infoedge
ventures

InfoEdge Ventures
AI Hackathon 2024

Hack the future - Innovate with AI

Powered by H2S

Team Details

a. Team Name : **whiterock**
b. Leader name : **Packia vinslin D**
c. Statement : **Make the books & Documents easier to read using LLMs(EX. hard into simple lang editions)**

Brief about the idea

To develop an application that simplifies complex English or native language text found in PDF books into easy-to-understand English. The application will also identify difficult words within the text and provide their simpler definitions or translations in the reader's native language, integrating these simplified versions seamlessly into the original paragraph.

This tool aims to improve readability and comprehension for users who struggle with complex language, thereby promoting better learning and understanding.

Innovate methods for handling image-based or photocopied PDFs using computer vision techniques. The aim is to develop advanced solutions that enhance the extraction and processing of text from such PDFs, ensuring accurate recognition and improved readability.

How different is it from any of the other existing ideas?

- our approach automates book lang simplification and translation, using OCR and **GEMINI APIs** for efficiency, consistency, and scalability, **unlike manual rewriting.**

- You can create your own simplified version of the book without any burden **Especially documents.**

- The whole thing is in automation**, no labour work** ,no paperwork and also you can create easy readable version for **any Book**

How will it be able to solve the problem?

- **Automates Text Simplification:** Transforms complex English text into simpler language efficiently, reducing the need for manual rewriting the whole book.

- **Handles Image-Based PDFs:** Uses OCR and computer vision to extract text from image-based or photocopied PDFs, making all content accessible.

- **Reduces Costs**: Leverages automation and APIs to lower costs compared to manual text simplification and rewriting, ensuring a cost-effective solution.

USP of the proposed solution :

Efficient and Cost-Effective Book & Document Simplification:

The solution automates the process of transforming complex English text and translating difficult words, handles image-based PDFs using OCR, and integrates seamlessly with APIs. This combination ensures consistent, scalable, and affordable text simplification, making complex content accessible with **minimal manual effort and cost.**

List of features offered by the solution

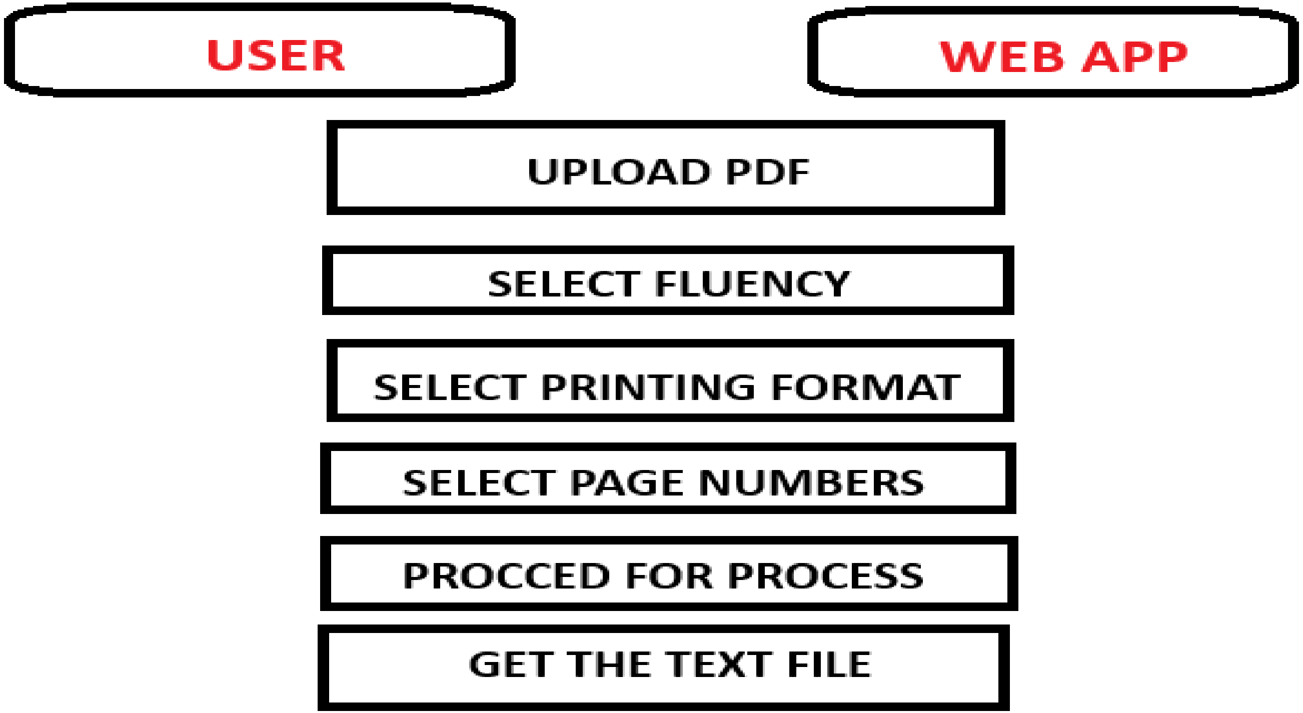
- **Beginner and Intermediate Modes**: Offers two levels of simplification based on user fluency ,beginner mode and intermediate mode.

