# 知能情報実験 III（データマイニング班）
# Twitter 上のテキスト文を対象とした「コロナで何が困っているのか」を見つける

グループの学籍番号 205759A, 205720E, 205763J, 195719J

提出日：2022 年 6 月 9 日

# 目次

# 1　はじめに

## 1.1　概要

本実験では、グループでテーマを決めて、半年間でデータ解析に取り組むグループワークである。グループワークを通して、機械学習や実験再現のためのドキュメント作成等を目指す。

## 1.2　テーマ：Twitter 上のテキスト文を対象とした「コロナで何が困っているのか」を見つけるとは

本グループでは Twitter 上のテキスト文を対象として「コロナで困っていること」を見つけることを対象問題として設定した。SNS を使ってコロナで発生している問題を可視化し、件数が多い項目を整理することで、その後の改善策を見出して今後に応用できる。それによって現在発生している問題に対処しつつ今後同様の問題が発生する際により効果的な対策ができるようになることが期待できる。

# 2　実験方法

## 2.1　実験目的

コロナで何が困っているのかがわかることで、その後の改善策を見出すことができ、今後に応用するために、Twitter 上のテキスト文を用いて解析する。

## 2.2　データセット構築

Twitterscraping.py を実行することで、Python の snscrape ライブラリ・モジュールを使用して、Twitter 上のテキスト文のなかから「コロナ」という単語が含まれた文章を抽出し、それらをデータセットとして sample.csv に出力した [2]。次に、extract_content.py を実行して sample.csv を読み込み、content カラムだけを取り出すことで、tweet の文章だけを content.csv に出力した。negaposi.py を実行することで、この csv ファイルと、日本語評価極性辞書を利用した Python 用 Sentiment Analysis ライブラリ oseti を使ってネガティブとポジティブに分けた [3]。そして、ネガティブに判定された tweet だけを抽出し、nega.csv に出力した。

## 2.3　モデル選定

実験目的に沿って、教師なし学習で文書のクラスタリングを行うため LDA(トピックモデル) を使用した。また、トピックモデルは pyLDAvis を用いてインタラクティブに可視化できる部分も LDA を選んだ理由である。

### 2.4　パラメータ調整

クラスター数 (カテゴリ数) の最適数を見つけるため、クラスター数３～５０の結果をメンバーで確認し、トピック名をつけやすいクラスター数を探した。他のパラメータは参考にしたサイトの値をそのまま使用した [4]。

## 3　実験結果

pyLDAvis を使用してトピックモデリングの内容を html に出力した結果は以下の通りである。(リンク) (途中)

## 4　考察

実験結果より、クラス 1,3 は「コロナ感染症状」、クラス 2 は「政治」、クラス 4 は「メディア」についてのトピックと考えた。また、今回データセットの構築を行う際「コロナ」という単語を含めたツイートを取得していたため、トピックごとに占めるコロナという単語数の割合が多いトピックが今回の実験目的に即していると考える。したがって、コロナによって困っていることのうち、改善する優先度の高いトピックは、コロナ感染症状、ニュース、政治の順であると考える。しかし、snscrape を用いて「コロナ」という単語を含めたツイートを取得する際、snscrape は単語だけでなく、ユーザ名やタグからも取得することに加え、同じ tweet にコロナという単語が複数回使われている可能性もある。そのため、一概にコロナという単語数の割合が多ければ優先度の高いトピックとはならない可能性もある。また、今回の実験は参考資料をもとに行っていたため、LDA のパラメータを理解し、その調整などを行えていない。より、今回の実験に合わせたモデルの構築を行うことができればより良い結果が出るのではないかと考える。さらに、データの数としてネガティブとして判定されたツイート数は約 1 万だった。今回の実験としてこのデータ数が良いのかということもこれから考えなければならない。(これに実験結果を踏まえた部分の考察も書き加える)

### 4.1　考察のトピック数について

## 5　意図していた実験計画との違い

## 6　まとめ

## 参考文献

[1] レポート作成の手引き　レポートの基本的形式に関するガイド, https://www.kanazawa-u.ac.jp/wp-content/uploads/2015/01/tebiki2.pdf, 2020/07/02.

[2] Men of Letters（メン・オブ・レターズ） - 論理的思考/業務改善/プログラミング, https://laboratory.kazuuu.net/using-python-to-scrape-social-networking-sites-using-snscrape/, 2022/07/02.
[3] oseti による日本語の感情分析, https://note.com/npaka/n/n3c7722d2e4bc, 2022/07/02.
[4] Shingo の数学ノート, http://mathshingo.chillout.jp/blog27.html, 2022/07/05.