Data Understanding Report

Overview- Nature and Scope of the Data

The primary dataset for this project is the Kenya Companies Act, 2015, a comprehensive legal framework governing company law in Kenya. This text serves as the foundation for training and powering the Retrieval-Augmented Generation(RAG) system. The Act consists of over 1,000 sections and multiple schedules, providing detailed provisions on the formation, operation and regulation of companies in Kenya.

Data Sources and Acquisition Strategy

The data for this project will be sourced from authoritative and publicly accessible platforms to ensure legal credibility and compliance with open-access terms. These sources include:

i) **Kenya Law (KenyaLaw.org)** — the official online repository of Kenyan legislation.
ii) **Business Registration Service (BRS)** — which provides downloadable PDFs of the Act.
iii) **Government Gazette Publications** — for amendments or official interpretations.
iv) **Manual user input/Q&A logs (optional for tuning)** — if collected during testing phase.

All data sources are compliant with public domain or open-access terms, making them legally viable for research and innovation.

Data Acquisition and Preprocessing Plan

The data acquisition follows a structured plan. The end-to-end data ingestion and transformation pipeline includes:

i) Document Retrieval- downloading official PDFs of the Companies Act from Kenya Law and BRS.
ii) Text Extraction- converting PDFs to machine-readable text using OCR or text extraction tools.
iii) Text segmentation- chunking and preprocessing the text using Python (with libraries like spaCy) into logical units (e.g., section-wise).
iv) Embedding Generation and Vector Indexing- embedding these chunks into a vector database (e.g., FAISS or Pinecone) using transformer-based models.
v) Validation- Continually validating data integrity and accuracy during preprocessing and indexing stages.

Structure of the Legal Document

The Kenya Company Act is organised into 16 parts, each addressing specific aspects of company law:

Part I- Preliminary

Sections 1-4 cover the short title, commencement, objects of the Act, and key definitions.

Part II- Types of Companies

Sections 5-11 define various company types, including private, public, companies limited by shares or guarantee, and unlimited companies.

Part III- Company Formation

Sections 12-39 detail the process of incorporating companies, including requirements for registration, company names, and the issuance of certificates of incorporation.

Part IV- Company Constitution

Sections 40-55 discuss the company's constitution, including articles of association and their alteration.

Part V- Company Members

Sections 56-104 address membership, including the rights and obligations of members, and the maintenance of registers.

Part VI- Company Directors

Sections 105-169 outline the appointment, duties, and liabilities of directors, emphasizing accountability and fiduciary responsibilities.

Part VII- Company Secretaries

Sections 170-180 specify the qualification, appointment, and duties of company secretaries.

Part VIII- Company Meetings and Resolutions

Sections 181-225 provide guidelines on meetings, voting procedures, and passing of resolutions.

Part IX- Company Accounts and Audit

Sections 226-266 mandate the preparation of financial statements, auditing requirements, and filing of annual returns.

Part X- Company Charges

Sections 267-289 deal with the registration and satisfaction of company charges.

Part XI- Company Management and Administration

Sections 290-349 cover various administrative aspects, including record-keeping and company communications.

Part XII- Company Investigations

Sections 350-396 provide for the investigation of company affairs by inspectors appointed bythe Registrar.

Part XIII- Company Arrangements and Reconstriuctions

Sections 397-408 address schemes of arrangement, mergers, and takeovers.

Part XIV- Company Winding Up

Sections 409-510 details the procedures for voluntary and compulsory winding up of companies.

Part XV- Foreign Companies

Sections 511-538 regulate the registration and operation of foreign companies in Kenya.

Part XVI- General Provisions

Sections 539-1042 include miscellaneous provisions, penalties, and transitional arrangements.

Data Elements to Be Collected and Processed

The following legal components will be extracted, segmented, and stored for efficient retrieval and analysis:

i) Full Legal text of all sections, clauses and schedules of the Act.
ii) Legal definitions, duties, and procedures related to incorporation, governance, and compliance.

iii) Frequently asked legal questions (FAQs) from credible sources such as Kenya Law Reports or BRS offices.
iv) Legal summaries and commentary for potential fine-tuning and context expansion.

Feature Representation and Description

In the context of legal NLP, the following feature types are relevant:

i) Textual features- section headings, clause structures, definitions, and legal conditions.
ii) Metadata features- section numbers, tags (e.g., “private company”, “winding up”), and document hierarchy.
iii) Query features- keywords, named entities and legal intent extracted from user input (e.g., company type, director, registrar).

These features will support semantic searching, grounded answer generation and context-aware legal summarization.

Existing Work and Value Proposition

While similar tools exist globally- such as DoNotPay, LegalRobot, or ChatLaw- there is no known project applying a RAG architecture to the Kenya Companies Act.

Some institutions and researchers have explored digitizing Kenyan legislation and Legal chatbots for general Q&A.

However, this project is unique in that it applies modern LLM and RAG architecture (retrieval + generation) specifically to Kenyan corporate law, supports multilingual interaction ans targets high legal accuracy and grounding building up on NLP and legal-tech innovations to create a first-of-its-kind, legally grounded assistant focused on Kenya’s corporate regulatory environment.