**Capstone Project ideas**

**1. Deepfake Detection**

* Problem: Deepfake techniques, which present realistic AI-generated video of people doing and saying fictional things, have the potential to have a significant impact on how people determine the legitimacy of information presented online. This is also a critical concentration in news outlets, who try to keep their news unbiased and eliminate the fake news. The main idea is to build a system which can detect the deepfakes and manipulated media.
* Data: from Kaggle (<https://www.kaggle.com/c/deepfake-detection-challenge/overview>)

**2. News feeds sentimental analysis**

* Problem: Different news outlets have their opinions toward a dedicated news event, but it’s hard for people to understand the whole and detail point of view for the special event, especially when there is huge amount of news feeds around us. I want to build a system which can analyze the sentiment of every piece of news articles and give a general evaluation toward a news event or toward special countries.
* Data: AG news(<https://registry.opendata.aws/fast-ai-nlp/>)

<https://nlp.stanford.edu/sentiment/>

**3. Stock price prediction**

* Problem: According to historical price ticks, people, particularly those who use technical analysis, can predict the future price. Also, there is great impact from social media to financial market. I want to setup a system using ML to combine the historical price fluctuation and social media (twitter) article to predict the future stock price.
* Data: need further investigation.

<https://www.quandl.com/data/EOD-End-of-Day-US-Stock-Prices/usage/quickstart/api>

* Note: I want to try something in this field, but I don’t know whether my idea is viable.

Dataset

<https://datasetsearch.research.google.com/search?query=News%20feed%20sentimental&docid=6srYA%2Bi3qZssY3ztAAAAAA%3D%3D>

### **Context**

This is the sentiment140 dataset. It contains 1,600,000 tweets extracted using the twitter api . The tweets have been annotated (0 = negative, 4 = positive) and they can be used to detect sentiment .

### **Content**

It contains the following 6 fields:

1. **target**: the polarity of the tweet (0 = negative, 2 = neutral, 4 = positive)
2. **ids**: The id of the tweet ( 2087)
3. **date**: the date of the tweet (Sat May 16 23:58:44 UTC 2009)
4. **flag**: The query (lyx). If there is no query, then this value is NO\_QUERY.
5. **user**: the user that tweeted (robotickilldozr)
6. **text**: the text of the tweet (Lyx is cool)

<http://archive.ics.uci.edu/ml/datasets/News+Popularity+in+Multiple+Social+Media+Platforms>

**Data Set Information:**

This is a large data set of news items and their respective social feedback on multiple platforms: Facebook, Google+ and LinkedIn.  
The collected data relates to a period of 8 months, between November 2015 and July 2016, accounting for about 100,000 news items on four different topics: economy, microsoft, obama and palestine.  
This data set is tailored for evaluative comparisons in predictive analytics tasks, although allowing for tasks in other research areas such as topic detection and tracking, sentiment analysis in short text, first story detection or news recommendation.  
  
Further details on the process of building the data set are provided in the article mentioned in the 'Relevant Papers' section.  
  
An .R file is provided to provide a simple introduction to handling the data set.

<http://archive.ics.uci.edu/ml/datasets/Reuters-21578+Text+Categorization+Collection>

**Data Set Information:**

From the original readme file (please consult it for more information):  
-------------------------  
The documents in the Reuters-21578 collection appeared on the Reuters newswire in 1987. The documents were assembled and indexed with categories by personnel from Reuters Ltd. (Sam Dobbins, Mike Topliss, Steve Weinstein) and Carnegie Group, Inc. (Peggy Andersen, Monica Cellio, Phil Hayes, Laura Knecht, Irene Nirenburg) in 1987.  
  
In 1990, the documents were made available by Reuters and CGI for research purposes to the Information Retrieval Laboratory (W. Bruce Croft, Director) of the Computer and Information Science Department at the University of Massachusetts at Amherst. Formatting of the documents and production of associated data files was done in 1990 by David D. Lewis and Stephen Harding at the Information Retrieval Laboratory.  
  
