Project Proposal

Team member's details:

- Group Name: IntiLP (Intelligent Language Processing)
- Group Members:
  - Disha Lamba
  - Kiarash Rastegar
  - Somasundaram Palaniappan
  - Mercy
- Email:
  - Disha: jb.dishalamba@gmail.com
  - Kiarash: krastegar0@gmail.com
  - Somu: plsaran97@gmail.com
  - Mercy: creativeart19@icloud.com
- Country:
  - Disha: New York
  - Kiarash: California
  - Somu: England
  - Mercy: London
- College/Company:
  - Disha: MS in Computer Science and Engineering @ New York University
  - Kiarash: MS in Bioinformatics @ San Diego State University
  - Somu: MS in AI and Robotics @ University of Hertfordshire
  - Mercy: MSc in Finance @ UCL
- Specialization:
  - NLP

Problem description:

With the ongoing struggle for combating cancer, it has become increasingly important to find underlying causes that may be related to the onset of the disease. Researchers in the biomedical field are constantly looking for new biological markers that may be incorporated into their cancer detection assays. Here we define a biological marker as either a gene or protein that has been observed in that specific disease or cancer. One major issue is that scientists must read through large amounts of articles to find new candidates of biological markers for their cancer assays. Here we propose an approach to use a deep learning NLP model that may be able to extract important information from biomedical literature and return a list of biological markers related to that disease. Here we will focus on 3 major types of cancer: lung, pancreatic, colon.

Github Repo Link:

dldisha/IntiLP (github.com)

EDA Performed on Data:

- After EDA analysis we know cancer, lung, and tumor are the most used words.
- Non-small cell lung cancer(NSCLC) is the most detected cancer
- Oncogene is the most common gene.

Final Recommendation:

- After our EDA analysis, we can say that there are a few keywords that will be most important for the classification.
- Cancer, tumor, lung, NSCLC, and Oncogene can be some of the words to focus on.