DEVANAGARI DOCUMENT ENCODING CONVERTER

A Project Report

Submitted by

Sanghamitra Khobragade	**110803026**
Komal Kharat	**110803057**
Rupali Patil	**110808053**

in partial fulfilment for the award of the degree

of

B.Tech Computer Engineering/ Information Technology

Under the guidance of

Prof. Abhijit A. M.

College of Engineering, Pune

DEPARTMENT OF COMPUTER ENGINEERING AND INFORMATION TECHNOLOGY, COLLEGE OF ENGINEERING, PUNE-5

May, 2012

DEPARTMENT OF COMPUTER ENGINEERING AND INFORMATION TECHNOLOGY,
COLLEGE OF ENGINEERING, PUNE

CERTIFICATE

Certified that this project, titled “DEVANAGARI DOCUMENT ENCODING CONVERTER” has been successfully completed by

Sanghamitra Khobragade	**110803026**
Komal Kharat	**110803057**
Rupali Patil	**110808053**

and is approved for the partial fulfilment of the requirements for the degree of “B.Tech. Computer Engineering/Information Technology”.

SIGNATURE

ABHIJIT A. M.
Project Guide
Department of Computer Engineering
and Information Technology,
College of Engineering Pune,
Shivajinagar, Pune - 5.

SIGNATURE

DR. JIBI ABRAHAM
Head
Department of Computer Engineering
and Information Technology,
College of Engineering Pune,
Shivajinagar, Pune - 5.

Abstract

People has been using many available encodings and Non-Unicode fonts since long ago to write different devanagari documents. To make them universally accessible, searchable over the network, these documents must use unicode encoding and unicode font. Hence there is a need to convert an encoding and font of a Non-Unicoded devanagari document to unicode. The goal of this project is to provide a software, which converts an encoding and font of Non-Unicoded document to unicode. This software converts an encoding of Non-Unicode document to unicode by using unix command and Non-Unicode font of document to unicode font by byte-by-byte mapping.

This results in free, open source, encoding and font converter for Non-Unicode Devanagari document.

Contents

List of Figures

Chapter 1

Introduction

Computers understand only binary data. For representing character, encoding comes in picture which represents every character with unique number. There are many encodings which can represents certain set of character and use their own encoding standard. Hence there is problem that other encodings will identify these characters according to their own encoding. Hence there is a need for an encoding standard which can uniquely identify each character in the world which is Unicode standard.

Even though Unicode standard is widely acceptable over Internet, there are many software for writing devanagari documnet, which still do not support it. They use their own encoding standards. With the use of such software many more Devanagari documents has designed. Those documents are encoded using Non-Unicode standards.

Those Non-Unicode documents are neither web searchable nor sortable as web supports Unicode only. They are readable with only specific fonts and/or availability of particular software like shivaji font, akruti software respectively. Also they are not viewable on those operating systems which do not support those particular Non-Unicode standards.

Many documents which exist presently are Non-Unicode. Hence to cope up with above problems, those Non-Unicode documents would be converted to Unicode. There are software available to do this work. But they are not open source and free. Hence there is a need of designing an open source software.

1.1 Various software used to type Marathi language

1.1.1 Akruti

[6] Akruti is a multilingual software. Memory required to install akruti multifont engine is very less, about 8 MB. Using the multifont engine, akruti provides many keyboard layouts for many indic languages. The typing methodology is simple and very easy to learn. It also supports Unicode. It has dictionary, multiengine, spell checkers and font converter which helps to type in any font from Akruti family. Akruti fonts are compatible with various applications like Lotus Smartsuit, MS Office, Notepad, Corel Draw, Pagemaker, etc. It supports Unicode and extends to popular non-Unicode fonts. To type in indic language Scroll-lock must be on. For that it requires On-screen keyboard.

1.1.2 Baraha

[7] Baraha is a word processing application. It is used for creating documents in many Indian languages including Marathi by using a phonetic keyboard. Baraha pad is needed to edit or create documents in free version. Registered version can be used to type in any application. Baraha can export data in various data formats such as ANSI text, Unicode text. Baraha comes with True type fonts used to format the ANSI text. They are not Unicode based.

1.1.3 Google Transliterate

[9] Google Transliterate is an online application provided by Google. It is very easy to use and learn. It uses phonetic keyboard. Google Transliterate can switch between English and transliteration language. It can change the language of standard input. It

does not require specific application as in baraha. Google Transliterate uses Unicode to encode the characters.

1.1.4 Shree-lipi

[10] Shree-lipi contains various number of packages and these packages has number of fonts. It is multilingual supporter. It is compatible for many popular applications. It facilitates conversion of one font format to another. It also supports conversion of various types of files. It has all popular keyboard layout. It is Windows based software. Spellcheckers and official language dictionaries are available.

Chapter 2

Literature Study

Non-Unicode Devanagari fonts, various encoding schemes, Unicode are needed to be studied.

