Experiments With Topic Models

Team member names: Ragy Costa de Jesus, Diego Grisales, Oscar Fox

COLAB's

Unigram:

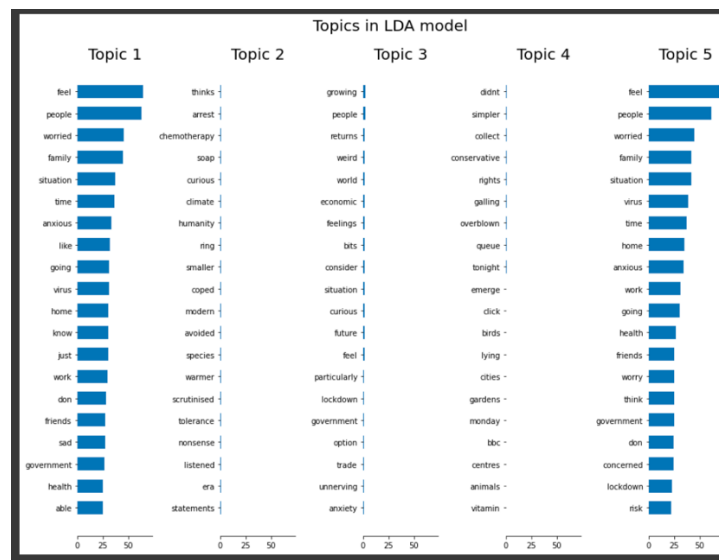https://colab.research.google.com/drive/1PSC83ImA4qQynZp6-A-0AF7Pe74I2g2V?usp=sharing

Trigram:

https://colab.research.google.com/drive/1Q2LtWTb1H-ms9WdiZQSEqjOyADJ0MHTq

Result & Observations

The experiment gave us totally different results depending on whether we used unigram or trigram. In this experiment we used TF-IDF vectorizer on text from covid data. First, we start by using the unigrams (dataset 1) to separate the words and most of the results ended up in topic 2. We concluded that the vocabulary in topic 2 resembles human experiences that they go through. In addition to the unigram, we also conducted the trigram experiment which resulted in more evenly data across the board as shown in dataset 2. The trigrams resulted in more specific emotions being displayed in the topics based on some human experience. For example, topic 3 vocabulary has more anger emotions in them than topic 1. It also can be that they all are displaying different levels of sadness and anger.

During our experiment with Latent Dirichlet Allocation (LDA) we encountered variations of topics depending on the parameter and which engram to use. For example, one of the experiments we performed was using a max_df of 0.95, a min_df of 2, and excluding stop_words. The results showed that topic 2 had most of the vocabulary. However, when we commented out the min_df and max_df the vocabulary split between topics 1 and 5. The words with the most frequency was similar in both topics as you can see in image 1. In addition, we also did an experiment using the stop words and most of the vocabulary shifted to topic 3 which were mostly stop words on top of the frequency table.

LDA topic outputs:

Unigrams:

Topic 0: relaxing, unknowns, updates, communication, element, options, resulting, interacting, create, gloves

Topic 1: feel, people, worried, family, situation, time, virus, anxious, home, going

Topic 2: delivered, exciting, failed, riding, folk, responsibilities, released, adhered, lease, warmer

Topic 3: availability, difficulty, answers, scientific, reset, numb, amounts, observing, dislike, middle

Topic 4: epidemic, afraid, annoys, ban, nights, horror, sufficient, resolve, 65, savings

Trigrams:

Topic 0: worried family members, day day life, lost loved ones, observing social distancing, situation makes feel, prime minister intensive, minister intensive care, don know going, makes feel anxious, worried contracting virus

Topic 1: seeing family friends, having stay home, doing right thing, spend time family, feel quite anxious, seeing friends' family, able spend time, anxious current situation, family loved ones, family friends feel

Topic 2: people aren taking, feel sad people, losing loved ones, high risk category, aren taking seriously, toll mental health, people ignoring government, feel anxious situation, mental health issues, following lockdown rules

Topic 3: people taking seriously, underlying health conditions, feel angry people, feel like lot, aren taking seriously, people following rules, people aren taking, year university student, feel anxious long, worried current situation

Topic 4: social distancing rules, feel like people, family friends worried, worried catching virus, husband key worker, light end tunnel, need stay home, corona virus situation, worried family friends, sad people lost

