**Scope**

This projects objective is to analyze the different properties of Reddit posts and build a model that predicts the score or upvote percentage of a post.

**Overview**

This document describes the basic data project pipeline. What tools, libraries, data sources are going to be used in each step.

**General flow**

1. Extract data
2. Exploratory data analysis
3. Feature engineering - vectorization
4. Partition the data to train, validation and test sets
5. Build a model
6. Train the model
7. Test and evaluate the model
8. Improve the model by repeating steps 3, 5-7

**Software Architecture - components**

1. Data extraction:

* Pushshift REST api to search and extract Reddit submissions  
  https://reddit-api.readthedocs.io/en/latest/
* Complete missing data using PRAW python library to extract most current Reddit data https://praw.readthedocs.io/en/latest/

1. Exploratory data analysis: analyze extracted data  
   Python libraries:

* numpy, scipy, pandas for data manipulation and statistical analysis
* matplotlib, seaborn for data visualization
* nltk (maybe others) for natural language processing

1. Feature engineering: Use data and analysis inferences to build a feature vector.  
   Python libraries:
   * numpy, scipy, pandas for data manipulation
   * nltk (maybe others) for natural language processing
   * input: data extracted from Reddit
   * output: a vector of features for each submission
2. Model:

* input: feature vector
* output: predict score / upvote ratio
* The model will be built with PyTorch machine learning framework