパワハラLINE「死刑死刑死刑…」「実質支配の兼重親子との関係断つための民事再生法」山岸久朗弁護士の解説【MBSニュース解説】（2023年7月31日）
21	jhDMUN1OXyk	HikakinGames	2023-08-01T11:00:11Z	UCX1xppLvujI03ubLio8jslyA	10年間ありがとうございました。
22	v7ePavkHFh0	サントリー公式チャンネル （SUNTORY）	2023-07-31T12:24:06Z	UCyhBnB7_27IhyaYgjCf9Z4g	【プロ野球界のレジェンドが集結】サントリードリームマッチ2023アーカイブ配信
23	CYSisYBrLis	にこちゃん放送局	2023-07-31T11:30:03Z	UCmWkowGJ1I7PRjlWWz85xTw	【裏側密着】ディズニー＆ピクサー映画の声優を務めることになりました
24	aRWr1a915ys	THE RAMPAGE from EXILE TRIBE	2023-07-31T12:00:07Z	UCYI3GYW6aVsxr0NZW86KXQ	THE RAMPAGE / Summer Riot ～熱帯夜～ (MUSIC VIDEO)
25	SW5IAMcjUy4	Apex Legends	2023-07-31T15:00:11Z	UC0ZV6M2THA81QT9hrVWJG3A	Apex Legends: Resurrection Launch Trailer	Kill Code - Part 2
26	6-vMeNZJlus	椎名唯華 / Shiina Yuika	2023-07-31T12:42:57Z	UC_4tXjqecqox5Uc05ncxpxg	【#にじ甲2023】＃９　にじさんじ高校3年目春～【にじさんじ/椎名唯華】
27	K_f1aUbPxa0	斎藤裕	2023-07-31T00:02:37Z	UCeK1fLPjfHJYZ-Ucr0VK4cw	超RIZIN2の感想
28	JOpWVU_F13c	RKB毎日放送NEWS	2023-07-31T09:11:54Z	UCgrmHAblzN-QzMMgI2oNRog	ビッグモーター　現役店長と社員が証言　「ＬＩＮＥで１日２０００件の業務連絡」「罪１本生えていたら減点」
29	t4MqKdlQR20	oricon	2023-07-30T11:33:35Z	UCbZvkG2uAgr6Olva4FytscQ	【超RIZIN2】朝倉未来、ケラモフに衝撃の一本負け　試合直後の心境明かす「強かったです…」『超RIZIN.2』試合後インタビュー
30	YhI7F0sKwPA	ChroNoiR	2023-07-31T10:00:12Z	UCz6vnlbgiqFT9xUcD6Bp65Q	【このツイート誰クイズ】にじさんじライバーのバズったツイートを学ぼう！ #くろなん
31	tM_3DNXB41c	らっだぁ	2023-07-31T09:03:42Z	UCINdVSK1uy2xFHTaXZYXopw	気付いたら自分以外屋医者でワンオペ状態になっていたらっだぁ【VCRグラセフ/VCRGTA】
32	FIcKd9T2B7Q	ONE PIECE公式YouTubeチャンネル	2023-07-30T01:00:04Z	UCdAHaWcKdpbT5XkN2Er6BUQ	ONE PIECE 1071話予告「ルフィの最高地点　到達！"ギア5。」
33	dGITkGxhKxM	ホモサピ	2023-07-30T13:28:44Z	UCd0hscDvJvzRbo8Rk7JPQMA	カメムシ潰してレモンを絞るとコーラの香りになる
34	Uhk6afwm2lw	おろちんゆー	2023-07-30T08:28:45Z	UCypYz0HzH-2oL7mdP-JK2yQ	猫にセミ奪われかけたwww
