機械学習特論

LDA(Latent Dirichlet Allocation)に用いた

日本・韓国
YouTube Video Topic 分析

システム情報科学府　情報学専攻
冨浦研究室　M1

韓　範錫 (ハン　ボムソク)

目次

- データセット
- LDAとは
- LDA分析の結果
  - 映画・スポーツ・ゲーム・ニュース
  - ゲーム

データセット

https://www.kaggle.com/datasnaek/youtube-new

Trending YouTube Video Statistics (kaggle)

- 流行したYoutube Videoの情報 (2017~2018)
- USA・Britain・Germany・Canada・Japan・Koreaなど
- 動画のViews・Likes・Unlike・Published_Date・Comment_Countなど：「Numeric Data」
- 動画のTitle・Tags・Description：「Text Data」

今回使用したデータセット

- 「日本・韓国のYouTube Videos」　（日本：2018年、韓国：2017・18年）
- Title + Tags + Description の「Text_data」から、「名詞」だけを抽出
  - Mecab-ipadic-neologd
  - Mecab-ko-dic　　で日本語・韓国語の形態素分析を行う。

LDAとは

- 「Latent Dirichlet Allocation」の考え方
  - 人々は文章を書く時、いくつかの「Topic」の中で一つを、確率的に選ぶ。
  - 選ばれたTopicに似合う「単語」を確率的に選ぶ。
  - 選ばらた単語で、文を書く。

すでに生成されている文章から、上記した「確率」を追跡すること。

文の集合から、TOPICを抽出するプロセス

LDAとは

- 文章を書くための「確率表」
  - Doc3の著者は、Topic1とTopic2を、「33%：61%」考えている。
  - Topic1に似合う単語 ✳ ：99.7%、Topic2は ✳ ：27.5% 、 🐱 ：72.3%

	Topic 0	Topic 1	Topic 2
Document 0	0.290	0.507	0.203
Document 1	0.007	0.986	0.007
Document 2	0.681	0.007	0.312
Document 3	0.051	0.333	0.616
	Topic 0	Topic 1	Topic 2

	Topic 0	Topic 1	Topic 2
✳	0.003	0.997	0.275
🎩	0.995	0.001	0.002
🐱	0.003	0.001	0.723
	Topic 0	Topic 1	Topic 2

- Topicの選び・単語の選びを繰り返し作った文章

Document 3 🐱✳✳✳🐱✳✳✳🐱✳✳🐱🐱✳🐱🐱🐱✳🐱🐱✳🐱🎩✳🐱✳🐱✳✳🐱✳✳🐱

https://lettier.com/projects/lda-topic-modeling/

LDA分析の結果（日本・韓国）

- 使用したCategory：映画・スポーツ・ゲーム・ニュース
  - Category別の「Text_info」の数　　：約1000～3000個
  - Topicの数のセッティング　　：８個

日本 ：Topic・Topicに対する単語の寄与度

```
(0, '0.015*"日本" + 0.011*"選手" + 0.009*"仮想通貨" + 0.009*"日大" + 0.009*"海外の反応" + 0.007*"動画" +
(1, '0.013*"日本" + 0.009*"スイング" + 0.007*"報道" + 0.007*"ゴール" + 0.007*"特注" + 0.006*"日本代表" +
(2, '0.019*"的中" + 0.011*"馬連" + 0.011*"予想" + 0.011*"ポケモン" + 0.010*"映画" + 0.010*"競馬" + 0.007
(3, '0.010*"さん" + 0.008*"山口達也" + 0.008*"コチラ" + 0.008*"アニメ" + 0.006*"ニュース" + 0.006*"監督" +
(4, '0.031*"釣り" + 0.024*"動画" + 0.014*"登録" + 0.011*"リスト" + 0.009*"素材" + 0.009*"公式" + 0.008*"
(5, '0.035*"ワンピース" + 0.019*"ハイライト" + 0.015*"大谷翔平" + 0.012*"大谷" + 0.011*"試合" + 0.011*"動画
(6, '0.006*"作詞" + 0.005*"人生相談" + 0.005*"編曲" + 0.005*"作曲" + 0.005*"アニメ" + 0.005*"ナウ" + 0.0
(7, '0.016*"羽生結弦" + 0.012*"動画" + 0.010*"ＮＨＫ" + 0.009*"選手" + 0.008*"羽生" + 0.008*"平昌五輪" +
```