- **Output as Text File**: Provides the simplified and translated text as an output text file for easy access and use.

- **Hard Word Identification**: Uses the Gemini API(replaced by any LLMs) to identify difficult words and provide simpler definitions.

- **Automated Text Simplification**: Converts complex English text into easier-to-understand language automatically.

Use-case diagram



USER

WEB APP

UPLOAD PDF

SELECT FLUENCY

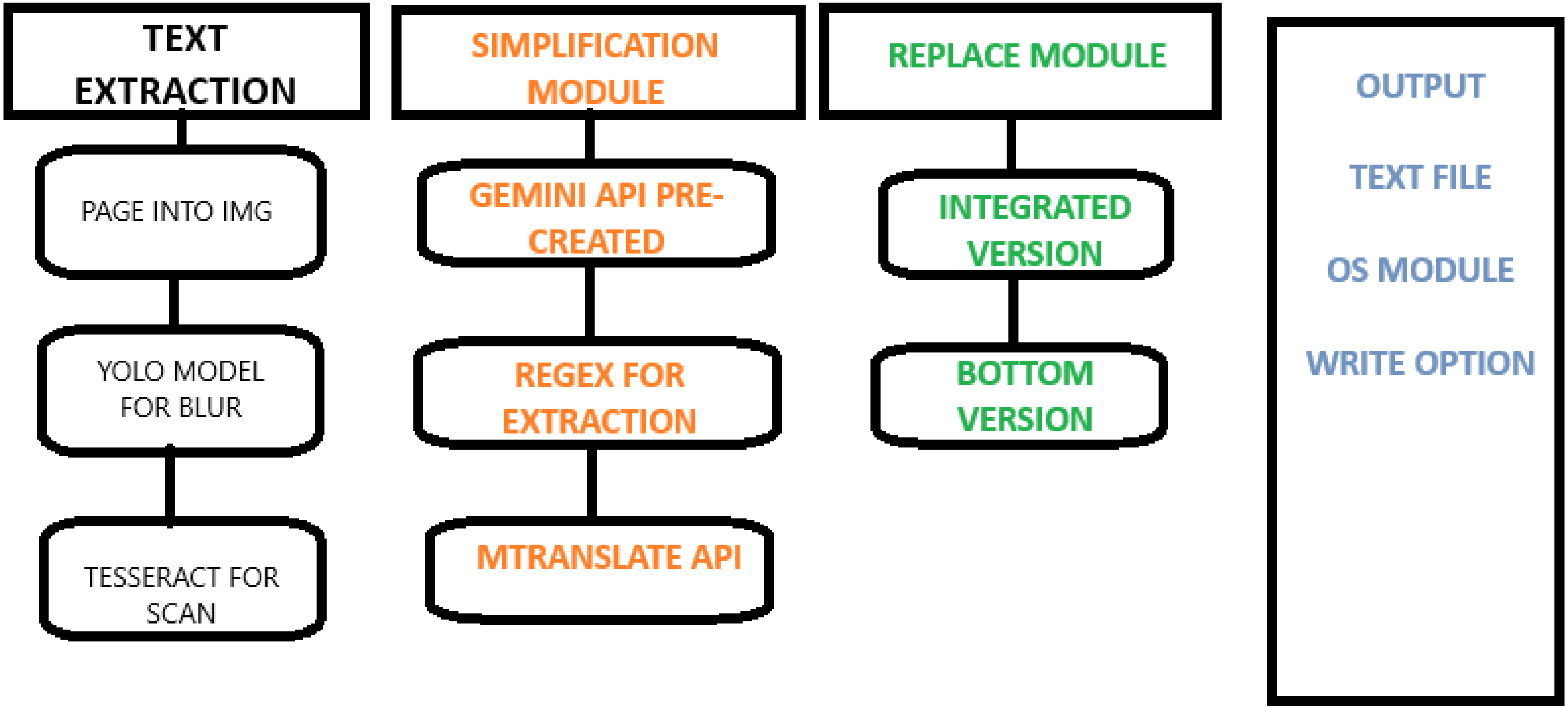
SELECT PRINTING FORMAT

SELECT PAGE NUMBERS

PROCCED FOR PROCESS

GET THE TEXT FILE

