**Comprehensive Text Mining Report on Data Engineering Roles in the UK**

**1. Introduction**

Data Engineering has emerged as a pivotal discipline in today’s data-driven landscape. Organizations rely on Data Engineers to design pipelines, ensure data quality, and optimize cloud infrastructure. To uncover the most sought-after skills, tools, and qualifications for Data Engineering roles in the UK, we conducted a text mining project on job descriptions scraped from LinkedIn. This report outlines the methodology—from data cleaning through advanced analytics methods—and highlights insights derived from clustering and **Social Network Analysis (SNA)**. Equipped with these findings, job seekers can better tailor their skill sets, and hiring managers can gain clarity on the evolving demands of the market.

**2. Methodology**

Our text mining pipeline consisted of four primary stages:

**2.1 Data Cleaning**

1. **Dataset Collection**
   * We initially scraped **1,000** Data Engineering job postings from LinkedIn, specifically targeting roles based in the UK.
   * After removing duplicates, incomplete listings, and near-identical entries, we consolidated the dataset to **644** unique records.
2. **Text Pruning**
   * Non-relevant information—such as company intros, office locations, and benefits—was excised to isolate skill requirements.
   * We removed recruitment-domain “stopwords” (e.g., “benefits,” “remote,” “company”) in addition to standard English stopwords.
3. **Normalization**
   * **Lemmatization** unified variations of words (e.g., “engineers” → “engineer”).
   * Whitespace, punctuation, and special characters were standardized. Segmentation errors such as “DataEngineer” were corrected to “Data Engineer.”
4. **Deduplication**
   * Any repeated or near-duplicate postings were identified and removed, culminating in a high-quality, domain-specific corpus of 644 entries.

**2.2 Keyword Extraction**

We employed **KeyBERT**—powered by the distilbert-base-uncased model—to extract keywords and short phrases from each job description. **Maximum Marginal Relevance (MMR)** ensured a balance between keyword relevance and diversity. Formally:

where is typically cosine similarity, is the set of candidate phrases, is the set of already chosen phrases, and is the document.

**2.3 Semantic Clustering**

After extracting keywords, we generated semantic embeddings using **SentenceTransformer** (all-MiniLM-L6-v2). These embeddings capture the contextual meaning of each keyword. We then performed **Hierarchical Clustering** with Ward’s linkage and **cosine similarity** as the distance metric:

1. **Cosine Similarity**

Cosine similarity effectively measures the angular distance between embedding vectors and is particularly robust for text-based tasks.

1. **Ward’s Linkage**  
   Ward’s method aims to minimize the total within-cluster variance:

where ​ represents cluster , is the centroid of cluster , and is the squared Euclidean distance of each embedding to the cluster centroid ​. By integrating these measures, we achieved coherent grouping of semantically related keywords.

**2.4 Social Network Analysis (SNA)**

For the **SNA**, we treated each **keyword** as a **node**, linking them by **edges** representing co-occurrences within the same job descriptions:

1. **Network Construction**
   * We built a co-occurrence matrix where counts how often keywords and appear together.
   * Low-frequency connections were filtered out to eliminate noise.
2. **Undirected, Weighted Graph**
   * Our resulting graph is undirected, with edge weights corresponding to co-occurrence frequency.
   * Node size in Gephi often correlates with the node’s **degree** or **weighted degree**.
3. **Network Metrics**
   * **Degree Centrality**: Indicates how many connections a node has.
   * **Betweenness Centrality**: Shows the extent to which a node serves as a “bridge” within the network.
   * **Clustering Coefficient**: Reflects how interconnected a node’s neighbors are.

**3. Results**

**3.1 Keyword Overview**

**(Insert “Data Engineer Keyword Word Cloud” image here)**

In addition to generating a word cloud, we tallied the **frequency** of each keyword across the **644** cleaned job descriptions (see partial data below). This helps contextualize which skills and terms appear most prominently in the dataset:

| **Keyword** | **Frequency** |
| --- | --- |
| data engineer | 200 |
| team deliver | 165 |
| python | 151 |
| data infrastructure | 115 |
| pipeline | 88 |
| sql | 78 |
| cloud infrastructure | 59 |
| snowflake | 43 |
| machine learning | 41 |
| business stakeholder | 36 |
| lake | 29 |
| highperformance trading | 27 |
| life cycle | 26 |
| aws | 25 |
| airflow | 22 |
| mentoring team | 22 |
| relationship delivering | 20 |
| crossfunctional | 19 |
| code maintaining | 19 |
| bash scripting | 18 |
| microsoft azure | 18 |
| software development | 17 |
| secure rightsized | 17 |
| version control | 17 |
| agile team | 17 |
| team responsible | 16 |
| pandas | 16 |
| analytics engineer | 16 |
| power bi | 16 |
| managing infrastructure | 15 |
| ... (others) | ... |

From this table, **“data engineer”** (200 occurrences) stands out as the most frequently mentioned term, reflecting its central role in the job descriptions. **“team deliver”** (165) appears surprisingly high, emphasizing the collaborative environment that Data Engineers operate in. **“python”** (151) reaffirms the predominance of programming expertise, while **“data infrastructure”** (115), **“pipeline”** (88), and **“sql”** (78) underscore the criticality of building and maintaining robust data pipelines.

