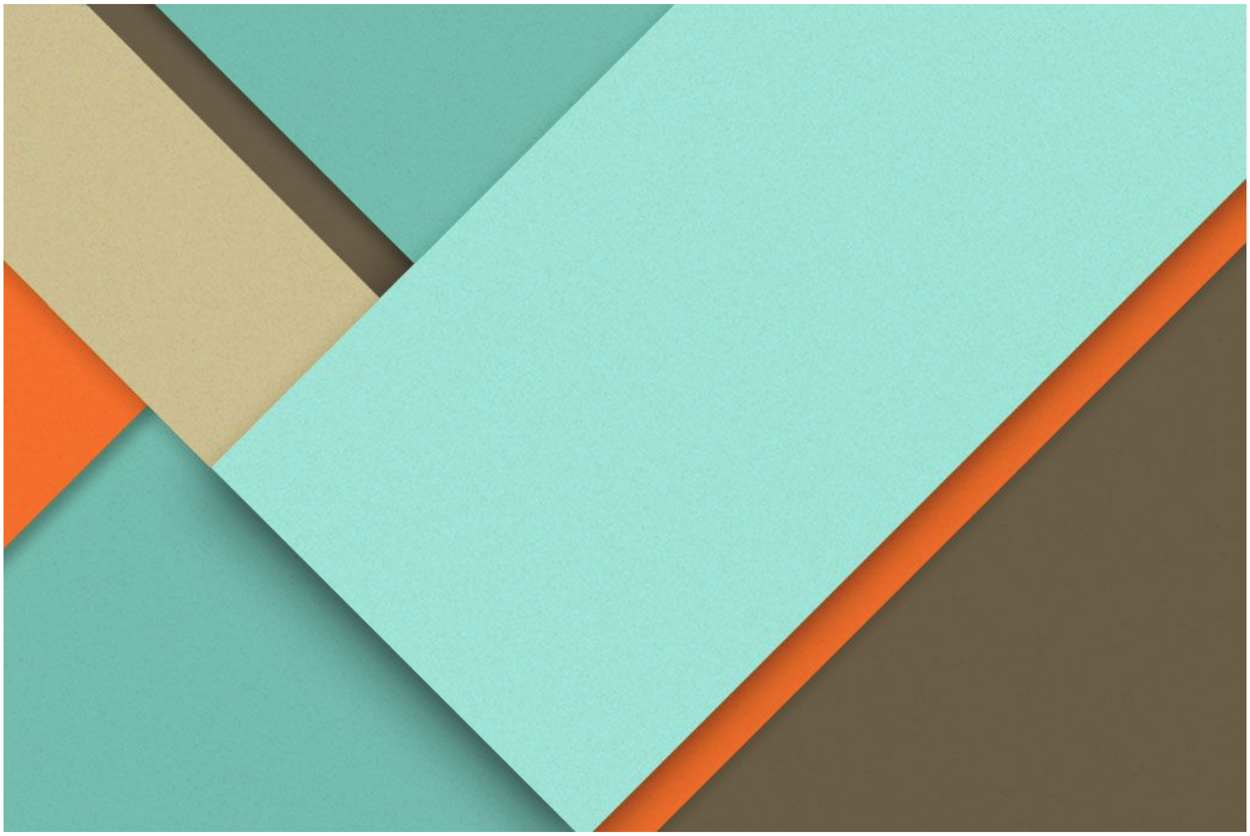

Machine learning Workflow to improve ISTHISAREALJOB.COM

19.10.2019

Oladokun Pelumi

HNG6 Machine learning Intern

(Slack username: @pucado)

Overview

This document contains the workflow to improve isthisarealjob.com. Isthisarealjob.com is a website used to predict if a job is real or just a scam. Currently the website does not use any machine learning technology and the database for both real and fake jobs is updated manually. This project proposes a new method for job seekers to ascertain whether the job invite received is real or fake.

New Proposed Method

Currently the website relies on its database being updated manually and this is not a very efficient way to manage the database. I propose a new method in which data is collected on a per use basis i.e the database is updated by the information given by each job seeker who uses the website to confirm if the job invite received is real or a scam.

The User is first asked to paste the full invite received and in the format received i.e User should not make any changes to invite and paste as received. A Natural language processing model then runs inference on this invite and matches it with the database containing a list of fake invites to find similarities (the database is also updated with the response). The model also checks for basic grammatical errors as that is a common indicator that a job invite is a scam or not. Using NLP also, the model obtains the name of organization and address and matches them out via the Google Places API. If the check matches out, there is a good chance that the invite is real and if it doesn't check out there is a good chance that the invite was a fake but for a more robust approach to the problem, this is not the only check to confirm if invite is real or not.

Second, the website asks the user if he/she applied for the job. There are three options in this phase; yes, no or maybe. Yes means the User applied for the job , no means the user did not apply for the job and in this case if company name and address also does not match google maps record there is a very good chance that the job is a scam. Maybe indicates the user isn't sure if he\she applied for the role, perhaps a relative sent an invite on their behalf. Finally, the user is asked for the source of invite, options include via social media, email, text, phone call and then gives a prediction based on the information given.

NLP Model

The natural language processing model will be built using the Spacy library on python. Spacy has built in features that can be used to analyse the text inputted by the user to obtain the address of the Organization and the name of the organization. First spacy is imported and the English tokenizer, tagger, parser and word vectors are imported. The text inputted by the User is then parsed through the loaded tokenizer. The syntax is then analyzed to obtain the entities present in the text. The entities of concern being the Name of Organization and address of Organization. The entities obtained are then parsed to the Google Places API to confirm if the name and address of the organization checks out. The fields parsed to the API is the address of the establishment to return the name of the establishment which is then matched with the name of organisation obtained from the user input text.

An extension of the Spacy library Hunspell is also used on the text input by the user. Hunspell is used to check for grammatical errors in the message received by the user. The input text is first tokenized into respective words and is parsed through the spell checker to check for errors in spelling words. Next the grammatical context of the whole text is considered.

The results obtained from the natural language processing of the inputted invite from the user can to a certain extent make a prediction on whether or not the job invite was a scam or not. If name of organization matches the name of establishment obtained from the Google Places API, there is a good chance that job is real if otherwise we run other checks to ascertain. We also use information about whether or not user applied for job role as some of the time invites are received for jobs not applied for and information on source of invite is also important in making a prediction on whether or not invite is real or not. If source of invite was a text message, address and name of organization did not check out and errors were found in text then system can predict that the invite is a scam.

Improved Data Collection

As more and more users use the new system to confirm job invites, the database of the website is updated. The features obtained being; job title, address of company, name of company, state and city, date of interview, source of invite. The features can be obtained from the use of SpaCy to process the invite. This new database matched with whether or not job was a scam or not can then be trained to build additional machine learning models to make predictions.