Architecture diagram of the proposed solution



TEXT EXTRACTION

PAGE INTO IMG

YOLO MODEL FOR BLUR

TESSERACT FOR SCAN

SIMPLIFICATION MODULE

GEMINI API PRE-CREATED

REGEX FOR EXTRACTION

MTRANSLATE API

REPLACE MODULE

INTEGRATED VERSION

BOTTOM VERSION

OUTPUT

TEXT FILE

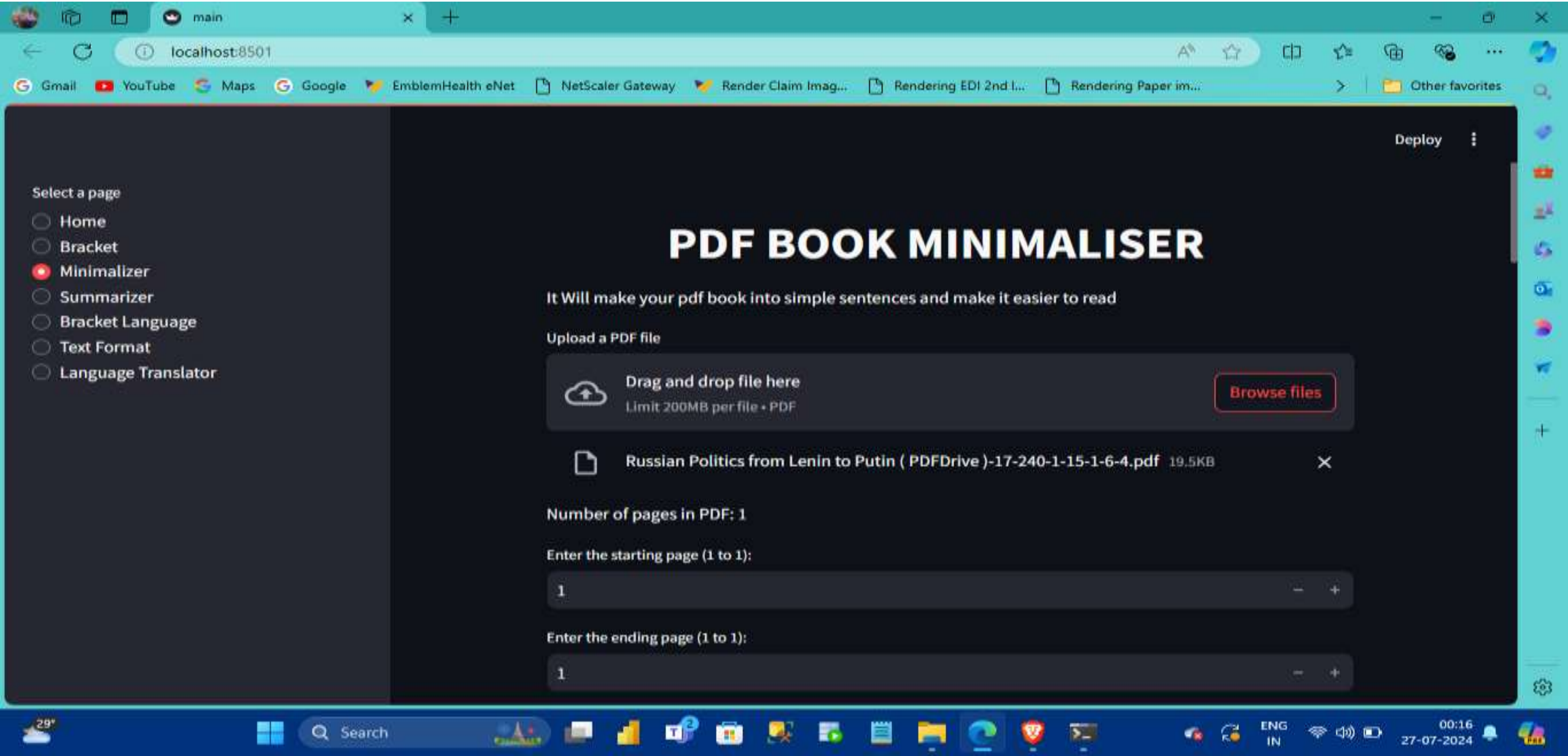
OS MODULE

WRITE OPTION

Technologies to be used in the solution:

- **TESSERACT OCR** – For text extraction
- **YOLO V8** – To find and remove unwanted objects like diagram
- **GEMINI AI ,NLTK** -- To find hard words and rephrase the paragraph
- **MTRANSLATE API** -- Translate one lang into another