2.1 Study of various Non-Unicode Devanagari fonts:

[8] Various Non-Unicode Devanagari fonts and their character mappings need to be studied because using these software, document in a particular Non-Unicode Devanagari font or encoding is going to convert to Unicode. Character mappings of different encodings are reqiure to map their characters with Unicode charaters.

Non-Unicode Devanagari fonts:

- Akruti monolingual
- Akruti bilingual
- Kautilya
- DV-TTSurekh-Normal
- Kruti Dev 010
- Shivaji

- Shreelipi
- Tirkas
- Kiran
- Xdvng

2.2 Character Encoding Schemes

Character encoding schemes are to be studied to understand the use and various types of available character encodings, and way of representing characters by them. **Encoding :** The rules used to make assignment between character and bit pattern are called encoding schemes. These rules are to map each character to a numerical code that is then converted to binary and represented as one or several bytes. There are many encodings are available today but three (ASCII, EBCDIC, Unicode) are well known encoding techniques.

Character sets : It is range of characters handled by that encoding.

Coded Character Set : It is a character set where all characters have unique numbers by some method.

Some of available encoding standards are as follows:

2.2.1 EBCDIC

Extended Binary Coded Decimal Interchange Code; used by IBM mainframes and AS/400 machines. It represents each character by 8 bits. It encodes only 256 characters. Mainly used on IBM Mainframe and IBM midrange computer operating systems.

2.2.2 ASCII

American Standard Code for Information Interchange; used by almost all other hardware. The original ASCII character set includes 128 characters. The mapping of this character set to 7-bit numeric value. New characters are the extension of the ASCII character set from 128 to 256 characters. Out of 128 characters, 95 are printable and remaining 33 including space are non printable.

2.2.3 UNICODE

It is widely accepted by software and over network. Unicode will be seen in detail in next chapter.

Chapter 3

Unicode

Unicode is universal encoding standard. It uses unique value to represent a single character[5].

There are lots of encodings, like ISO-8859 and fonts like shivaji, shreelipi fonts, akruti fonts, etc are available to write devanagari documents. Each of these encoding provides their own standard and uses their own unique values to represent devanagari characters. Hence there may be a situation that, one character may has different numeric values in different encodings; for example, let an encoding A use 32 to represent character 'x' and an encoding B may use 23 to represent same character 'x'. There may be a case that two different characters may has same value in different encodings; for example, let an encoding A uses 32 to represent character 'x' and an encoding B may use 32 to represent character 'y'. But none of encoding is universal except Unicode. Many devanagari documents are not in Unicode encoding. Hence when data passes through different encoding systems or platforms, there may be chances of loss in data. So there is need to change their encoding to Unicode. Before Unicode, none of an encoding have enough characters to represent single language like English. An encoding ASCII can represent only 128 characters. For representing special characters, mathematical symbols in english, requires other encodings. Hence more than one encoding is required to represent single language like English. For this, single computer needs to support many encoding standards.

Unicode solve all above problems, by providing support to a large number of characters. Hence there is no need of many encodings to represent single language like English. It also supports many scripts like devanagari, latin, etc. *Unicode transformation format*(UTF) is a mapping of Unicode numeric value to a unique byte sequence.

Unicode is a standard for encoding, representation and handling text. Unicode is implemented by using different encodings like UTF-8, UTF-16, UTF-32.

3.1 UTF-8

[3, 11] It is Byte oriented and variable length encoding. It is compatible with ASCII. World Wide Web uses UTF-8. It can encodes any Unicode character. If splitting and concatenation of string is not proper, then it may lead to loss of data. For some languages, it takes more space.

Applications - World Wide Web, .txt, .html, .xml.

3.2 UTF-16

[1, 11] It uses 16 bits to represent single character and surrogates to encode more characters. Surrogates are code points from two special ranges of Unicode values, reserved for use in combination with other value. They are called surrogates, since they do not represent characters directly, but only used in pair. It has Byte Order Mark. A byte order mark (BOM) consists of the character code U+FEFF at the beginning of a data stream, which can be used as a signature defining the byte order and encoding form. Data types longer than a byte can be stored in computer memory with the most significant byte (MSB) first which is big-endian or last which is little-endian. It may be big endian

or little endian. It requires more memory to represent ASCII characters.

Applications - files on Linux, .NET environment variables and telnet.

3.3 UTF-32

[2, 11] It uses 32 bits to represent single character. Actually only 21 bits are used, remaining are masked, hence it wastes more space. It also requires more memory than above encoding techniques.

Application : HTML5.

Depending on the application, suitable Unicode Transformation Format is used. Unicode maps the Devanagari characters in the range 0900097F. Text files on Linux uses UTF-8 encoding. UTF-8 requires 1 byte for characters in range 0000-007F, 2 bytes for 0080-07FF range and 3 bytes for 0800-FFFF range. As Devanagari characters are between 0800-FFFF, UTF-8 takes 3 bytes to represent single devanagari character.

Chapter 4

Requirements Specification

Devanagari Document Encoding Converter should convert Devanagari document written using non-Unicode encodings to a document with Unicode encoding. It should convert text as well as odt files. It should maintain file format of odt files after conversion. GUI should be created for converting input files as well as given input text. New font table should be added to the converter, so that it can convert newly added font by user. Debian and rpm package of converter should be created, to run converter through command line. For this input file should be provided by user.

