INFM600: Assignment 1- Info Seeking

# [Dataset-1] Online News Popularity

This is a dataset of articles published by [Mashable](http://www.mashable.com/). It contains various details related to a news article like the title of the article, date on which it was published, the category that it belongs to (lifestyle, entertainment, technology, etc.)

## Source

UCI Machine Learning Repository: Data Set. (2015, May 31). *Online News Popularity Dataset.* Retrieved from [https://archive.ics.uci.edu/ml/datasets/Online News Popularity#](https://archive.ics.uci.edu/ml/datasets/Online%20News%20Popularity%23%20) on Sep 11, 2016 08:49:23 PM.

## Terms of Use / License

There is no explicit Terms of Use or License for this particular dataset. But it has been taken from University of California Irvine Machine Learning Repository and it is directed for use by students, educators and researchers all over the world as a primary source of machine learning data sets (refer: <https://archive.ics.uci.edu/ml/about.html>)

## Description

This dataset includes around 61 attributes about a news article. It not only includes trivial data like the title and the URL of the articles but also details like the rate of positive words, number of shares of the article, etc. This opens up a wide arena of questions which can be answered using this data. This is what makes the dataset interesting.

## Potential Data Users / Decision-makers

The potential data users are the media companies: be it print media or digital media. They can figure out the best ways to increase their clientage using this dataset.

## Prospective Queries

Following are the prospective queries that can be answered using this dataset:

* Do the readers prefer news from the entertainment category over technology category?
* Do readers share more of positive news or negative?
* What is the best day to publish a piece of news so that it becomes popular?

# [Dataset-2] Yelp Academic Dataset

This is dataset is basically published by the popular site Yelp in the form of a challenge- Yelp Dataset Challenge. It includes millions of reviews and tips from thousands of users for various businesses like restaurants, pet stores, etc. Moreover, information related to 10 different cities across 4 countries is covered by this dataset.

## Source

Yelp: Dataset (n.d.). *Yelp Dataset Challenge*. Retrieved from <https://www.yelp.com/dataset_challenge> on Sep 11, 2016 11:12:07 PM

## Terms of Use / License

As the dataset is available as a part of an academic challenge, the Terms of Use as well as License details are defined clearly [here](https://www.yelp.com/html/pdf/Dataset_Challenge_Academic_Dataset_Agreement.pdf).

## Description

Since this dataset contains data from various countries, it can be used to study the patterns of thinking of people in different countries, For example, we can figure out what factors people from different countries or even cities look for when they go out for a meal. This is what makes this dataset very interesting.

## Potential Data Users / Decision-makers

Yelp, business owners and the common public constitute the potential decision-makers/data users of this dataset as it can be used to figure out business models, marketing strategies or even sales strategies.

## Prospective Queries

Following are the prospective queries that can be answered using this dataset:

* Do people in Canada prefer bars to cafes?
* What is the busiest hour for pet stores?
* How much percentage of people give significance to the ambience when they go out to eat?

# [Dataset-3]

## Source

United States Department of Health and Human Services, Centers for Disease Control and Prevention and National Cancer Institute. (2016). *United States Cancer Statistics: 1999 - 2013 Incidence, WONDER Online Database*. Accessed at <http://wonder.cdc.gov/cancer-v2013.html> on Sep 12, 2016 10:35:32 AM.

## Terms of Use / License

CDC WONDER is a public service developed and operated by the Centers for Disease Control and Prevention, an agency of United States federal government. The public web site at http://wonder.cdc.gov is in the public domain, and only provides access to public use data and information. The information may be accessed freely, used, copied, distributed or published without additional or explicit permission, provided a proper citation is mentioned to credit the authors and/or data providers.

## Description

This website contains data regarding cancer incidence as well as cancer mortality for about 50 states in the United States, classified by age, year, state, metropolitan area, race, ethnicity, gender, etc. This website contains an online database wherein we can put our request and get the data in the form of tables, charts, maps and data extracts. Using this interactive interface, we can search for specific data easily without having to go through a huge magnitude of data. Moreover, this can used to study the cancer patterns around the country.

## Potential Data Users / Decision-makers

The potential data users or decision makers can be the general public of United States as well as researchers and medical staff (doctors, surgeons, nurses, etc.) The reason for including medical staff is that they can use this data and find out where the requirement of their skills is the maximum.

## Prospective Queries

Following are the prospective queries that can be answered using this dataset:

* What is the average cancer mortality rate in Puerto Rico?
* Has the cancer incidence rate in children increased or decreased over the last decade?
* Is liver cancer seen more in males than in females?