- **REGEX** -- Python ‘ re ’ library

SNAPSHOTS OF THE PROTOTYPE



Select a page

○ Home

○ Bracket

● Minimalizer

○ Summarizer

○ Bracket Language

○ Text Format

○ Language Translator

PDF BOOK MINIMALISER

It Will make your pdf book into simple sentences and make it easier to read

Upload a PDF file

Drag and drop file here

Limit 200MB per file • PDF

Browse files

Russian Politics from Lenin to Putin (PDFDrive)-17-240-1-15-1-6-4.pdf 19.5KB

Number of pages in PDF: 1

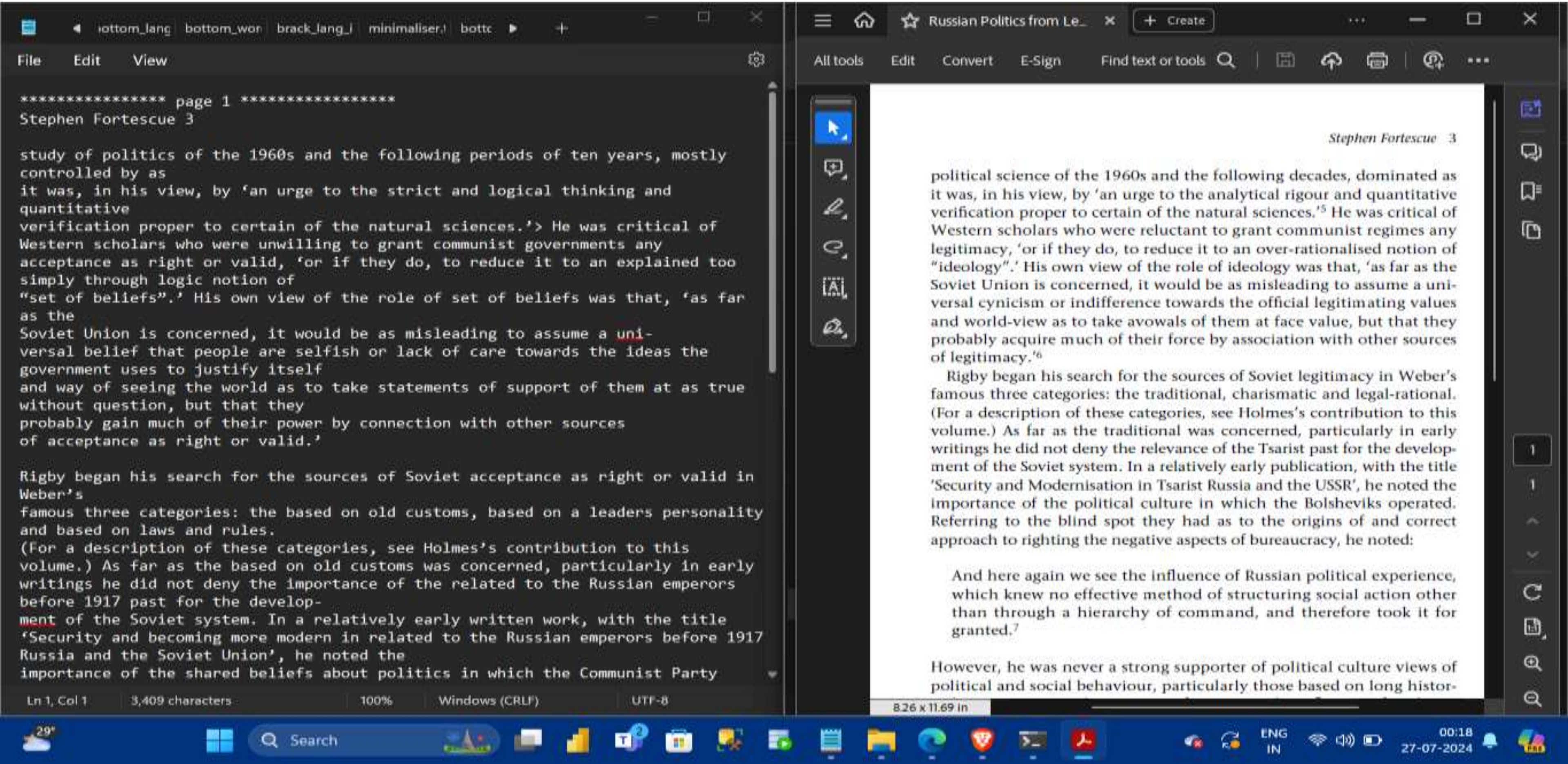
Enter the starting page (1 to 1):