Gender COLAB's

Male:

https://colab.research.google.com/drive/1WYes253p-J3NyKJZbAzkQj3XOTNfG1oh?usp=sharing

Female:

https://colab.research.google.com/drive/1-MzwTBSyVXDY9xqCyQKBUsDjpYZlWtNA?usp=sharing

Result & Observations

The second parts of the experiment we divided the dataset into two groups female and male. In this experiment we used TF-IDF vectorizer on text from covid data. We started with the female unigram and trigram. First, we ran the unigram without stop words and a max_df = 0.95 and a min_df = 2, because we wanted to compare the result with our previous experiment. As a result, the female unigram concentrated into topic 5 considerably. Most of the words resemble family-oriented concerns which make sense. Secondly, we did the trigram for the females we kept the same parameters as for the unigram and the vocabulary spread evenly across all the different topics. The topics seem to reflect different levels of emotions but closely related to anxiousness.

The male experiments were also run under the same parameter list as the female experiment. However, the results of the unigram shifted from topic 2 to topic 1 which is what we expected. In our opinion, the results illustrate the different concerns or through process of between the two different genders. In addition, during the male trigram experiment the result were like the female trigram. However, the difference was that the male vocabulary was more evenly distributed in frequency than the female. In other words, the female was repeating the same vocabulary, therefore increasing the word frequency.

Other experiments we conducted were female and male without max_df and min_df on the unigrams. The female unigram resulted with the most vocabulary concentrated in topic 4 while the male unigrams was more concentrated in topic 1.

LDA topic outputs:

Unigrams (Female):

Topic 0: unnecessary, ruined, shelves, unknowns, areas, wages, attitude, statistics, escalated, truth

Topic 1: emerge, leaders, restricting, employees, recovery, didnt, city, task, forseeable, establishment

Topic 2: feel, people, worried, family, situation, time, anxious, virus, home, going

Topic 3: pregnancy, celebrate, straight, hardly, reacting, overseas, treating, tension, deliberately, swiftly

Topic 4: list, answers, created, write, letting, flatten, exercising, effecting, excuse, doors

Trigrams (Female):

Topic 0: worried friends family, husband key worker, social distancing rules, worried family friends, feel anxious going, doing right thing, self employed people, feel anxious family, mental health issues, family member friend

Topic 1: people aren taking, aren taking seriously, day day life, lost loved ones, family loved ones, seeing friends family, situation makes feel, people flouting rules, corona virus situation, relaxed corona situation

Topic 2: worried family members, aren following rules, worried family friends, anxious worried family, feel anxious long, able work home, family members don, advice staying home, worried long situation, concerned mental health

Topic 3: feel sad people, feel angry people, high risk category, losing loved ones, feel quite anxious, sad people dying, don know long, anxious current situation, parents high risk, social distancing guidelines

Topic 4: underlying health conditions, having stay home, people lost lives, feel sad people, prime minister intensive, minister intensive care, sad people lost, family friends especially, people taking seriously, national health service

Unigrams (Male):

Topic 0: feel, people, time, situation, worried, work, life, home, family, able

Topic 1: global, response, anger, lead, using, said, disgusted, specifically, countries, relatively

Topic 2: people, feel, worried, situation, family, virus, health, going, work, home

Topic 3: people, feel, government, media, time, situation, lot, pandemic, virus, quite

Topic 4: stronger, toilet, paper, research, 000, recovery, safest, frowned, pushed, lasts

Trigrams (Male):

Topic 0: high risk category, seeing friends family, government handling situation, life return normal, able work home, miss seeing friends, people taking situation, following lockdown rules, things taken granted, country world economy

Topic 1: spend time family, ignoring social distancing, light end tunnel, people following guidelines, social distancing rules, people ignoring social, family friends worried, key worker work, ignoring government instructions, feel like people

Topic 2: things return normal, people taking seriously, feel sad people, worried contracting virus, mental physical health, feel like government, worried family getting, feel little worried, feel anxious future, worried corona situation

Topic 3: mental health issues, underlying health conditions, worried current situation, worried health family, pre existing conditions, think government doing, health family friends, feel government doing, number people dying, concerned death rate

Topic 4: social distancing measures, worry family friends, feel bit anxious, feel sorry people, angry people listening, observing social distancing, losing loved ones, worried elderly relatives, feel hopeful future, feel anxious economic