Chapter 5

Design Of Software

5.1 Programming language

[4, 14] Python is used as programming language for the project. As Python has more features. It is easy to use and understand. All modules required for the project are easily available in Python.

5.2 GUI

wxFormBuilder[12] is used to create GUI. It provides drag and drop facility to create forms. It uses wxPython[13] module of python.

5.3 Work Done so far

Mapping tables are created, which contain Non-Unicode characters and corresponding Unicode characters. As mapping table is UTF-8 encoded, Unicode characters written are also UTF-8 encoded and thus it uses 3 bytes per devanagari character. Input file can be odt or text file. For input odt file, output file must be odt. Convert programme does byte by byte mapping of input file characters to Unicode charcters using given font table.

Following are the list of Non-Unicode Devanagari fonts which can be converted using this programme.[8]

- Akruti monolingual
- Akruti bilingual
- Kautilya
- DV-TTSurekh-Normal
- Kruti Dev 010
- Shivaji
- Shreelipi
- Tirkas
- xdvng
- kiran

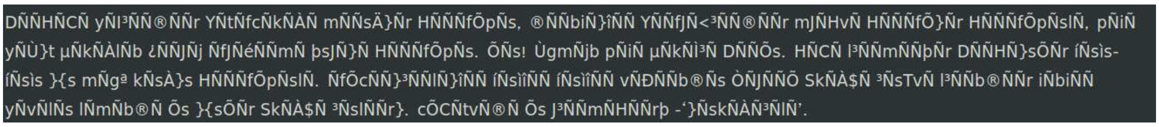
5.3.1 Command-line

Text File :

convertenco -i inputfilename.txt -o outputfilename.txt -f fontname This is a command to convert Non-Unicode text file to Unicode file text file.

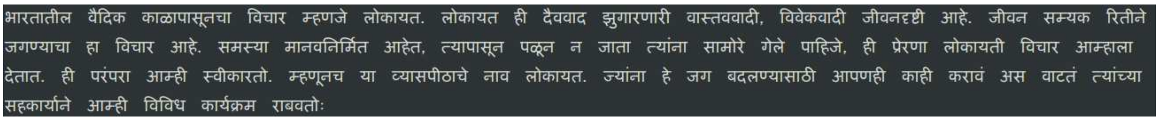
convertenco - It is a python program which actually converts Non-Unicode file to Unicode file. inputfilename.txt - This is input text file. outputfilename.txt - It is the Unicode output file which is desired. fontname - It is a mapping table which is a name of Non-Unicode Devanagari font in which the input file is encoded.

Similarly a pdf file can be converted. Only change done in above command is instead of inputfilename.txt inputfilename.pdf.



```
DÑÑHÑCÑ yÑl³ÑÑ®ÑÑr YÑtÑfcÑkÑÀÑ mÑÑsÄ}Ñr HÑÑÑfÕpÑs, ®ÑÑbiÑ}îÑÑ YÑÑfJÑ<³ÑÑ®ÑÑr mJÑHvÑ HÑÑÑfÕ}Ñr HÑÑÑfÕpÑslÑ, pÑiÑ
yÑÙ}t µÑkÑÀlÑb ¿ÑÑJÑj ÑfJÑéÑÑmÑ þsJÑ}Ñ HÑÑÑfÕpÑs. ÕÑs! ÙgmÑjb pÑiÑ µÑkÑÌ³Ñ DÑÑÕs. HÑCÑ l³ÑÑmÑÑþÑr DÑÑHÑ}sÕÑr íÑsìs-
íÑsìs }{s mÑgª kÑsÀ}s HÑÑÑfÕpÑslÑ. ÑfÕcÑÑ}³ÑÑlÑ}îÑÑ íÑsìîÑÑ íÑsìîÑÑ vÑÐÑÑb®Ñs ÒÑJÑÑÕ SkÑÀ$Ñ ³ÑsTvÑ l³ÑÑb®ÑÑr iÑbiÑÑ
yÑvÑlÑs lÑmÑb®Ñ Õs }{sÕÑr SkÑÀ$Ñ ³ÑslÑÑr}. cÕCÑtvÑ®Ñ Õs J³ÑÑmÑHÑÑrþ -‘}ÑskÑÀÑ³ÑlÑ’.
```

Figure 5.1: Input file to the programme convert encoded using Akrutimonolingual font