1

Enter the ending page (1 to 1):

1

MINIMALIZED WORDS



```
**************** page 1 *****************
Stephen Fortescue 3

study of politics of the 1960s and the following periods of ten years, mostly
controlled by as
it was, in his view, by ‘an urge to the strict and logical thinking and
quantitative
verification proper to certain of the natural sciences.’> He was critical of
Western scholars who were unwilling to grant communist governments any
acceptance as right or valid, ‘or if they do, to reduce it to an explained too
simply through logic notion of
“set of beliefs”.’ His own view of the role of set of beliefs was that, ‘as far
as the
Soviet Union is concerned, it would be as misleading to assume a uni-
versal belief that people are selfish or lack of care towards the ideas the
government uses to justify itself
and way of seeing the world as to take statements of support of them at as true
without question, but that they
probably gain much of their power by connection with other sources
of acceptance as right or valid.’

Rigby began his search for the sources of Soviet acceptance as right or valid in
Weber’s
famous three categories: the based on old customs, based on a leaders personality
and based on laws and rules.
(For a description of these categories, see Holmes’s contribution to this
volume.) As far as the based on old customs was concerned, particularly in early
writings he did not deny the importance of the related to the Russian emperors
before 1917 past for the develop-
ment of the Soviet system. In a relatively early written work, with the title
‘Security and becoming more modern in related to the Russian emperors before 1917
Russia and the Soviet Union’, he noted the
importance of the shared beliefs about politics in which the Communist Party
```

