Project Proposal 297 -1 (Natural Language Processing )

By: Alavi Khan, Anoushka Gade, Shrinidhi Kota Shreeshapuranik

Problem Statement:

We attempt to create a web-based solution for the public to make queries about the defense contracts of the United States of America and provide them closest answer.

Motivation:

We think this project will enable us to learn about the latest techniques used for question-answering on unorganized data. Our project will provide a quick way to gather information about the army's available contracts.

Dataset:
We will use the defense contracts dataset (https://www.defense.gov/News/Contracts/) having news and updates about the contract acquisitions for the US Army, Navy, and Air force. The contracts' information is updated daily, so we intend to scrape the website for the last four months, aggregating it into data set of 100+ documents.
We will use python and libraries like beautiful soup, requests, and urllib3 to scrape, parse and structure the contracts.

Experimental Plan:
We will start with cleaning the data and tokenizing the dataset to create tokens. Further, annotation of data if required. We will transfer processed data to a Natural language processing model. The model will take the user query and provide the closest possible answer. Finally, we will develop a web application that provides a user interface.

Anticipated challenges:

Data is raw.

A lot of numerical data is present with relevance.

We might need domain knowledge to understand defense terminologies.

Scraping the website is a challenging task.

Real-time data feed consumption is a real challenge.