Innov8 2.0 hackathon

Team Zeal

# Analysis of data provided

## 1. Absence of Labelled Data

- **Challenge of Unsupervised Context**: The unavailability of labelled data, such as predefined categories of fraudulent and genuine applicants, eliminates the feasibility of employing supervised learning techniques. In supervised learning, models are trained on annotated data to recognize patterns and predict outcomes. Without this, reinforcement learning (RL) or unsupervised methods become essential.
- **Need for Reinforcement Learning**: RL is particularly suited for scenarios where an agent learns optimal actions through interaction with an environment to maximize cumulative rewards. In this context, an RL-based model can be designed to learn from feedback or rewards associated with correctly identifying fraudulent vs. genuine applications. The model iteratively improves by adjusting its policy based on rewards, even in the absence of direct labels.

## 2. Irregular Data Formatting

- **Incompatibility with Standard Parsing Tools**: The resumes and recommendation texts provided lack a consistent structure. This inconsistency means that typical parsing tools, which rely on predefined templates (e.g., reading CSV files with specific columns), cannot be utilized.
- **Natural Language Processing (NLP) Requirements**: Instead of traditional parsing, more sophisticated NLP techniques must be employed to extract meaningful information. These include methods for sentence segmentation, named entity recognition (NER), and syntactic parsing to decipher information like work experience, skills, and endorsements from unstructured text.

## 3. Variability Among Applicants

- **Skill Sets and Experience Levels**: Applicants differ in their professional expertise, as evidenced by the diverse range of skills listed and the variation in years or types of work experience. These variations can be quantified and analyzed to detect anomalies or inconsistencies in the information presented.
- **Recommendations as a Differentiator**: The nature and specificity of recommendations received by applicants are significant indicators. A pattern of vague or overly positive endorsements, especially when not substantiated by the applicant's credentials, could signify an attempt at manipulation.

## 4. Irrelevance of Institutional and Organizational Information

- **Obfuscation of Details**: The dataset appears to have deliberate ambiguities in the names of educational institutions and organizations, making it impossible to use these attributes reliably. This could have been introduced to anonymize data or protect privacy.
- **Exclusion from Analysis**: Since these details do not contribute to distinguishing between legitimate and fraudulent entries, they are excluded from the primary analysis, focusing instead on skills, experience, and endorsement content.

# Conceptual Framework

## 1. Detection of Identity Manipulation

- **Impact of Manipulation on Language**: Manipulation or falsification of identity, whether by the applicant or the recommender, generally results in language that is either overly vague or suspiciously precise. For instance, fabricated information often lacks the subtlety and depth found in genuine entries.
- **Language as a Signal**: The model aims to capture these nuances by evaluating the linguistic features of the text. Unusual patterns, such as excessive use of technical jargon or overly generic language, can serve as red flags.

## 2. Quantifying Vagueness

- **Role of TF-IDF in Measuring Vagueness**: The TF-IDF (Term Frequency-Inverse Document Frequency) vectorizer is used to gauge how specific or vague a piece of text is. It calculates the importance of words in a document relative to their frequency across a broader corpus.
- **Interpretation of Scores**:
    - A higher TF-IDF score indicates that a word is relatively rare in general language but prevalent in the given text, suggesting the use of domain-specific terminology.
    - Conversely, a lower score indicates common, non-specific language usage. The overall score for a document helps determine whether the language used is contextually rich or intentionally vague.

## 3. Identification of Reciprocal Endorsements

- **Pattern Recognition**: Reciprocal endorsements—where two candidates endorse each other—are identified using a brute force approach. This method involves scanning the dataset for pairs of applicants who have exchanged endorsements, which could indicate collusion.
- **Algorithmic Detection**: A simple nested loop algorithm is implemented to cross-check endorsements. This approach compares every applicant-recommender pair, flagging suspicious reciprocal patterns for further scrutiny.

# Methodology

## 1. Data Extraction

- **Use of Pyresparser Framework**:
    - Pyresparser is an open-source tool built on top of popular NLP libraries like spaCy, designed specifically for extracting structured data from resumes.

- It uses the PyPDF library to read PDF files, extracting textual content which is then processed by spaCy to identify relevant sections like experience, education, and skills.

- **spaCy for Context Extraction**:
  - spaCy provides advanced NLP capabilities, such as part-of-speech tagging, named entity recognition, and dependency parsing. It helps in segmenting the resume into different components and understanding the context around each skill or experience mentioned.
  - By analysing linguistic features like nouns, verbs, and their dependencies, the tool categorizes the content under labels like summary, experience, skills, education, etc.

