memes lol

Or,

Predicting Viral Spread Characteristics of Internet Memes from Textual Descriptions

Abstract

A model has been developed to fit the SIR epidemiological model to meme popularity data acquired from Google Trends. This enables finding spreading characteristics - new infections per person per day, average recovery time, and population size. This project aims to create a language model to predict these parameters from text descriptions of internet memes. Such a model would enable the estimation of the popularity, longevity, and general virality of memes using input - a description - which can be obtained trivially from any meme. In addition to learning about the ideal characteristics of a meme, the model would have applications in viral marketing and news.

SIR Fit Model

The basic idea behind the project is that internet memes, and ideas in general, spread through a population like viruses and can be modeled as being subject to natural selection. Thanks to this idea, the ideas developed in epidemeology can be applied to memes.
The previously developed model uses manually collected data from Google Trends as the ground truth. The data describes searches over time, giving an approximation of popularity. This data is then used to fit a model defined by the following system of differential equations:

$$\dot{S} = -\beta I S$$

$$\dot{I} = \beta I S - \frac{1}{d} I$$

$$\dot{R} = \frac{1}{d} I$$

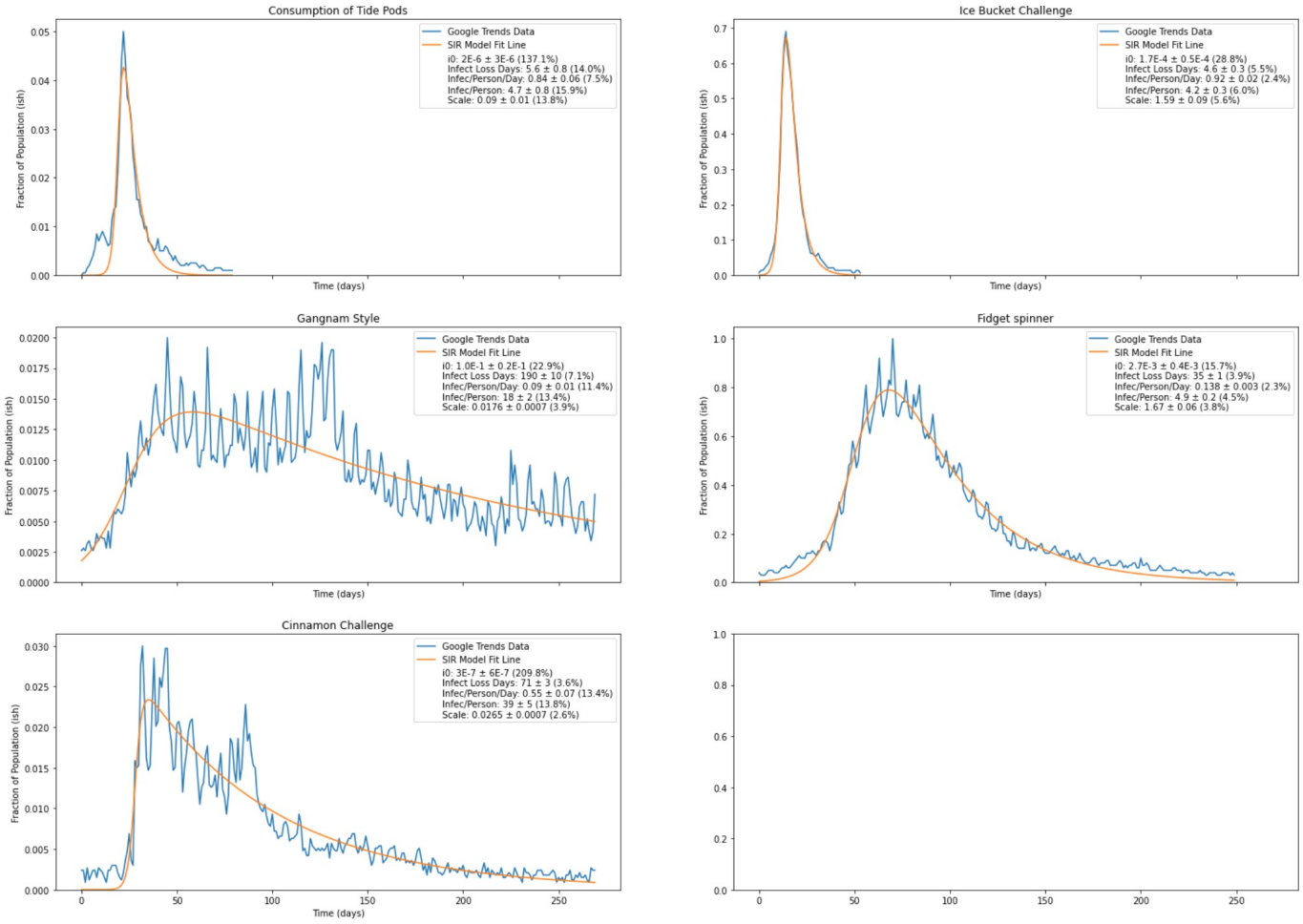
Where S is the fraction of the population which is susceptible, I fraction which is infected, R is the fraction which has recovered (and is now immune), β is the number of new infections per infected individual per day, and d is the recovery time of an individual infection.