Stephen Fortescue 3

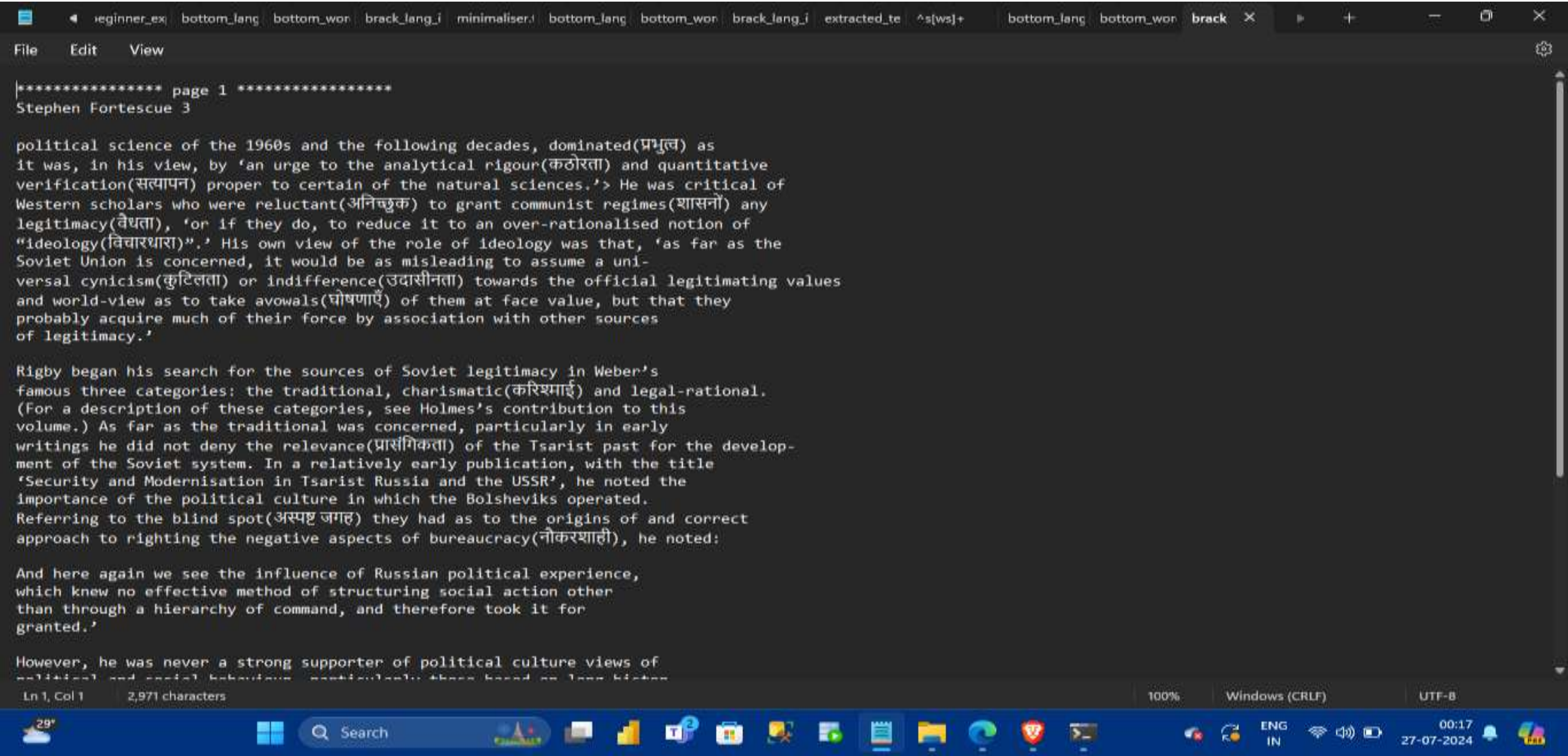
political science of the 1960s and the following decades, dominated as it was, in his view, by ‘an urge to the analytical rigour and quantitative verification proper to certain of the natural sciences.’[5] He was critical of Western scholars who were reluctant to grant communist regimes any legitimacy, ‘or if they do, to reduce it to an over-rationalised notion of “ideology”.’ His own view of the role of ideology was that, ‘as far as the Soviet Union is concerned, it would be as misleading to assume a universal cynicism or indifference towards the official legitimating values and world-view as to take avowals of them at face value, but that they probably acquire much of their force by association with other sources of legitimacy.’[6]

Rigby began his search for the sources of Soviet legitimacy in Weber’s famous three categories: the traditional, charismatic and legal-rational. (For a description of these categories, see Holmes’s contribution to this volume.) As far as the traditional was concerned, particularly in early writings he did not deny the relevance of the Tsarist past for the development of the Soviet system. In a relatively early publication, with the title ‘Security and Modernisation in Tsarist Russia and the USSR’, he noted the importance of the political culture in which the Bolsheviks operated. Referring to the blind spot they had as to the origins of and correct approach to righting the negative aspects of bureaucracy, he noted:

> And here again we see the influence of Russian political experience, which knew no effective method of structuring social action other than through a hierarchy of command, and therefore took it for granted.[7]

