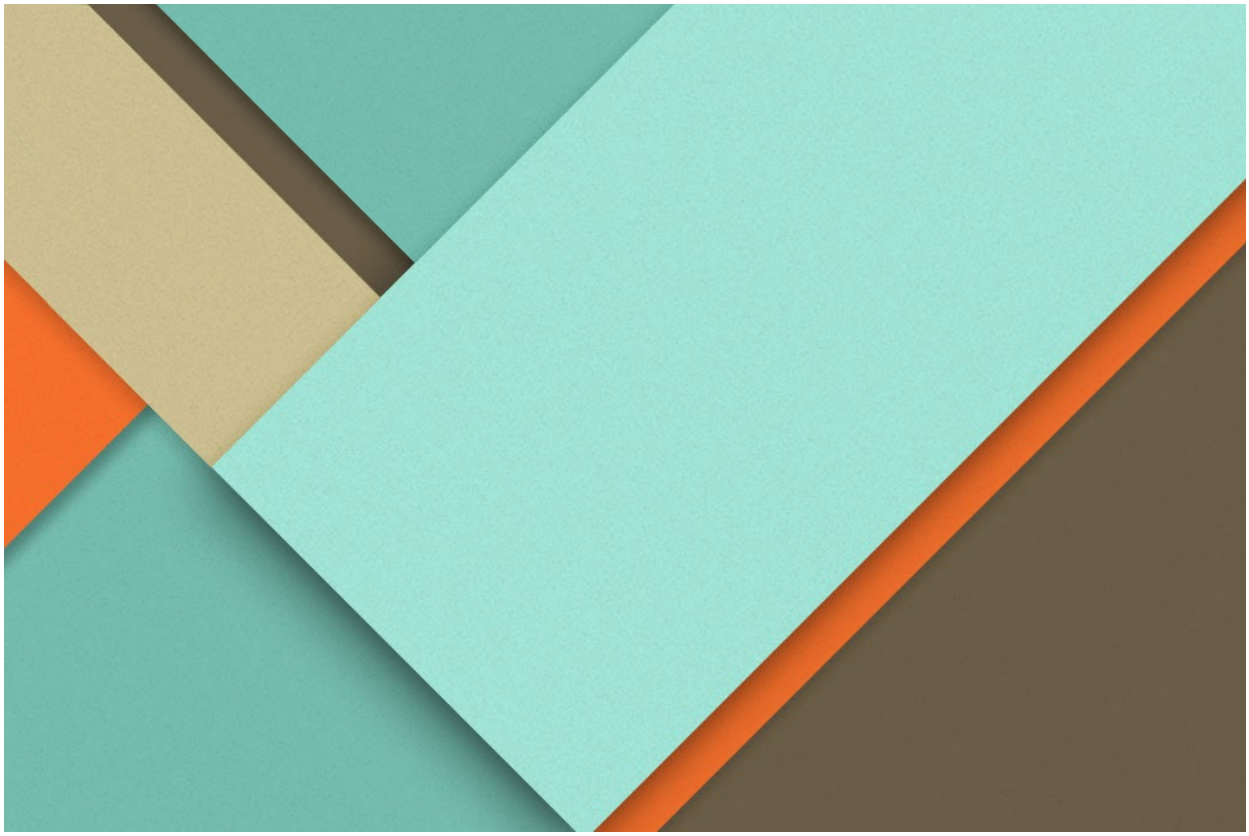

Capstone 2

Natural Language Processing Topics

Student - N.Swathi Priya **Mentor - Jatin Dixit**

Project Topics

1. Spambase Data Set

Classifying Email as Spam or Non-Spam.

Data Set Information: The "spam" concept is diverse: advertisements for products/websites, make money fast schemes, chain letters, pornography.

The collection of **spam e-mails** came from the postmaster and individuals who had filed spam. The collection of **non-spam e-mails** came from filed work and personal e-mails, and hence the word 'george' and the area code '650' are indicators of non-spam. These are useful when constructing a personalized spam filter. One would either have to blind such non-spam indicators or get a very wide collection of non-spam to generate a general purpose spam filter.

2. Real or Not? NLP with Disaster Tweets

Predict which Tweets are about real disasters and which ones are not.

In this project tweets are categorized in 2 classes: Tweets about disasters and tweets that are not about disasters. The data contains a Keyword, Location and the text around 10000 tweets. Natural Language Processing (NLP) techniques are used to make predictions with the text data. There are about 7000 Tweets in the train set and about 3000 tweets in the test set. The results are based on the F1 score as an evaluation metric.

3. KeyWord Extracrtion

Multi-label classification for Tag predictions.Identify keywords and tags from millions of questions

All of the data is in 2 files: Train and Test.

Train.csv contains 4 columns: Id,Title,Body,Tags

- Id - Unique identifier for each question
- Title - The question's title

- Body - The body of the question
- Tags - The tags associated with the question (all lowercase, should not contain tabs '\t' or ampersands '&')

4. Question-Answer Dataset

Finding the closest question and then answering it using NLP.

Being able to automatically answer questions accurately remains a difficult problem in natural language processing. The dataset has everything needed to try our own hand at this task. The project is to correctly generate the answer to questions given the Wikipedia article text where the question was originally generated from.

Data: There are three question files, one for each year of students: S08, S09, and S10, as well as 690,000 words worth of cleaned text from Wikipedia that was used to generate the questions.