Case Study – Legal Document Analysis

Primary Technology to use: Python, NLP, OCR

Goal: By solving this case, you should be able to demonstrate fundamental skills in Python, ML, OCR and NLP.

Feel free to reuse any appropriate code available on Github or other sources.

You should be able demonstrate and explain the steps taken to achieve the below results and how challenges were overcome. Please ask any questions that might arise while solving this case.

Part-1

1. Create a judgment knowledge base from all the supreme court judgments available on https://drive.google.com/open?id=10VLxsZXPDvui2qc2oyCZ_DMbUykXX1Tu

2. Applying NLP & ML concepts, train the machine to answer the following closed-domain questions on these judgments knowledge base (Q&A). Remember, there could be more than one answer to a given question sometimes.

    a. What is Courts Martial?
    b. What are the different types of Courts Martial?
    c. What types of Courts Martial does Section 108 provide for?
    d. Can an Arbitrator exercise the powers given to a Tribunal under section 11A?
    e. Can a temporary employee be replaced by another temporary employee?
    f. Who can replace a temporary employee?
    g. Which Section of the Act prohibits any person from erecting or re-erecting any building without written permission from the Corporation
    h. What is Judicial Review as described by Lord Brightman?
    i. What does the expression "undue hardship" relate to?
    j. How does [1953] S.C.R. 302 Industrial Disputes Act define "industry"?

Part-2

3. Using OCR Auto-read the attached writ petition document (SKT writ.pdf)
4. Auto-Extract text from pdf into word document and save with appropriate file name
5. Auto-Read the entire writ document in point 4 and Auto-summarise into maximum 2000 characters (excluding spaces) automatically using python and NLP. Eliminate all general English and focus on important specifics only. From the summary, the important points of the case should be clear. Besides the grievances raised by the petitioners, the summary should include the area of law, sections, acts, etc.

Part-3

6. Automatically perform ALL the following tasks:

    a. Using NLP techniques, Auto-Read the summary in point 5 above
    b. Using NLP techniques, Auto-Identify and Auto-extract the important unique issues, grievances, area of law, sections, acts, etc. captured in the summary of point 5
7. Based on the extracted details in 6b, identify relevant judgments that match and address the crux of the matter presented in the writ petition. For example, if point 5 has the phrases such as “building permission” and “Article 14”, then all of these points should be addressed in the matching judgment pdfs.
8. Auto-Save the judgments that map to the petition’s issues with appropriate file name in a local folder
    a. Produce a summary that does not exceed 5 lines for each of the identified judgments and save them all in one word file numbered and in alphabetical order