Project Name: Proposal for automatic data collection and prediction

Objectives: The objective of this document is to show how https://isthisarealjob.com can solve the problem of manual data collection and better improve the site functionality and usability by using an automatic data collection process and using an ML algorithm to better classify job adverts and interview invite.

Executive Summary

This proposal is focused on how isitarealjob.com can use different approaches to transition from manual data manipulation by improving their data collection and manipulation process to populate their database and how this data can be used to make future predictions for unseen data based on specific information and keywords. This processes can be handled by data and software engineers along side Machine learning Engineers.

Data collection and Transformation

Extracting the Data.

To extract data, Isitarealjob.com will first have to get a data source. This data source can be a social network site, job advertisement sites, or by carrying out a user survey.

There are three ways data can be gotten from the data source

1. Write a script to scrape job adverts from various job promotion site and company news site as well as on social network e.g Facebook and twitter.

2. Write another script that scrapes jobs interview invitations and reviews about those interviews from sites like nairaland.

3. Another method is that the company can send out a survey form containing the following details.

    Company Name
    Interview Address Location
    Interview Invitation
    Real or Fake

Transform the Data

In this stage, the data collected from the various sources are transformed into usable formats and structured.

1. Write a script that automatically parses the data and fix them in the appropriate columns in a spreadsheet. This script can be scheduled to run at a specific time in a day or week or month.

2. The arranged data is then cleaned and processed.

- The cleaning process involves removing duplicates.
- Null columns are taken care of.
- Missing values are filled with NaN
- Data are categorised according to data types(int, DateTime, string. Etc).

Loading the Data.

The transformed and structured data are loaded into the database and can now be used for analysis and making predictions.

1. The arranged and structured data will then be directly inserted in the database.

2. This process can be done with a script that usually runs based on a schedule.

Once all this process is complete isitarealjob.com will be left with a clean database that is constantly updated as scheduled. This data can then be used to improve the way predictions are given to users.

Classification of Job Adverts

The structure and clean data loaded in the database can then be used to perform classification. The following process can be followed to achieve that.

1. The data will be loaded.

2. Since this is an unsupervised problem(the data doesn't have a column that identifies an advert or interview invitation as real or fake).

3. The data will then be preprocessed so we can work with it on a machine learning model.

- Preprocessing steps like removing stopwords and use of contractions.

- Vectorization and tokenization of words, turning them into integers.

4. Analysis and visualization will be carried out on the data to find hidden information about the data.

- Carry out statistical analysis, clustering and exploratory data analysis on the data.

5. A sentiment analysis model will then be built based on selected explanatory features of our dataset.

- Features like address, invite format, medium sent, and job advert review and invitation review will be used.

6. The model will then be tested for deployment.

- Testing by using flask API, Heroku and docker