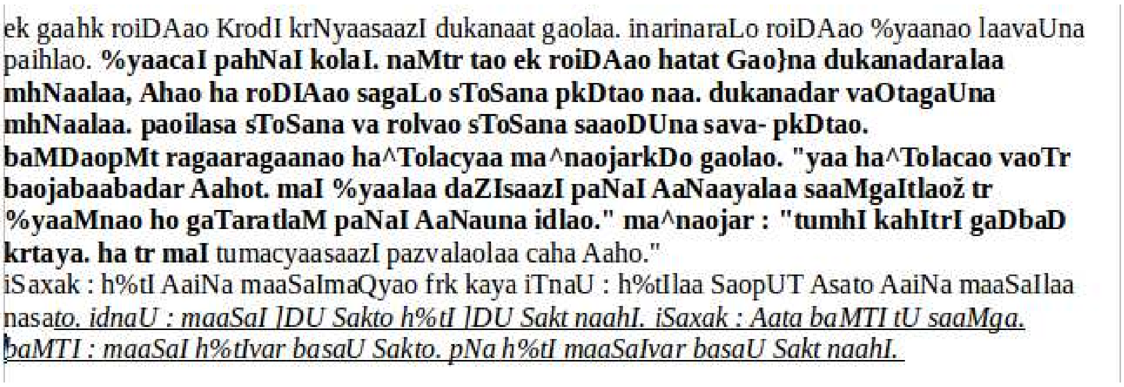
भारतातील वैदिक काळापासूनचा विचार म्हणजे लोकायत. लोकायत ही दैववाद झुगारणारी वास्तववादी, विवेकवादी जीवनदृष्टी आहे. जीवन सम्यक रितीने
जगण्याचा हा विचार आहे. समस्या मानवनिर्मित आहेत, त्यापासून पळून न जाता त्यांना सामोरे गेले पाहिजे, ही प्रेरणा लोकायती विचार आम्हाला
देतात. ही परंपरा आम्ही स्वीकारतो. म्हणूनच या व्यासपीठाचे नाव लोकायत. ज्यांना हे जग बदलण्यासाठी आपणही काही करावं अस वाटतं त्यांच्या
सहकार्याने आम्ही विविध कार्यक्रम राबवतोः

Figure 5.2: Output Unicode file of the program convert which is desired

Odt File :

convertenco inputfilename.odt outputfilename.odt fontname This is a command to convert Non-Unicode odt file to Unicode file odt file.

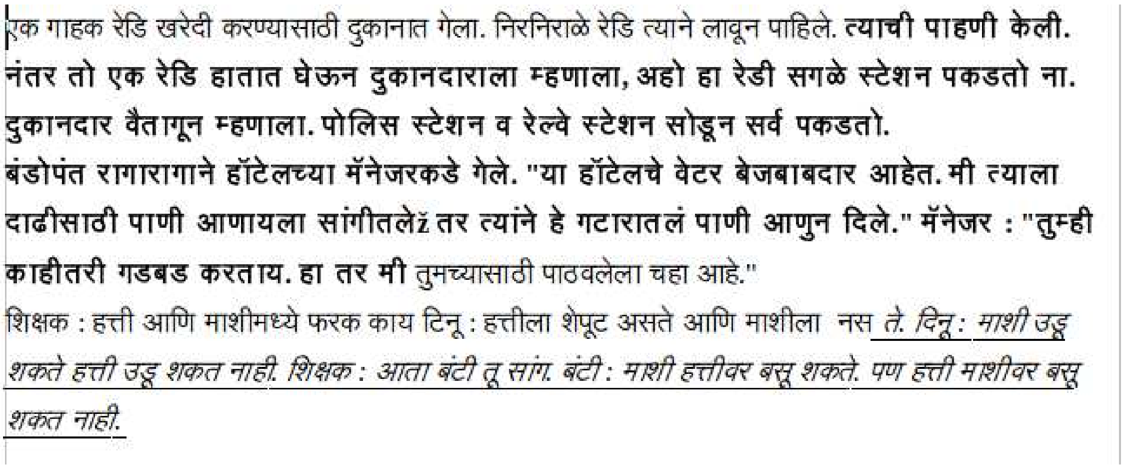
It maintains file format of odt file, by getting differnt tags from content.xml and style.xml of input odt file. Actual data is obtained from content.xml and it is passed to convert program. Font of file are changed to a Lohit Hindi, by editing both content.xml and style.xml.



ek gaahk roiDAao KrodI krNyaasaazI dukanaat gaolaa. inarinaraLo roiDAao %yaanao laavaUna paihlao. **%yaacaI pahNaI kolaI. naMtr tao ek roiDAao hatat Gao}na dukanadaralaa mhNaalaa, Ahao ha roDIAao sagaLo sToSana pkDtao naa. dukanadar vaOtagaUna mhNaalaa. paoilasa sToSana va rolvao sToSana saaoDUna sava- pkDtao. baMDaopMt ragaaragaanao ha^Tolacyaa ma^naojarkDo gaolao. "yaa ha^Tolacao vaoTr baojabaabadar Aahot. maI %yaalaa daZIsaazI paNaI AaNaayalaa saaMgaItlaož tr %yaaMnao ho gaTaratlaM paNaI AaNauna idlao." ma^naojar : "tumhI kahItrI gaDbaD krtaya. ha tr maI** tumacyaasaazI pazvalaolaa caha Aaho."

