Capstone-2 Project Ideas

1. Women's E-Commerce Clothing Reviews

General :

This is a Women’s Clothing E-Commerce dataset revolving around the reviews written by customers. It’s nine supportive features offer a great environment to parse out the text through it’s multiple dimensions. Because this is real commercial data, it has been anonymized, and references to the company in the review text and body have been replaced with “retailer”.

Aim : Aim is to make predictions based on customer reviews using NLP.

Content :

This dataset includes 23486 rows and 10 feature variables. Each row corresponds to a customer review, and includes the variables:

•**Clothing ID:** Integer Categorical variable that refers to the specific piece being reviewed.
•**Age:** Positive Integer variable of the reviewers age.
•**Title:** String variable for the title of the review.
•**Review Text:** String variable for the review body.
•**Rating:** Positive Ordinal Integer variable for the product score granted by the customer from 1 Worst, to 5 Best.
•**Recommended IND:** Binary variable stating where the customer recommends the product where 1 is recommended, 0 is not recommended.
•**Positive Feedback Count:** Positive Integer documenting the number of other customers who found this review positive.
•**Division Name:** Categorical name of the product high level division.
•**Department Name:** Categorical name of the product department name.
•**Class Name:** Categorical name of the product class name.

https://www.kaggle.com/nicapotato/womens-ecommerce-clothing-reviews

2. 100K Coursera's Course Reviews Dataset

General :

100K+ Scraped Course Reviews from the Coursera Website (As of May 2017)

Aim : Aim is to make predictions based on reviews using NLP.

Content :

For a 5-star rating, the review was labelled as Very Positive, Positive for 4-star, Neutral for 3-star, Negative for 2-star, and Very Negative for 1-star. There are 2 files, **reviews.tsv** and **reviews_by_course.tsv**. The **reviews.tsv** file has no grouping, just the course reviews and their corresponding label. For the **reviews_by_course.tsv**, they are grouped by the CourseId column.

reviews.tsv

Id - The unique identifier for a review.
Review - The actual course review.
Label - The rating of the course review.

reviews_by_course.tsv

CourseId - The course tag. This is in the URL of the course in the Coursera website. For example, in this URL, **machine-learning** would be the course tag.
Review - A review in a specific course.
Label - The rating of the course review.

https://www.kaggle.com/septa97/100k-courseras-course-reviews-dataset

3. Airbus Ship Detection Challenge

General :

This is a Kaggle competition asking competitors to build models detecting ships in satellite images as quickly as possible.

Aim : Aim is to build a model that detects all ships in satellite images as quickly as possible.

Content :

We’re asked to locate ships in images, and put an aligned bounding box segment around the ships you locate. Many images do not contain ships, and those that do may contain multiple ships. Ships within and across images may differ in size (sometimes significantly) and be located in open sea, at docks, marinas, etc.

For this metric, object segments cannot overlap. There were a small percentage of images in both the Train and Test set that had slight overlap of object segments when ships were directly next to each other. Any segments overlaps were removed by setting them to background (i.e., non-ship) encoding. Therefore, some images have a ground truth may be an aligned bounding box with some pixels removed from an edge of the segment. These small adjustments will have a minimal impact on scoring, since the scoring evaluates over increasing overlap thresholds.

The `train_ship_segmentations.csv` file provides the ground truth (in run-length encoding format) for the training images. The `sample_submission` files contains the images in the `test` images.

https://www.kaggle.com/c/airbus-ship-detection