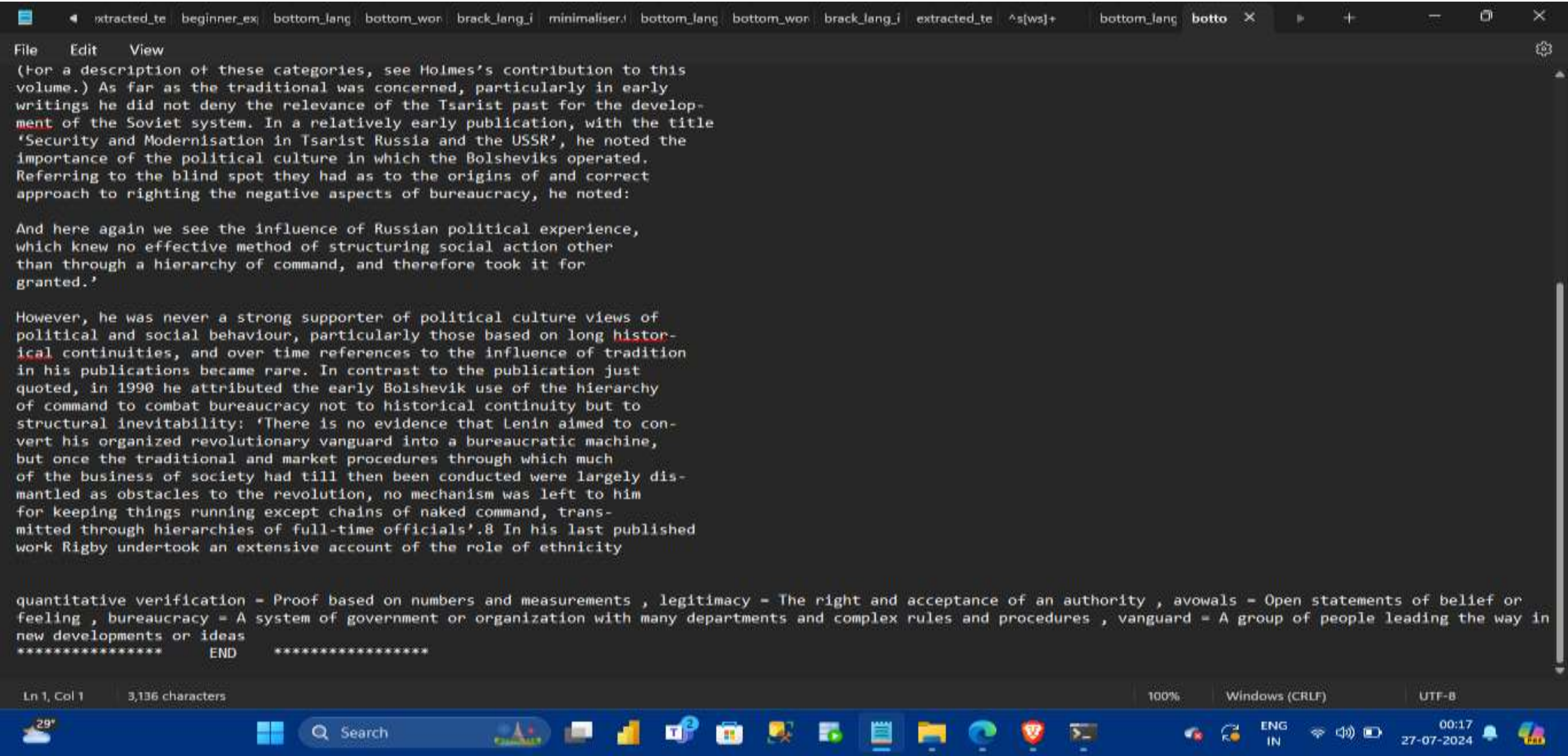
However, he was never a strong supporter of political culture views of political and social behaviour, particularly those based on long histor-