35	zEIgogugvqc	ヒロヤとニッシーちゃんねる	2023-07-30T15:30:23Z	UCsH9FFOI1sY5ItGEanCpgkQ	【超RIZIN.2】試合を終えて｜ヒロヤVS伊藤裕樹
36	Kb7H7yB5osg	스브스케이팝 ZOOM	2023-07-30T09:00:10Z	UCM3jwNR15-WBVzgT6DsaEQ	[안방1열 직캠4K] 정국 'Seven (feat. Latto)' (Jung Kook FanCam) @SBS Inkigayo 230730
37	OHJA5AnA3ck	東京03 Official YouTube Channel	2023-07-30T03:00:07Z	UConlcs8o0z5vYTamGtq-Lsw	東京03 -「スイッチ」／『第23回東京03単独公演「やな因果」』より
38	i-PnC1tYnGM	Snow Man	2023-07-28T14:00:09Z	UCuFPaemAaMR8R5cHzjy23dQ	Snow Man「Dangerholic」Music Video YouTube Ver.
39	Tg4SVNgsJdg	RIZIN FIGHTING FEDERATION	2023-07-30T14:08:11Z	UCZZOUGjWsRdM8_5bsqtxYaQ	【勝者と敗者】試合直後の選手の素顔に密着【 超RIZIN.2 / RIZIN 】
40	SRk0t1_uQQE	King & Prince	2023-07-30T03:00:13Z	UCSxwcQnzA5K6DvofvBn6ATA	King & Prince『なにもの』@「音楽の日2023」
41	ocDDLPLKe70	BANGTANTV	2023-07-29T13:00:08Z	UCLkAepWjdylmXSltofFvsYQ	[슈취타] EP.15 SUGA with 정국
42	Yq7j1rn41Fo	金原正徳の金ちゃんTV	2023-07-30T12:03:27Z	UCtX7oQLlXYNoqb2gdJ3s-nw	【速報】壮絶決着！波乱こそ格闘技の醍醐味【超RIZIN2】
43	t_grw5cgOiE	はじめしゃちょー（hajime）	2023-07-30T10:55:54Z	UCgMPP6RRjktV7krOfyUewqw	【ポケカ】黒炎の支配者を100ボックス開封してみた！！！
44	uc4gOOGHNn0	sasatikk	2023-07-30T11:00:40Z	UC79849JZ6Up4hMgQbZnU3nA	予想だにしない結末を迎えた山本署長663円誘拐事件【スト鯖GTA】
45	NkfVifYqZ40	タイピー日記/taipi	2023-07-30T11:00:13Z	UCIDOHrFd4wI8P6pPyPKlG2A	全身ノミのうんちだらけの子猫をシャンプーしてみたら。。
46	gmShyA236KM	スーツ 交通 / Suit Train	2023-07-30T11:00:08Z	UCxBR2bnAFAavDHpHtQrTA9Q	珍しく東海道新幹線が遅れているので乗れません！
47	-8OPhlg8zmE	スカイピース	2023-07-31T11:00:12Z	UC8_wmm5DX9mb4jrLiw8ZYzw	【ありえない】横浜アリーナのライブで問題が発生しすぎていた件
48	uol3dg7BrMo	ぺえ	2023-07-30T11:00:43Z	UC0YsAIBJC0yx1sCbshABidA	【仲里依紗＆ぺえ】カラフル女が集まったら盛り上がりすぎて3150
49	AqEqeiqskEI	Marine Ch. 宝鐘マリン	2023-07-30T13:03:20Z	UCCzUftO8KOVkV4wQG1vkUvg	【#宝鐘マリン生誕LIVE2023】3rd Generation Girls Band !!!【ホロライブ/宝鐘マリン】