韓国 ：Topic・Topicに対する単語の寄与度

```
(0, '0.080*"영화" + 0.024*"추천" + 0.015*"리뷰" + 0.012*"영상" + 0.010*"무비" + 0.009*"손석희" + 0.009*"
(1, '0.031*"마인" + 0.030*"베이" + 0.030*"크래프트" + 0.027*"블레이드" + 0.024*"구독" + 0.021*"게임" + 0.0
(2, '0.033*"문재인" + 0.021*"김어준" + 0.019*"회담" + 0.015*"정상" + 0.014*"이명박" + 0.012*"정치" + 0.00
(3, '0.015*"김정은" + 0.015*"기자" + 0.012*"트럼프" + 0.009*"미국" + 0.008*"정규재" + 0.008*"드루" + 0.00
(4, '0.024*"인피니티" + 0.022*"영화" + 0.020*"어벤져스" + 0.018*"마블" + 0.016*"영상" + 0.013*"리뷰" + 0.0
(5, '0.060*"북한" + 0.048*"메카드" + 0.041*"평창" + 0.036*"공룡" + 0.027*"시즌" + 0.025*"올림픽" + 0.022
(6, '0.031*"라이더" + 0.028*"워치" + 0.027*"영상" + 0.021*"오버" + 0.020*"로얄" + 0.020*"구독" + 0.019*"
(7, '0.042*"대통령" + 0.013*"대표" + 0.012*"기자" + 0.010*"타운" + 0.010*"은행" + 0.010*"김성태" + 0.009
```

LDA分析の結果（日本・韓国）

- 使用したCategory：映画・スポーツ・ゲーム・ニュース

T	重要単語	
1	日本・選手・海外の反応	スポーツ
2	山口達也・さん・ニュース	ニュース
3	ワンピース・ハイライト	映画
4	羽生結弦・動画・オリンピック	スポーツ
5	的中・馬連・予想	スポーツ ゲーム
6	釣り・動画・登録	スポーツ
7	作詞・編曲・アニメ	映画

日本

T	重要単語	
1	映画・おすすめ・レビュー	映画
2	マイン・クラフト・ゲーム	ゲーム
3	ムンジェイン・イミョンバク	ニュース
4	キムジョンウン・トランプ	ニュース
5	マーブル・アベンジャーズ	映画
6	北朝鮮・オリンピック・平昌	スポーツ
7	オーバーウォッチ・動画	ゲーム

韓国

差異：映画への認識・Youtubeでのニュース

LDA分析の結果（日本・韓国）

- 使用したCategory：ゲーム
  - Category別の「Text_info」の数　　：約1500個づつ
  - Topicの数のセッティング　　：５個

T	重要単語
1	マイクラ・トミー・カンタ ・水溜りボンド(Youtuber名)
2	モンスト・ファイトリーグ・ モンスターストライク
3	ポケモン・攻略 ・やまだ(Youtuber名)
4	ゲーム実況・再生・プレイ

日本

T	重要単語
1	オーバー・ウォッチ・ ラナ(Youtuber名)
2	マイン・クラフト・ アゴ(Youtuber名)
3	壺ゲーム・アホ・ ・デド(Youtuber名)
4	クラッシュ・ロワイヤル・動画

韓国

流行りのゲーム・Youtuberがわかる

考察

- EDA(Exploratory Data Analysis)の不在にも関わらず、悪くない結果
  - データセットの詳しい情報まで把握できなかった。
  - リンク（http://…）・記号・１文字単語を削除する簡単の前処理を行った。
  - Topicの数も詳しく考えずに、様々なことを試してみた。
  - 日本・韓国のYouTubeユーザーの使い方の違い、興味のあるニュースなどが把握できた。

  - 個人的にゲームというCategoryは、馴染みのないこと。
  - 当時（17・18年）どのようなゲームが流行し、どのYouTuberが有名であったかがわかった。

機械学習特論

ご清聴ありがとうございます。