LANGUAGE ORIENTED



```
***************** page 1 ******************
Stephen Fortescue 3

political science of the 1960s and the following decades, dominated(प्रभुत्त्व) as
it was, in his view, by ‘an urge to the analytical rigour(कठोरता) and quantitative
verification(सत्यापन) proper to certain of the natural sciences.’> He was critical of
Western scholars who were reluctant(अनिच्छुक) to grant communist regimes(शासनों) any
legitimacy(वैधता), ‘or if they do, to reduce it to an over-rationalised notion of
“ideology(विचारधारा)”.’ His own view of the role of ideology was that, ‘as far as the
Soviet Union is concerned, it would be as misleading to assume a uni-
versal cynicism(कुटिलता) or indifference(उदासीनता) towards the official legitimating values
and world-view as to take avowals(घोषणाएँ) of them at face value, but that they
probably acquire much of their force by association with other sources
of legitimacy.’

Rigby began his search for the sources of Soviet legitimacy in Weber’s
famous three categories: the traditional, charismatic(करिश्माई) and legal-rational.
(For a description of these categories, see Holmes’s contribution to this
volume.) As far as the traditional was concerned, particularly in early
writings he did not deny the relevance(प्रासंगिकता) of the Tsarist past for the develop-
ment of the Soviet system. In a relatively early publication, with the title
‘Security and Modernisation in Tsarist Russia and the USSR’, he noted the
importance of the political culture in which the Bolsheviks operated.
Referring to the blind spot(अस्पष्ट जगह) they had as to the origins of and correct
approach to righting the negative aspects of bureaucracy(नौकरशाही), he noted:

And here again we see the influence of Russian political experience,
which knew no effective method of structuring social action other
than through a hierarchy of command, and therefore took it for
granted.’

However, he was never a strong supporter of political culture views of
```

SAME LANGUAGE EXPLANATIONS



(For a description of these categories, see Holmes's contribution to this volume.) As far as the traditional was concerned, particularly in early writings he did not deny the relevance of the Tsarist past for the development of the Soviet system. In a relatively early publication, with the title 'Security and Modernisation in Tsarist Russia and the USSR', he noted the importance of the political culture in which the Bolsheviks operated. Referring to the blind spot they had as to the origins of and correct approach to righting the negative aspects of bureaucracy, he noted:

And here again we see the influence of Russian political experience, which knew no effective method of structuring social action other than through a hierarchy of command, and therefore took it for granted.'

However, he was never a strong supporter of political culture views of political and social behaviour, particularly those based on long historical continuities, and over time references to the influence of tradition in his publications became rare. In contrast to the publication just quoted, in 1990 he attributed the early Bolshevik use of the hierarchy of command to combat bureaucracy not to historical continuity but to structural inevitability: 'There is no evidence that Lenin aimed to convert his organized revolutionary vanguard into a bureaucratic machine, but once the traditional and market procedures through which much of the business of society had till then been conducted were largely dismantled as obstacles to the revolution, no mechanism was left to him for keeping things running except chains of naked command, transmitted through hierarchies of full-time officials'.8 In his last published work Rigby undertook an extensive account of the role of ethnicity

quantitative verification = Proof based on numbers and measurements , legitimacy = The right and acceptance of an authority , avowals = Open statements of belief or feeling , bureaucracy = A system of government or organization with many departments and complex rules and procedures , vanguard = A group of people leading the way in new developments or ideas

**************** END *****************

Prototype Performance report/Benchmarking :

- **Accuracy and Efficiency** : The prototype effectively extracts text from both digital and scanned PDFs with over 90% accuracy using Tesseract OCR. Text simplification and translation show a high accuracy rate, ensuring readable and coherent outputs.
- **Processing Speed** : On average, the app processes a 100-page PDF in under 5 minutes, making it suitable for quick turnarounds. The use of efficient algorithms and APIs contributes to this speed.

Future Development :

- **LLM for Book Rephrasing** : Integrate a Large Language Model (LLM) like GPT-4 to rephrase complex text, ensuring high-quality, readable, and natural simplifications. This can provide users with varying levels of rephrasing options, from basic to comprehensive.

- **Book Language Translator Using LLM** : Use an LLM to translate entire books into multiple languages accurately, maintaining fluency and contextual appropriateness. This will enable users to access simplified content in their native languages.

- **Chapter-Wise Processing** : Implement chapter-wise processing, allowing users to upload, review, and edit books incrementally, enhancing workflow efficiency and making it easier to handle large texts.

GitHub link : https://github.com/vinslin/pdf_minimaliser/tree/main

Demo video URL : https://www.youtube.com/watch?v=EqXR-VegBmU

infoedge
ventures

InfoEdge Ventures

AI Hackathon 2024

Hack the future - Innovate with AI

Powered by H2S

THANK YOU!