

TECH ALPHARETTA

2972 Webb Bridge Rd, Alpharetta, GA 30009

**Project # 13 : Tracking Disease Outbreaks Using Hews Headlines**

**Problem Statement**

Congratulations! You have just been hired by the American Institute of Health. The Institute monitors disease epidemics in both foreign and domestic lands. A critical component of the monitoring process involves analysis of published news data. Each day, the Institute receives hundreds of news headlines describing disease outbreaks in various locations. The news headlines are too numerous to be analyzed by hand.

Your first assignment is as follows; you will process the daily quota of news headlines and extract the locations mentioned within. You will then cluster the headlines based on their geographic distribution. Finally, you will review the largest clusters within the United States and outside of the United States. Any interesting findings should be reported to your immediate superior.

**Dataset Description**

The file 'headlines.txt' contains the hundreds of headlines that you must analyze. Each headline appears separately on an individual line within the file.

**Overview**

* In order to address the problem at hand we will need to know how to:
* Cluster datasets using multiple techniques and distance measures.
* Measure distances between locations on a spherical globe.
* Visualize locations on a map.
* Extract location coordinates from headline text.