# 2. Language Vagueness Detection

- **Implementation of TF-IDF Vectorization**:
  - TF-IDF vectorization involves two main components:
    - **Term Frequency (TF)**: How frequently a term appears in a document.
    - **Inverse Document Frequency (IDF)**: How rare the term is across all documents. This penalizes common words and highlights terms that are unique to the document in question.
  - The product of these values results in the TF-IDF score for each term, allowing the system to prioritize contextually significant words over generic ones.
  - **Text Scoring**: By averaging the TF-IDF scores of all words in a document, the overall precision of the language is quantified. The score ranges from 0 to 20, where a higher score indicates a higher level of specificity in the text.

# 3. Detection of Reciprocal Endorsements

- **Brute Force Algorithm**:
  - The algorithm checks each applicant-recommender pair for mutual endorsements. If both individuals endorse each other, the pair is flagged as suspicious.
  - This method, while computationally intensive, is straightforward and effective in small datasets where sophisticated graph-based or machine learning algorithms may not be necessary.

# 4. Scoring Mechanism

- **Formula for Composite Score**:

The final score for each applicant is calculated using the following formula:

$$\text{Final Score} = (\text{Resume Score} \times 10) + (\text{Cycles} \times \text{Recommender Score})$$

Where:

- **Resume Score**: Reflects the precision and quality of the candidate's resume, as determined by the TF-IDF analysis.
- **Recommender Score**: Evaluated based on the credibility and specificity of endorsements. Suspicious patterns reduce this score.
- **Cycles Parameter**: Takes a value of 1 for non-suspicious candidates and -1 for those flagged due to reciprocal endorsements. This parameter adjusts the weight of

the recommender score based on the perceived trustworthiness of the recommendations.

- **Rationale**:
  - The formula gives more importance to the resume score, but penalizes applicants heavily if they are suspected of engaging in reciprocal endorsements. This approach balances the evaluation of self-presented data (resume) with external validation (recommendations).

## 5. Recommender Network Analysis

- **Network Connectivity**:
  - The network of endorsers associated with an applicant is analysed to determine the breadth and depth of their professional connections.
  - A legitimate candidate is expected to have a diverse network, while suspicious patterns might indicate tightly-knit groups of mutual endorsements.

## 6. Ranking system

- The scores are arranged in descending order and hence in descending order of performance
- The fraudulent claims are flagged by using a threshold value of score generated . this threshold is chosen between the highest score among the entries present in suspicious and the next closest greater score, which ensures that the trivial anomalies are captured.
- Upon obtaining the division among the entries, it was found that the last 96 entries in the sorted dataset needed to be flagged (in python notebook) to provide an optimally tight bound.

# Proposed Improvements

## 1. Assessment of Endorser-Endorsee Collaboration

- **Clustering and Feature Extraction**:
  - By applying clustering algorithms like K-means or hierarchical clustering to features extracted from resumes, groups of similar applicants can be identified.
  - The proximity of an applicant to their endorser(s) within this feature space can indicate the likelihood of collusion. For example, if an applicant and their endorser have almost identical career trajectories or skill sets, it may raise a red flag.
- **Proximity Analysis**:
  - Calculating distances between data points in the feature space helps to understand the closeness of collaborations. If applicants are clustered too closely with their endorsers, this could indicate a potential conflict of interest or fraudulent behavior.

## 2. Level of Collaboration Analysis

- **Network and Path Analysis**:

- The network of an applicant is compared against those of other applicants within their cluster. This can help identify abnormal patterns of collaboration or endorsement that might not be apparent from individual resumes alone.
- Tools like network graph analysis can visualize connections and identify tightly-knit subgroups that deviate from typical collaboration networks.

# 3. Cross-Validation of Skills

- **Skill Mapping to Vector Space**:
  - Skills listed in the resumes are mapped to a vector space using techniques like Word2Vec or BERT embeddings. These embeddings capture the semantic similarity between different skills.
  - **Inner Product Analysis**: By calculating the inner products (or cosine similarity) between skill vectors, the alignment of an applicant's skills with those of the cluster can be quantified. Significant deviations might suggest exaggeration or fabrication of skills.
- **Qualitative Analysis**:
  - The model can use qualitative thresholds to compare an applicant's skill set against the central tendency of their cluster. Skills that are common in the cluster but absent in the applicant’s resume might indicate a gap, while unique skills might be scrutinized for authenticity.

    These refinements aim to strengthen the fraud detection framework, providing a more granular and nuanced understanding of applicant data and their professional networks.