DEVELOPMENT OF FILTERS FOR TRANSCRIPTION BETWEEN ENGLISH AND BENGALI DOCUMENTS

SUSHMITA SEN (EXAM ROLL : 111105027)

SURAJEET BHARATI (EXAM ROLL : 111105044)

PROJECT GUIDE : MANAS HIRA

What is it ?

- Transcript Bengali Document to equivalent English Script and vice-versa
- Look and Feel will be unchanged.
- Support Multiple format conversion (HTML, PDF, DOC)

- It will be a web application
- Any device with a browser can use

Dataset Design

- Storing and searching in a dictionary containing each and every word and its transcribed form will be wastage of resources

- The following approach is better in terms of resources

বালক	:	baalak
ব	-	b
া	-	aa
ল	-	la
ক	-	k

Dataset Sample

ক = k		খ = kh		গ = g		ঘ = gh		ঙ = ng
চ = ch		ছ = chh		জ = j		ঝ = jh		ঞ = ng
ট = t		ঠ = th		ড = d		ঢ = dh		ণ = n
ত = t		থ = th		দ = d		ধ = dh		ন = n
প = p		ফ = f,ph		ব = b		ভ = bh,v		ম = m
য = j		র = r		ল = l		শ = sh,s		ষ = sh,s
স = s		হ = h		ড় = r		ঢ় = rh		য় = y
ৎ = t		ং = ng		ঃ = :		ঁ = n		

Use of Lexical Analyzer

- The Data-set will be stored using Lexical Analyzer instead of any data structure

- Longest Match in input file will be replaced by its transcribed form

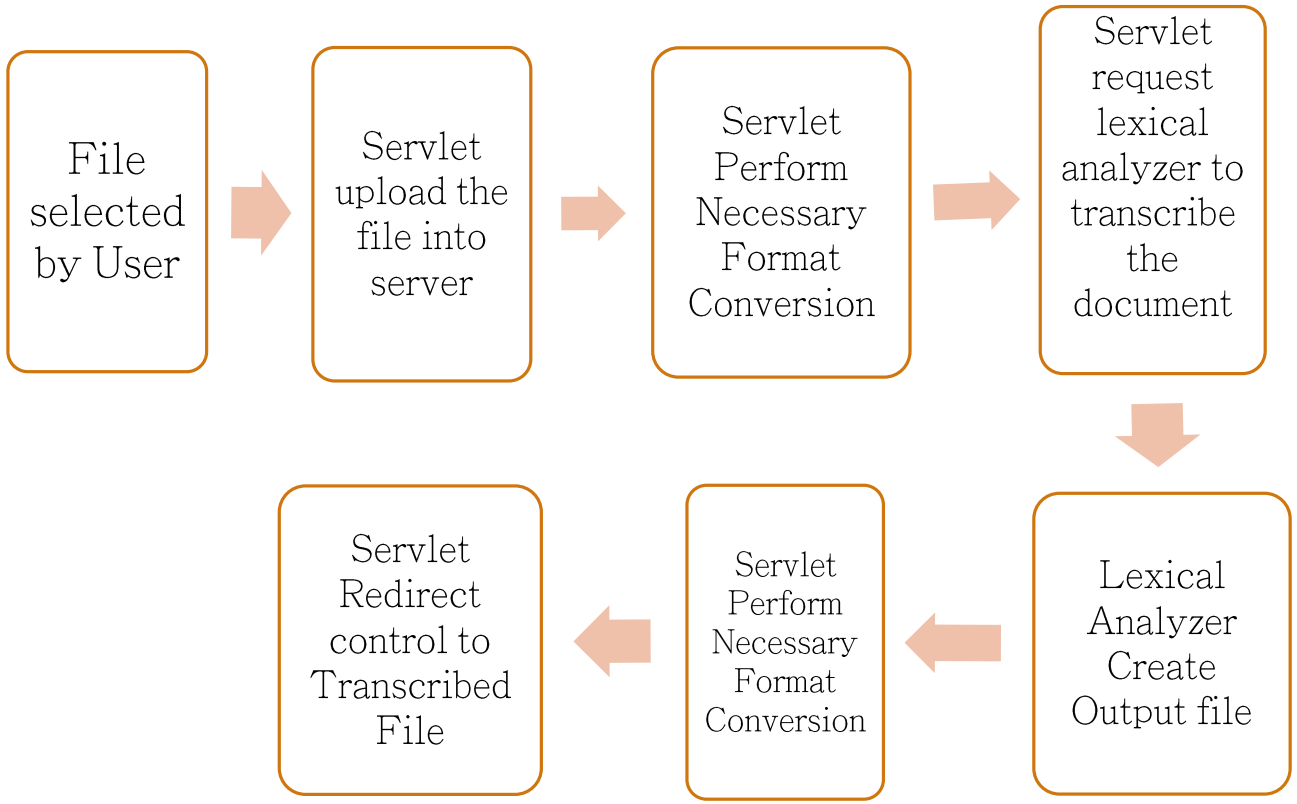
Implementation

Implementation

- Used JAVA Servlet to implement Back-end
- Used HTML5 and CSS3 to design front-end

- Used Apache Tomcat 7 as Java Web-server

Transcription Process



Transcription Between Different Formats

HTML to HTML

- Uploaded file will be transcribed using Lexical Analyzer
- Lexical analyzer parses the HTML file and replaces longest match

HTML to PDF

- Uploaded file will be transcribed using Lexical Analyzer
- Transcribed HTML output will be converted to PDF using WKHTMLTOPDF

HTML to DOC

- Uploaded file will be transcribed using Lexical Analyzer
- Transcribed HTML output will be converted to PDF using WKHTMLTOPDF
- Output PDF will be converted to DOC file using PDFBox API

PDF or DOC to Other Formats

- Convert to equivalent HTML file using text extraction API
- Do Transcription and format conversation as said before

- Look and feel will not be preserved.

Limitation

- Limitation in Data-set

  - Language Specific anomalies can not be covered.

- Limitation in File Format Conversion

  - During conversation for PDF and DOC input file.

Future Scope

- Improvement in data-set
- Improvement of the regular expressions in the lexical analyzer

- Use of better API for improvement in file format conversion.