**Analyzing Real-world Job Data with Big Data Analytics**

**Introduction:**

Recruitment is a critical function for any company, and the Human Resources (HR) department plays a pivotal role in finding and hiring the right talent. With the increasing demand for Big Data professionals, it becomes essential to optimize the recruitment process for Big Data job roles. This project aims to develop a Big Data solution that leverages vast amounts of data gathered from real-world job posts published online to enhance the HR department's ability to identify, categorize, and assess candidates for Big Data positions.

**Objectives:**

The primary objectives of this project are as follows:

**Identify Big Data Job Families:** Analyze the given dataset to identify various Big Data job families based on the job titles extracted from real-world job posts.

**Group Big Data Skills:** Identify homogeneous groups of Big Data skills highly valued by companies, allowing HR to focus on key competencies during the recruitment process.

**Characterize Job Families:** Characterize each Big Data job family according to the level of competence required for each skill set, providing a comprehensive understanding of the skills needed for different roles.

**Optimize Recruitment:** Provide valuable insights to the HR department for more effective and targeted recruitment efforts for Big Data job roles.

**Related Work**

**Big Data Skill Analysis:**

Prior research has explored various aspects of Big Data skills and job roles. Traditional approaches involve keyword analysis and clustering techniques to identify commonalities among job titles and required skills. However, with advancements in Natural Language Processing (NLP) and machine learning, more sophisticated methods, such as BERTopic and TF-IDF analysis, have emerged.

**Clustering Techniques:**

Clustering techniques, including K-means and BERTopic, have proven effective in grouping similar job roles based on their descriptions. These methods provide a nuanced understanding of the skills associated with each cluster, aiding in the categorization and characterization of job families.

Moreover, clustering techniques not only assist in grouping similar job roles but also contribute to the identification of emerging patterns and trends within the Big Data job market. K-means, a traditional centroid-based clustering method, partitions job roles into clusters based on their similarity, offering a straightforward categorization. On the other hand, BERTopic, leveraging state-of-the-art language models like BERT (Bidirectional Encoder Representations from Transformers), provides a more advanced approach by considering the semantic meaning encoded in job descriptions. This not only enhances the accuracy of clustering but also captures subtle contextual nuances that may be missed by traditional methods. As a result, HR departments gain a richer understanding of the evolving landscape of Big Data roles, discerning not only the prevalent job families but also the intricate variations and specialties within each cluster. The nuanced insights derived from these clustering techniques empower HR professionals to tailor their recruitment strategies to the specific skill sets and expertise demanded by distinct segments of the Big Data job market, fostering a more targeted and effective recruitment process.

**TF-IDF Analysis**

Term Frequency-Inverse Document Frequency (TF-IDF) analysis is a widely used method for extracting meaningful information from text data. It helps convert textual information into numerical data, enabling the identification of key terms and their importance in a given context.

Additionally, TF-IDF analysis offers a quantitative measure of term importance by considering both the frequency of a term within a document (term frequency) and its rarity across the entire dataset (inverse document frequency). This dual consideration provides a more nuanced assessment of the significance of terms within the context of Big Data job roles. The method involves assigning a weight to each term, with higher weights indicating greater importance. Consequently, HR professionals gain not only a list of relevant terms but also an understanding of their hierarchical importance. By transforming textual information into a numerical representation, TF-IDF facilitates the identification of not only the most common skills but also the distinctive skills that set certain job roles apart. This depth of analysis enables HR departments to prioritize skills based on their actual impact, ensuring that recruitment efforts are tailored to the specific competencies crucial for success in Big Data positions. TF-IDF, as a versatile and powerful tool, adds a layer of precision to skill analysis, supporting HR in making informed decisions for strategic and effective recruitment.

**Methodology:**

**Data Preprocessing**:

The project begins with data preprocessing, including the removal of stop words, tokenization, and the extraction of relevant features such as job titles and skills from the dataset.

Moreover, data preprocessing involves ensuring the quality and consistency of the dataset. This includes handling missing values in job titles and skills, employing techniques such as imputation or exclusion based on the extent of missing data. Data cleaning steps also address potential inconsistencies or variations in the representation of job titles, ensuring uniformity in the subsequent analysis. Data preprocessing acts as the foundational step, ensuring that the subsequent analyses are conducted on a robust and standardized dataset, thereby enhancing the reliability and interpretability of the findings.

**Job Family Identification(N Grams & K Means clustering)**

Utilizing n-grams and frequency analysis, job titles are grouped into families, providing a broad overview of the different Big Data roles present in the dataset.

**Clustering**

K-means clustering is applied to categorize job roles into distinct clusters, each representing a specific job family. The most common job name within each cluster is identified to characterize the cluster.

**Homogeneous Skill Analysis(TF-IDF Vectorization)**

TF-IDF analysis is employed to convert skills into numerical data. The total TF-IDF values across all documents for each skill are calculated, providing insights into the significance of each skill in the dataset.

**Cluster Characterization(BERT Topic):**

We have used BERT Topic a topic modeling technique, which uses BERT (Bidirectional Encoder Representations from Transformers) to perform topic modeling. It automatically clusters and identifies prevalent job role patterns within our dataset. This model does not need any pre-processing of data like the removal of stop words, punctuations etc., as it comes with the pre-trained word embeddings. It captures the semantic meaning of a word in the sentence rather than relying on the bag of words.

**Results and Discussions**

**Job Family Identification**

The n-gram analysis successfully identified various Big Data job families based on the dataset. The clustering results, visualized through bar charts, showcase the distribution of job roles across different clusters, each associated with a common job name.

A graph with blue squares

Description automatically generated

**Skill Analysis(HOMOGENEOUS):**

TF-IDF analysis revealed the importance of different skills within the dataset. The obtained numerical representation of skills facilitates the identification of key competencies highly valued by companies seeking Big Data professionals.

A screenshot of a computer

Description automatically generated

**Cluster Characterization:**

The project successfully characterized each job family based on the competence level required for different skill sets using BERT Topic methodology. This provides HR with a detailed understanding of the skills associated with each job family, aiding in targeted recruitment efforts.

A computer screen with white text

Description automatically generated

A graph of data architect

Description automatically generatedA graph of data engineering

Description automatically generated

**Conclusion:**

In conclusion, the developed Big Data project offers a comprehensive solution for HR departments to enhance their recruitment processes for Big Data positions. The combination of n-gram analysis, clustering techniques, and TF-IDF analysis provides a holistic view of job families, skills, and competence levels, ultimately optimizing the recruitment process.