iSaxak : h%tI AaiNa maaSaImaQyao frk kaya iTnaU : h%tIlaa SaopUT Asato AaiNa maaSaIlaa nasa*to. idnaU : maaSaI]DU Sakto h%tI]DU Sakt naahI. iSaxak : Aata baMTI tU saaMga. baMTI : maaSaI h%tIvar basaU Sakto. pNa h%tI maaSaIvar basaU Sakt naahI.*

Figure 5.3: Input odt file encoded using Akrutimonolingual font



एक गाहक रेडि खरेदी करण्यासाठी दुकानात गेला. निरनिराळे रेडि त्याने लावून पाहिले. **त्याची पाहणी केली. नंतर तो एक रेडि हातात घेऊन दुकानदाराला म्हणाला, अहो हा रेडी सगळे स्टेशन पकडतो ना. दुकानदार वैतागून म्हणाला. पोलिस स्टेशन व रेल्वे स्टेशन सोडून सर्व पकडतो.**
बंडोपंत रागारागाने हॉटेलच्या मॅनेजरकडे गेले. "या हॉटेलचे वेटर बेजबाबदार आहेत. मी त्याला दाढीसाठी पाणी आणायला सांगीतलेž तर त्यांने हे गटारातलं पाणी आणुन दिले." मॅनेजर : "तुम्ही काहीतरी गडबड करताय. हा तर मी तुमच्यासाठी पाठवलेला चहा आहे."
शिक्षक : हत्ती आणि माशीमध्ये फरक काय टिनू : हत्तीला शेपूट असते आणि माशीला नस *ते. दिनू : माशी उडु शकते हत्ती उडु शकत नाही. शिक्षक : आता बंटी तू सांग. बंटी : माशी हत्तीवर बसू शकते. पण हत्ती माशीवर बसु शकत नाही.*

Figure 5.4: Output odt file encoded using Akrutimonolingual font

5.3.2 GUI

It helps to add new fonts to table and convert string and file. Home page of converter is as shown in following figure.

Browse for new font file and table gets directly added to table directory.

Browse for a input file, then give an output file name and select fontname from list. Click on convert, then dialogue box appears to save the output file. By selecting location to save, output file gets saved, as shown in following figure.

Input string is given in the textbox, and select the font name from list and click on convert. Converted text will appear in other textbox, as shown in following figure.

5.3.3 Web-based Converter

It is an another view of this software. Using web-based converter user can add new fonts to converter, and convert input string and file online, without actually installing the software own computer. Home page of converter website is as shown in following figure.

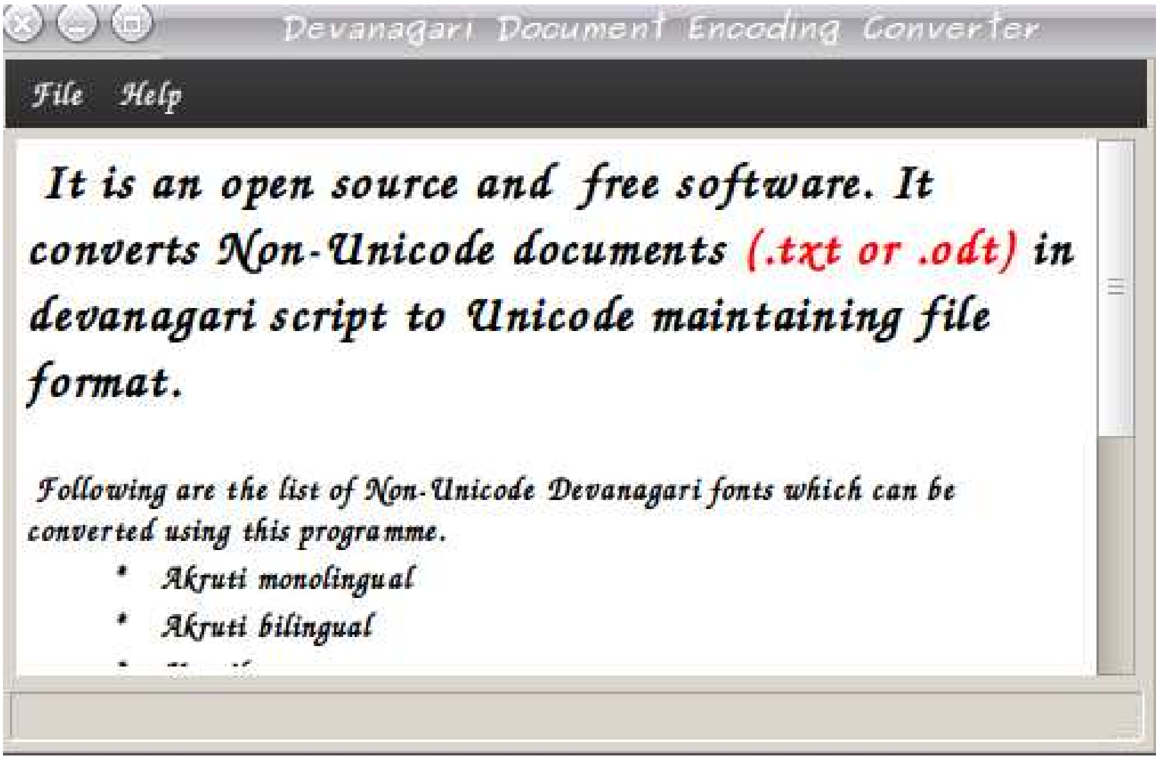
Browse for new font file and table gets directly added to table directory.