図１　今急上昇の動画 50 本のデータ

w_f_2023-08-03

keyword	frequency
MLB	1
メジャーリーグ	1
EPL	1
プレミアリーグ	1
SPFL	1
スコティッシュプレミアシップ	1
三笘薫	1
セリエA	1
エンゼルス	1
大谷翔平	1
ダルビッシュ有	1
吉田正尚	1
冨安健洋	1
FAカップ	1
古橋亨梧	1
YG Entertainment	1
YG	1
와이지	1
K-pop	1
트레져	1
트레저	1
TREASURE	1
CHOI HYUN SUK	1
JIHOON	1
YOSHI	1
JUNKYU	1
MASHIHO	1
YOON JAE HYUK	1
ASAHI	1
BANG YE DAM	1
DOYOUNG	1
HARUTO	1
PARK JEONG WOO	1
SO JUNG HWAN	1
최현석	1

図 2　急上昇の動画 50 本のデータからタグの出現数をまとめたデータ

w_f_2023-08-03_filt

keyword	frequency
マネーの虎	2
岩井良明	2
マネーの虎 岩井	2
マネーの虎 現在	2
投資	2
融資	2
資金調達	2
起業	2
ビジネス	2
経営	2
UCTyKZzmKi95wxmCg9rU-j6Q	2
モノリスジャパン	2
れいわのとら	2
TikTok	2
アイルトンモカ	4
本当は不良なのに	2
神回	2
お笑い	2
面白い	2
ゲーム実況	3
ヒカキン	2
ひかきん	2
ユーチューバー	2
YouTuber	2
Vtuber	4
葛葉	3
くずは	2
kuzuha	2
吸血鬼	2
にじさんじ	3
nijisanji	3
いちから株式会社	2
ichikara	2
anime	2
アニメ	2