Overall, these frequency counts align with the **semantic clusters** and **SNA** results, showing that both technical and collaborative dimensions of Data Engineering remain top priorities for employers.

**3.2 Clustering Outcomes**

**(Insert “Hierarchical Clustering Dendrogram” image here)**

Through hierarchical clustering, the keywords formed **six principal clusters**, as summarized below:

1. **Team Collaboration**
   * *Examples*: “team deliver,” “agile team,” “mentoring team,” “team responsible.”
   * Focus: Collaboration, leadership, and active team roles.
2. **Data Pipeline Tools**
   * *Examples*: “pipeline,” “snowflake,” “power bi,” “lake,” “airflow,” “life cycle.”
   * Focus: Orchestration platforms, BI reporting, and data-lake architectures.
3. **Stakeholder Engagement**
   * *Examples*: “nontechnical stakeholder,” “relationship delivering,” “secure rightsized,” “crossfunctional,” “highperformance trading.”
   * Focus: Communicating with diverse stakeholders, ensuring security, and specialized domains (e.g., trading).
4. **Cloud Infrastructure**
   * *Examples*: “data infrastructure,” “cloud infrastructure,” “aws,” “microsoft azure.”
   * Focus: Emphasis on deploying and maintaining scalable cloud-based solutions.
5. **Engineering & Software Development**
   * *Examples*: “data engineer,” “analytics engineer,” “software development.”
   * Focus: Bridging advanced engineering practices and data-centric roles.
6. **Core Technical Skills**
   * *Examples*: “python,” “machine learning,” “version control,” “sql,” “code maintaining,” “pandas,” “bash scripting.”
   * Focus: Programming, ML integration, DevOps fundamentals, and scripting languages.

**3.3 Expanded Social Network Analysis**

**(Insert “SNA Graph Visualization” image here)**

For our **SNA** in Gephi, each keyword is a node linked by co-occurrences in the same job description:

1. **High-Degree Nodes**
   * **Python** and **SQL** consistently appear in tandem with diverse tools and concepts, reflecting their foundational role in data workflows.
   * **AWS** also ranks high in node connections, highlighting the prevalent use of Amazon’s cloud services.
2. **Bridging Keywords**
   * **Python** demonstrates notable **betweenness centrality**, linking machine learning tasks with code maintenance and version control.
   * **Data infrastructure** forms a bridge between pipeline tools (Snowflake, Airflow) and multi-cloud requirements (AWS, Azure).
3. **Community Pockets**
   * A sub-community of ML-oriented keywords (e.g., “machine learning,” “pandas,” “analytics engineer”) co-occurs frequently, signaling the integration of ML tasks within Data Engineering pipelines.
   * Another cluster combines **cloud infrastructure** with pipeline tools, revealing a strong synergy between scaling architectures and orchestrating data flows.
4. **Soft-Skill Intersection**
   * **“team deliver,” “mentoring team,”** and **“crossfunctional”** connect to both technical keywords (pipeline, Python) and stakeholder-oriented terms, highlighting the interplay of communication and coding expertise.

Overall, the **SNA** reveals how frequently mentioned skills (e.g., Python, SQL) act as hubs, while related keywords (e.g., “airflow,” “machine learning”) form specialized clusters that branch out from these core competencies.

**4. Insights**

1. **Python and SQL as Cornerstone Skills**
   * The frequency table, clustering, and SNA collectively indicate that **Python** and **SQL** dominate job requirements, serving as the backbone for data manipulation and pipeline tasks.
2. **Evolving Cloud Demand**
   * High frequencies for **“cloud infrastructure”** (59), **“aws”** (25), and **“microsoft azure”** (18) confirm the push toward multi-cloud expertise.
   * Cloud solutions tightly integrate with pipeline tools such as **Snowflake** and **Airflow**.
3. **Collaboration & Stakeholder Focus**
   * Unexpectedly high frequencies for **“team deliver”** (165), **“relationship delivering”** (20), and **“crossfunctional”** (19) illustrate the necessity for strong communication and collaborative abilities in Data Engineering.
4. **Machine Learning On the Rise**
   * The presence of **“machine learning”** (41) alongside robust co-occurrences in SNA suggests a rising emphasis on operationalizing ML models within Data Engineering pipelines.

**5. Conclusion**

In this text mining project, we integrated **hierarchical clustering** and **Social Network Analysis** to uncover the trends shaping Data Engineering roles in the UK. The **keyword frequency table** highlights the preeminence of “data engineer,” “team deliver,” and “python,” underscoring the dual requirement of **technical prowess** and **collaborative capabilities**. Hierarchical clustering groups these keywords into thematic clusters, while the SNA visualization shows how skills and tools co-occur and form network communities in practical contexts.

Overall, the findings confirm that Data Engineering demands **strong coding fundamentals** (Python, SQL), **orchestration/pipeline expertise** (Airflow, Snowflake), **cloud proficiency** (AWS, Azure), and **cross-team collaboration**. As the data domain continues to evolve—integrating machine learning and advanced analytics—professionals who balance these capabilities are well-positioned to meet industry needs, and employers can optimize their hiring strategies by focusing on these core skill sets.