Browse for a input file, then give an output file name and select fontname from list. Click on convert, on new page output file will be made available to download.

Input string is given in the textbox, and select the font name from list and click on convert. Converted text will appear in other textbox on another page.

5.3.4 Packages

Debian and rpm packages of converter is created. To install on ubuntu operating system, use debian package and for redhat operating system use, rpm package. Debian package can be installed by command dpkg -i package-name.deb and removed by command dpkg -r package-name. Simillarly rpm package can be installed by command rpm -i package-



It is an open source and free software. It converts Non-Unicode documents (.txt or .odt) in devanagari script to Unicode maintaining file format.

Following are the list of Non-Unicode Devanagari fonts which can be converted using this programme.

- Akruti monolingual
- Akruti bilingual

Figure 5.5: Home page of converter

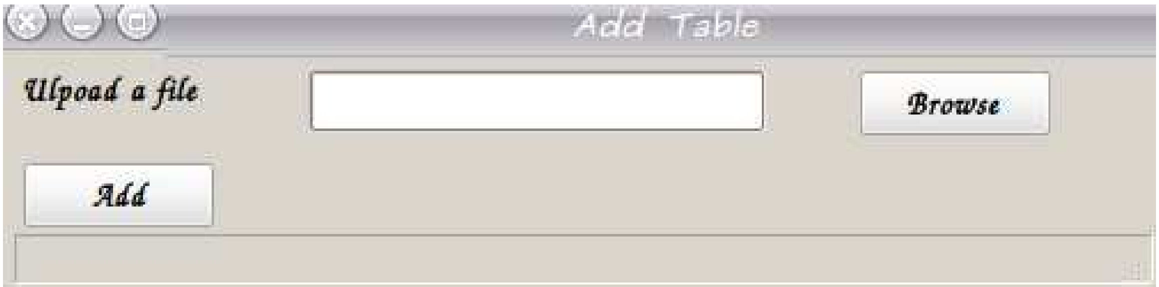


Figure 5.6: Adding new fonts to table

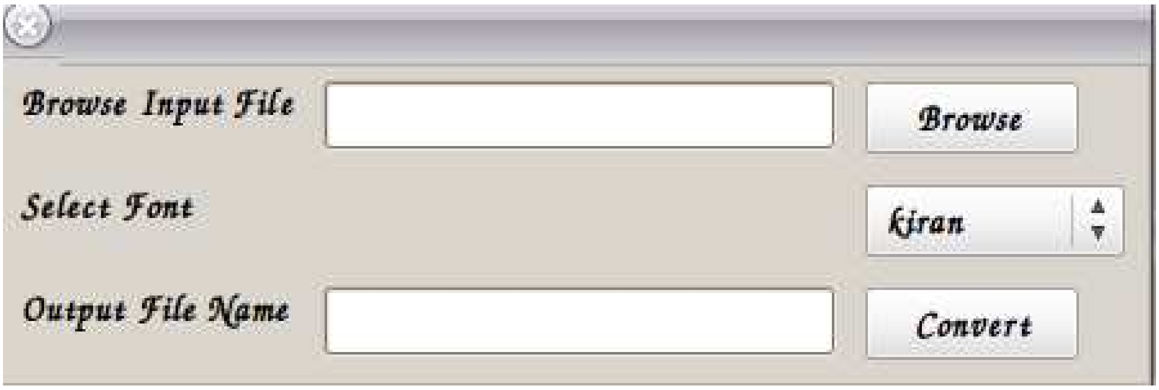


Figure 5.7: GUI for accepting file as a input

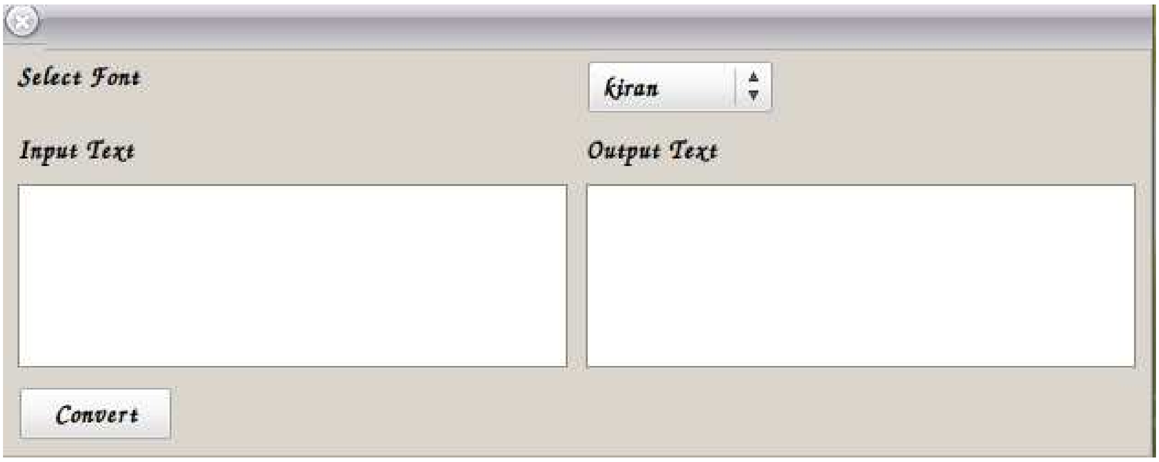
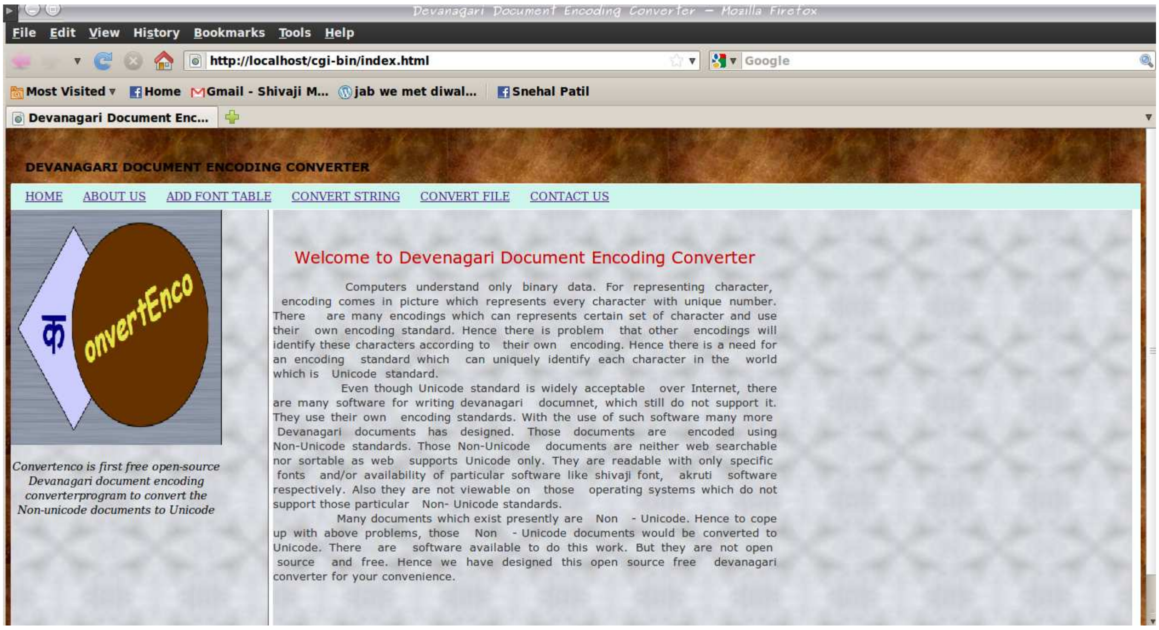


Figure 5.8: GUI for converting srting


Devanagari Document Encoding Converter – Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost/cgi-bin/index.html
Google
Most Visited
Home
Gmail - Shivaji M...
jab we met diwal...
Snehal Patil
Devanagari Document Enc...
DEVANAGARI DOCUMENT ENCODING CONVERTER
HOME
ABOUT US
ADD FONT TABLE
CONVERT STRING
CONVERT FILE
CONTACT US
क onvertEnco

Convertenco is first free open-source Devanagari document encoding converterprogram to convert the Non-unicode documents to Unicode

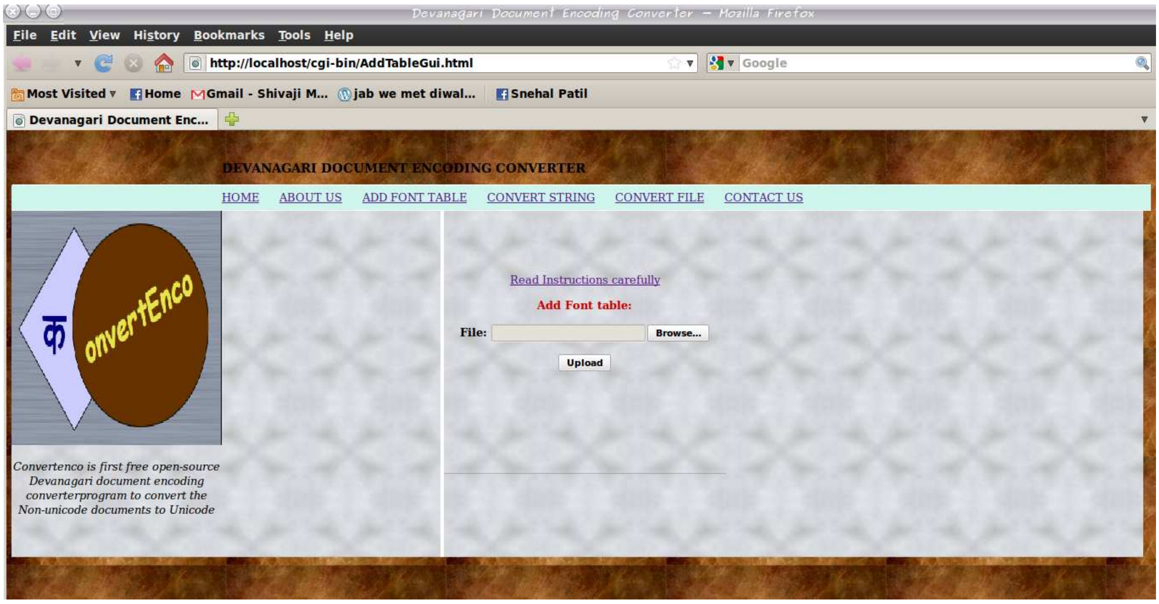
Welcome to Devenagari Document Encoding Converter

