

ソーシャルビッグデータを用いた 地域のトレンド分析

新美 舞[†] 小舘 亮之[†]

[†] 津田塾大学大学院理学研究科

1. はじめに

近年, SNS の普及により, ソーシャルデータは膨大な量となっている. また, SNS によって日々, 新たなトレンドが生み出され, ユーザ間で広く共有されており, 観光事業に用いられる例も多い. それに伴い, ソーシャルビッグデータを用いて観光情報を抽出する研究は盛んに行われている. [1][2]本研究は, ソーシャルビッグデータから地域ごとのトレンドを発見すること, トレンドの分布から文化圏を見出すことを目的として, Twitter や Instagram といった SNS 等からソーシャルビッグデータを収集し, 解析する.

2. 言語処理

地域ごとの言語による情報を収集するために, if-this-then-that (IFTTT)[3]を利用する. IFTTT とは, ソーシャルメディアやプラットフォームを仲介する, Web サービス連携ツールである. マイクロブログサービスである Twitter から, 図 1 のように, IFTTT を通じて特定の地域名を含む投稿 (ツイート) を収集し, Google spreadsheet に保存する. オープンソース形態素解析エンジンである MeCab を用いて形態素解析し, 名詞のみを抽出して, 図 2 のようなワードクラウドを作成した. 他の人のツイートを再び投稿する, ”リツイート”を意味する”RT”や, URL の一部である”http”, ”https”などのノイズが多く含まれていることがわかった.



図1. データ収集手順



図2. ワードクラウド

3. 画像解析

地域ごとの画像情報を収集するため, 4K Stogram[4]を利用する. 4K Stogram は Instagram のダウンローダーである. 図 1 のように, 4K Stogram のハッシュタグ検索によって, 地域名のハッシュタグを含む投稿の画像を Instagram から収集する. 収集した画像はピクセルごとに色を読み込み, Python の機械学習ライブラリである scikit-learn を用いて, k-means 法でクラスタリングする. クラスタ数はさしあたり 5 とした. 各クラスタの中心の座標の色をドミナントカラーとして抽出し, 1 枚の画像につき 5 色のドミナントカラーを得た. 抽出したドミナントカラー (HSV 値) の色相 (Hue) を 10 度ずつで分類し, 色のグループに分けた. この色グループから, 今後, トレンドを定義する.

5. まとめ

本稿では, IFTTT や 4K Stogram を利用することでソーシャルビッグデータを収集し, 地域のトレンドを分析する方法について検討した. ツイートのノイズ排除, クラスタ数, アウトプット手法は今後の課題とする. また, 画像解析に AI を用いて, 物体認識を取り入れ, 地域ごとに多く撮影される被写体を推測することを検討したい.

参考文献

- [1] 土田 崇仁, 遠藤 雅樹, 加藤 大受, 江原 遥, 廣田 雅春, 横山 昌平, 石川 博 : Word2Vec を用いた地域やランドマークの意味演算,
<<http://db-event.jpn.org/deim2016/papers/167.pdf>>
(参照 2019-02-08)
- [2] 佐伯圭介, 遠藤雅樹, 廣田雅春, 倉田陽平, 横山昌平, 石川博 : 外国人 Twitter ユーザの観光訪問先の属性別分析,
<<https://db-event.jpn.org/deim2015/paper/286.pdf>>
(参照 2019-02-08)
- [3] IFTTT<<https://ifttt.com/>>(参照 2019-02-08)
- [4] 4K Stogram
<<https://www.4kdownload.com/ja/products/product-stogram>>
(参照 2019-02-08)