STATEMENT OF WORK

NATURE OF PROJECT:	ML Challenge
PROJECT REFERENCE NAME:	*Analysis of comments on Glassdoor*

Section 1.

Description of Services to be completed by:

1. Project: ***Analysis of comments on Glassdoor***

   Glassdoor – Companies Feedback Analysis (Advantages, Disadvantages) (Web Scraping, NLP analysis, English and Spanish)

   Human resource needs

   ● Web Scraper Specialist
   ● Python developer
   ● Data Scientist
   ● MLOps

   The project should be completed in several stages:

   - **Stage 1** Will create a pipeline to extract Information of website Glassdoor
     - **Web Scraping**

       This endpoint will receive as input Web scrapping:

       - Open web page
       - Read the content
       - Extract content by globant
       - Save all in a data frame

   - **Stage 2** Will create a model to classifier and make a sentiment analysis

- **Text Preprocessing**

  This endpoint will receive as input Web scrapping result in a data frame

  - Create English-Spanish data frame
  - Analyze Separately Spanish and English data frame (Corpus)
  - Data cleaning
  - Stop words.
  - Lemmatization
  - N-grmas Distributions

- **Classification proposed**

  This endpoint will receive as input parameters of text preprocessing:

  - Construction model
  - End Date (date time)
  - Taring model
  - Calculation of grammatical probabilities

- **Extraction of main features**

  - Classification
  - Sentimental Analysis (pysentimiento vs vader)

- **Stage 3 Create a pipeline to MLOps**

- The MLOps part will be done with mlflow performing the following tasks

  - Log metrics
  - Model signatures
  - Save the plot and log it as an artifact
  - Tracking url (localhost)
  - Run MLOps

Section 2.

Other Notes and Considerations:

- Use a personal account per team to connect to git hub
- Best Practices matters

- Code should be continuously uploaded to a list of users will be provided to be added as collaborators to the repo.
- Validate input parameters (Web scraping and data frame) of all requests
- Exception handling is excepted

Section 3.

Documentation

Expected documentation should be in English and will include:

- Web Scraping
- Model Construction
- MLOps
- Instructions on how to pull the code from github, compile, run the solution and post a request. These instructions are targeted to technical.

Extra documentation (nice to have):

- Architecture Diagrams