**CAPSTONE 3 – PROJECT IDEAS**

1. **Autonomous tagging of StackOverflow Questions**

<https://www.kaggle.com/stackoverflow/stacksample>

### Description

#### **Problem**

Make a multi-label classification system that automatically assigns tags for questions posted on a forum such as StackOverflow or Quora. Dataset includes text of 10% of questions and answers from the Stack Overflow programming Q&A website.

#### **ML problem Category**

* Advanced NLP, Multi label text Classification problem, Predictive modelling

#### **Tasks**

#### Identifying tags from question text

#### Predicting whether questions will be upvoted, downvoted, or closed based on their text

#### Predicting how long questions will take to answer

#### **Data used**

This is organized as three tables:

* Total unique users – 468K
  + Questions contains the title, body, creation date, closed date (if applicable), score, and owner ID for all non-deleted Stack Overflow questions whose Id is a multiple of 10.
  + Answers contains the body, creation date, score, and owner ID for each of the answers to these questions. The ParentId column links back to the Questions table.
  + Tags contains the tags on each of these questions

1. **Yelp Review**

## <https://www.kaggle.com/yelp-dataset/yelp-dataset>

### Description

#### **Problem**

This dataset is a subset of Yelp's businesses, reviews, and user data. It was originally put together for the Yelp Dataset Challenge which is a chance for students to conduct research or analysis on Yelp's data and share their discoveries. In the dataset you'll find information about businesses across 11 metropolitan areas in four countries.

#### **ML problem Category**

* NLP, Text classification and Sentiment analysis

#### **Tasks**

* Yelp's reviews contain a lot of metadata that can be mined and used to infer meaning, business attributes, and sentiment.
  + What's in a review? Is it positive or negative? Is it fake?
* Explore Local Graph (i.e. Offer its users the ability to express their opinions on local businesses, and build social networks; users “friend” each other, build directed graphs based on

apparent shared preferences thus generating information about the influence patterns they present)

* + How do user's relationships define their usage patterns?
  + Where are the trend setters eating before it becomes popular?

#### **Data used**

In total, there are:

* 5,200,000 user reviews
* Information on 174,000 businesses
* The data spans 11 metropolitan areas
  + Note: The data is in json format and needs to be parsed in appropriate format before performing any DSM

1. **Google QUEST Q&A Labeling**

## <https://www.kaggle.com/c/google-quest-challenge>

### Description

#### **Problem**

Computers are really good at answering questions with single, verifiable answers ***whereas*** humans are better at addressing subjective questions that require a deeper, multidimensional understanding of context - something computers aren't trained to do well yet.

Questions can take many forms - some have multi-sentence elaborations, others may be simple curiosity or a fully developed problem. They can have multiple intents or seek advice and opinions.

The challenge is to use this new dataset to build predictive algorithms for different subjective aspects of question-answering.

#### **ML problem Category**

* Advanced NLP, Multi label Classification

#### **Tasks**

The data for this competition includes questions and answers from various StackExchange properties. Your task is to predict target values of 30 labels for each question-answer pair.

This is not a binary prediction challenge. Target labels are aggregated from multiple raters, and can have continuous values in the range [0,1]. Therefore, predictions must also be in that range.

#### **Data used**

Data files contain

* train.csv - the training data (target labels are the last 30 columns)
* test.csv - the test set (you must predict 30 labels for each test set row)
* sample\_submission.csv - a sample submission file in the correct format; column names are the 30 target labels
  + The list of 30 target labels is the same as the column names in the sample\_submission.csv file.
  + Target labels with the prefix **question\_** relate to the **question\_title** and/or **question\_body** features in the data.
  + Target labels with the prefix **answer\_** relate to the answer feature.