Further formatting and data file production was done in 1991 and 1992 by David D. Lewis and Peter Shoemaker at the Center for Information and Language Studies, University of Chicago. This version of the data was made available for anonymous FTP as "Reuters-22173, Distribution 1.0" in January 1993. From 1993 through 1996, Distribution 1.0 was hosted at a succession of FTP sites maintained by the Center for Intelligent Information Retrieval (W. Bruce Croft, Director) of the Computer Science Department at the University of Massachusetts at Amherst.  
  
At the ACM SIGIR '96 conference in August, 1996 a group of text categorization researchers discussed how published results on Reuters-22173 could be made more comparable across studies. It was decided that a new version of collection should be produced with less ambiguous formatting, and including documentation carefully spelling out standard methods of using the collection. The opportunity would also be used to correct a variety of typographical and other errors in the categorization and formatting of the collection.

<https://www.kaggle.com/therohk/global-news-week>

# Context

This dataset is a snapshot of most of the new news content published online over one week. It covers the 7 Day-period of August 24 through August 30 for the years 2017 and 2018.

Year 2017: **1,398,431** ; Year 2018: **1,912,872**

It includes approximately **3.3 million** articles, with **20,000 news sources** and **20+ languages**.

This dataset has just four fields (as per the [column metadata](https://www.kaggle.com/therohk/global-news-week/data)):

* **publish\_time** - earliest known time of the url appearing online in yyyyMMddHHmm format, IST timezone
* **feed\_code** - unique identifier for the publisher or domain
* **source\_url** - url of the article
* **headline\_text** - Headline of the article (UTF8, Any possible languages)

See the ["Basic Feed-Code Exploration"](https://www.kaggle.com/therohk/basic-feed-code-exploration) notebook for a quick look at the dataset contents.

# Inspiration

The sources include news feeds, news websites, government agencies, tech journals, company websites, blogs and wikipedia updates. The data has been collected by polling RSS feeds and by crawling other large news aggregators.

As of 2018, these 7-Day slices were selected as there wasn't any downtime or outage during the intervals. New news content is produced at this rate by publishers everyday, throughout the year.

<https://www.kaggle.com/aashita/nyt-comments>

### **Context**

New York Times has a wide audience and plays a prominent role in shaping people's opinion and outlook on current affairs and also in setting the tone of the public discourse, especially in the USA. The comment section in the articles is very active and it gives a glimpse of readers' take on the matters concerning the articles.

### **Content**

The data contains information about the comments made on the articles published in New York Times in Jan-May 2017 and Jan-April 2018. The month-wise data is given in two csv files - one each for the articles on which comments were made and for the comments themselves. The csv files for comments contain over 2 million comments in total with 34 features and those for articles contain 16 features about more than 9,000 articles.

### **Inspiration**

The data set is rich in information containing comments' texts, that are largely very well written, along with contextual information such as section/topic of the article, as well as features indicating how well the comment was received by the readers such as editorsSelection and recommendations. This data can serve the purpose of understanding and analyzing the public mood.  
The [exploratory kernel here](https://www.kaggle.com/aashita/nyt-comments-eda) can be used for a review of the features of the dataset and the [NB-Logistic model kernel](https://www.kaggle.com/aashita/starter-kernel-for-predicting-nyt-s-pick/log) for predicting NYT's pick can be used as a starter for building models on a range of ideas, some of which are:

1. Predicting the number of upvotes a comment will receive using the feature recommendations as the target variable. With enough training set for the model, we can make a guess of how a hypothetical comment on a certain topic will be received by the community of NYT readers' and this can be considered a tool to gauge public opinion. The design of this model will be very similar to the ones used in ranking the reviews based on guessing how many upvotes the reviews will receive.
2. Predicting whether a comment will be editor's pick using feature editorsSelection as the target variable. It gives a clue to what NYT considers worth promoting.
3. Based on a comment, guessing the topic (using sectionName and/or newDesk as the target variable) of the article.
4. Predicting how likely it is for a comment to get replies (using replyCount feature as the target variable).
5. Predicting how likely it is for an article to initiate discussion and get comments and upvotes as well as sentiment analysis of the comments' text.
6. Predicting the same as above for topics (indicated by the features sectionName and/or newDesk).
7. Analyzing behaviors of the top commenters such as which topics they most likely comment and the sentiment analysis of the comments.