BSHES 797R/GRAD 700R
Module B, Assignment 1

August 28, 2021

Assignment 1: Analysis of drug-related chatter

Outline:

In this assignment, you will be working on a simplified version of a real-life scenario.

Due to the COVID-19 pandemic, many people who use opioids (PWUO) have had difficulties accessing treatment. Two evidence-based medications for opioid use disorder (MOUD) are methadone and buprenorphine-naloxone (typically referred to by the trade name Suboxone). You have been approached by a researcher who is interested in knowing how the pandemic is affecting access to MOUDs. Due to the pandemic, there is no real-time data available and the researcher is worried that getting access to data from traditional sources (*eg.*, overdose death counts, emergency department visits etc.) will take time. She is worried that the impact of lack of access to treatment can be very bad. She also suspects that a large number of overdose deaths are happening as a result of fentanyl and its analogs (eg., carfentanil). She is interested in knowing some real-time information and thinks social media might provide some insights. There are a few things she is interested in:

- What are the general topics of conversation associated with methadone and Suboxone?
- What are the topics associated with fentanyl?
- Which of these three substances are being discussed the most?
- Are there big differences in chatter about these medications before and after the emergence of the pandemic?

Data

To derive insights from social media data, you are provided with a small sample of tweets (Note that in real life, the volume of information is much much larger). The tweets were collected using the Twitter streaming API. There are two files with data collected at different times. To reduce the sizes of the data, simplified CSV files are provided. One set was collected before there was much attention about COVID-19 and the other was collected after.

Datasets: Set 1; Set 2.

Tasks

To provide insights to the researcher, you will perform the following tasks and/or answer the following questions:

1. What are the date ranges for the two sets? What information are provided in the CSV files? What are the languages in which tweets have been posted? (1 point)
2. What is the total number of posts in set 1? What is the total in set 2? (1 point)
3. How many tweets are there for methadone, Suboxone and fentanyl in total? Tip: sometimes alternative expressions are used for substances (*eg.*, *fent* for *fentanyl*). (2 points)
4. Are there fentanyl analogs that are also being discussed (*eg.*, carfentanil)? (1 point)
5. What are some of the topics that are most closely associated with each of the three substances? Top 5-10 topics (if relevant) is acceptable. (2 points)
6. Among the three substances and the tweets containing them, which two sets of tweets are more similar than the other? There are several ways you can approach this problem. For example, you could take a sample of tweets mentioning each of these substances and represent each set as a single document. Or you may compare tweets from one set one-by-one with a sample of the tweets from another set. Any reasonable approach for this task is acceptable. Explain your method in the report (4 points).
7. Generate word clouds for each set, so that they can be shown to the researcher. (2 points)
8. Generate appropriate time-series figures to compare how the frequencies of mentions of these substances differ. (2 points)
9. Find the top 10 most frequent bigrams in each of the three sets. Plot a bar chart for these. (2 points)'
10. Write a report (described below) for your experiments and results. (3 points)

Submission requirements

You are required to submit a report (2 pages max plus figures/tables). Use Times New Roman, Georgia or Palatino Linotype font (11pt, single-spaced, justified)

- Describe the **methods** you used in the report and any intuition behind choosing specific NLP methods

- Present the **results** and a brief **discussion** of any insight you may have obtained.
- Your report will have to contain a link to your code. Ideally, the script/code should be on GitHub or Bitbucket. However, you will not be graded down for other mechanisms of code sharing.