図 3　図 2 において 2 回以上出現したタグをまとめたデータ

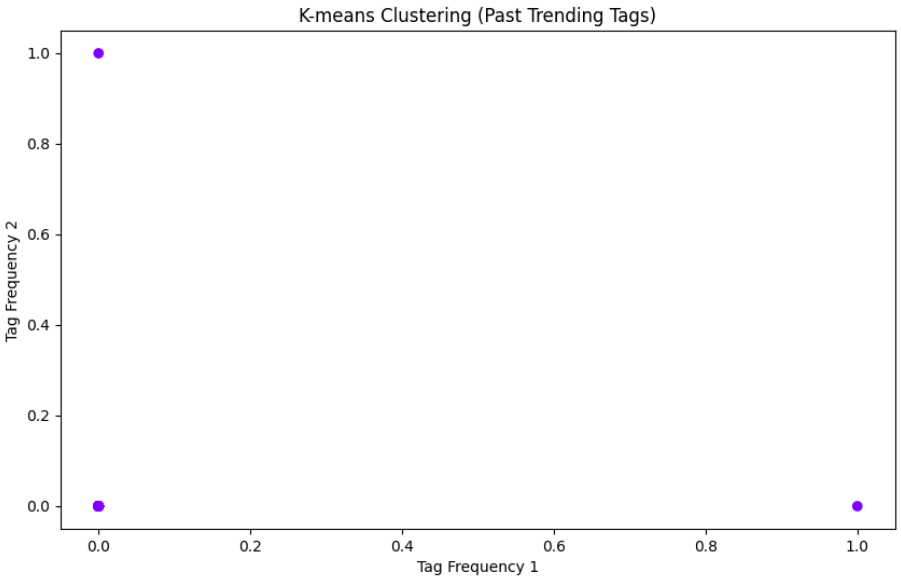

図 4　過去のデータのグラフ

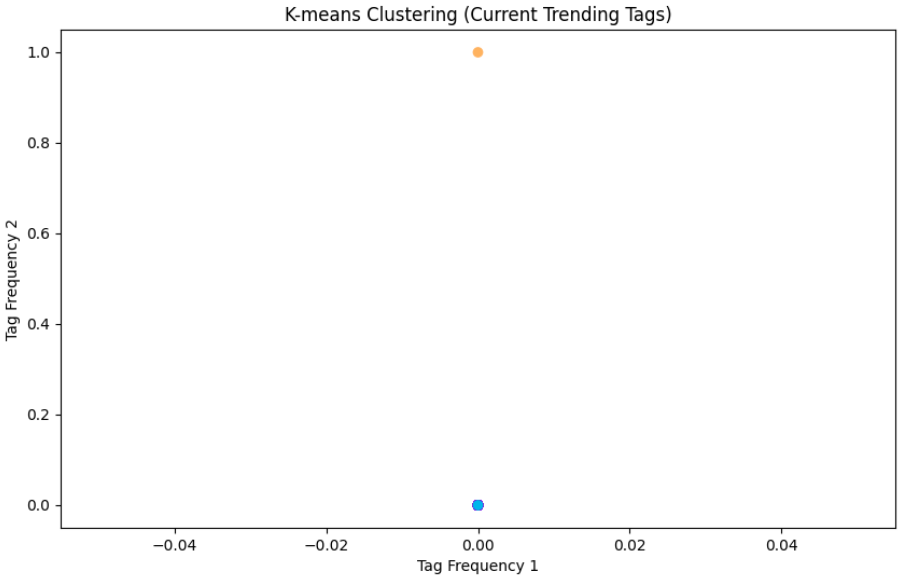

図 5　今週のデータのグラフ

5　考察

今の実験で採用した方法では、目的を達成できなかった。理由としては、以下の 2 つが考えられる。

1. 機械学習の学習方法が今回の目的に適していなかった。
   今回使用した機械学習は主に、クラスタリングの k-means 法と CountVectorizer である。本来、機械学習によってクラスタリングする時には単語を数値化させる必要があるが、今回の我々はタグの出現回数を単語の数値として扱ったため上記ような結果になってしまったと考える。また、CountVectorizer を用いた機械学習も並行して行ったが、そもそも CountVectorizer ではあるテキストから出現する単語のカウントを特徴量にする手法であるが、今回の実験では単語と単語同士で出現する CountVectorizer を使用してしまったため、出現した単語数がカウントしたベクトル化できなかったと考える。
2. データの量が不足していた。
   急上昇というジャンルは比較対象が 50 しかなく、機械学習を行うにはあまりにも少ないデータであった。

6　意図していた実験計画との違い

今回の結果を経て、反省点が多く挙がり、大きく感じた反省は、「目標の共有」である。一回目の授業では以下の目標ではなく、もう少しざっくりとした目標になっており、そこでは全員の目標が決まったと考えていた。しかし、メンバー各々が少しずつ違った目標をイメージしており、今回の目標が決定するまでの間、メンバーの認識に誤差が生まれてしまった。この修正に時間がかかってしまったことも今回のプロジェクトを完成できなかった要因の一つだと考える。

今回の失敗を踏まえ、より良い取り組みを行うために考えたことは 2 つである。一つ目は、事前準備をもっと行うことである。ざっくりとした目標定義でだけではなく、具体的な目標を決め、また、どう言った経緯で各々がどのような動きをするのかをよく話し合う事が必要であると考えた。二つ目は、メンバー間での話し合いである。動きが決まっているとしたとしても、急な変更や臨機応変に対応していかなければいけない場面はとても多く存在すると、今回の取り組みで実感した。行っている課題の進行状況やアイデアだけではなく、毎回授業初めに目標を確認したりすることで認識の誤差を減らすことが出来るのではないかと考えた。

7　まとめ

今回の実験で学んだことは、情報の共有の重要性である。各々の考えや思いの共有、話し合いを沢山しなかったことで、発生している問題が多々あった。これが出来ていれば、実験目標のズレだけでなく、1 人が用事や体調不良で休んだとしても遅れることなく作業が出来、発表でもスムーズに発表することが出来たと考えられる。

参考文献

[1] YouTube ヘルプ, <https://support.google.com/youtube/answer/7239739?hl=ja&sjid=14891511027695807438-AP>, 2023/06/08.

[2] ForestSeo, Python と YouTubeAPI で急上昇の動画のデータを取ってくる, <https://qiita.com/g-k/items/7c98efe21257afac70e9>, 2023/06/08.