To fit the data, all constants are chosen, then numerical integration is used to solve the initial value problem, finding $\hat{f}(t)$ which is some approximation (of unknown quality) for ground truth $f(t)$ describing searches per day over time as given by Google Trends. The cost of a given $\hat{f}$ is calculated via the following formula:

$$\int_{t_0}^{t_f} (\hat{f}(t) - f(t))^2 dt$$

Where t_0 and t_f are the start and end times of the Google Trends data.
The algorithm passes this cost function and parameters into `scipy.optimize.curve_fit`, which performs gradient descent to find the best-fitting parameters.
Reinfection was explored but determined to be negligible, and large-scale reinfection (secondary or repeat epidemics of similar scale to the original) were determined to be too difficult to predict, as they are often influenced significantly by external events.

This model gives the following results:



Finding Data

Currently, there are several options for finding textual data. These include human transcriptions from reddit, tags from knowyourmeme.com, or wikipedia/wikidata. The full dataset will be aquired through web scraping.
Google Trends data can be aquired automatically through the `pytrends` API.

Model Design and Pre/Postprocessing

This will be one of the greatest challenges. The data will need to be preprocessed - no model can take in raw text - and there are multiple viable options for this. The current plan is to do

significant research to find text vectorization solutions, but a starting point will be using `word2vec`.
The spread characteristic data will also have to be preprocessed. The scales of each parameter is different, which could cause problems with vanishing gradients. One current idea for a solution would be to scale the characteristics according to the output of PCA on the dataset of SIR parameters, but the optimality of this solution is unclear.
As for the model, further research is once again needed. A good starting point is a simple, fully-connected neural network, but this is likely far from ideal. Other possibilities are LSTMs or Transformers.

Division of Labor

The plan for the division of labor is to have Linus work on procuring output data - scraping google trends data, scaling it, fitting the SIR model to it, and collecting a dataset of those results - while I procure the output data, scraping knowyourmeme text data, vectorizing it, and collecting a dataset of those results. We will then work together to determine model architecture, and test/tune said architecture and hyperparameters on our own. We are both able to train models locally, so we will each test different things.

Appendix

Budget

As this is a purely software project, there will most likely be no spending neccessary.

Presentation Style

Traditional would be preferred.

Gantt Chart

See Gantt Chart here: https://docs.google.com/spreadsheets/d/1VRGk68cJCC-mdMbF9ECsIXnESskWSfuoc_APEnIDZTA/edit?usp=sharing

References

"The SIR Model for Spread of Disease." *Duke.edu*, 2000, https://services.math.duke.edu/education/ccp/materials/diffcalc/sir/sir1.html.