Computers understand only binary data. For representing character, encoding comes in picture which represents every character with unique number. There are many encodings which can represents certain set of character and use their own encoding standard. Hence there is problem that other encodings will identify these characters according to their own encoding. Hence there is a need for an encoding standard which can uniquely identify each character in the world which is Unicode standard.

Even though Unicode standard is widely acceptable over Internet, there are many software for writing devanagari documnet, which still do not support it. They use their own encoding standards. With the use of such software many more Devanagari documents has designed. Those documents are encoded using Non-Unicode standards. Those Non-Unicode documents are neither web searchable nor sortable as web supports Unicode only. They are readable with only specific fonts and/or availability of particular software like shivaji font, akruti software respectively. Also they are not viewable on those operating systems which do not support those particular Non- Unicode standards.

Many documents which exist presently are Non - Unicode. Hence to cope up with above problems, those Non - Unicode documents would be converted to Unicode. There are software available to do this work. But they are not open source and free. Hence we have designed this open source free devanagari converter for your convenience.

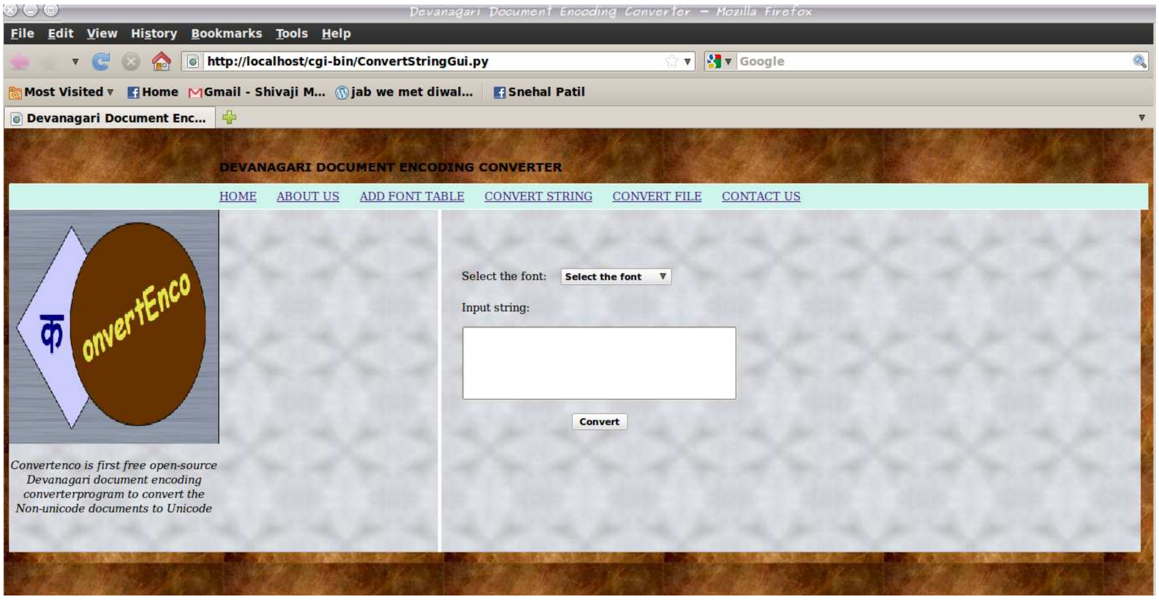
Figure 5.9: Home page of converter



Devanagari Document Encoding Converter - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost/cgi-bin/AddTableGui.html
Google
Most Visited Home Gmail - Shivaji M... jab we met diwal... Snehal Patil
Devanagari Document Enc...
DEVANAGARI DOCUMENT ENCODING CONVERTER
HOME ABOUT US ADD FONT TABLE CONVERT STRING CONVERT FILE CONTACT US
क onvertEnco
Read Instructions carefully
Add Font table:
File:
Browse...
Upload

Convertenco is first free open-source Devanagari document encoding converterprogram to convert the Non-unicode documents to Unicode

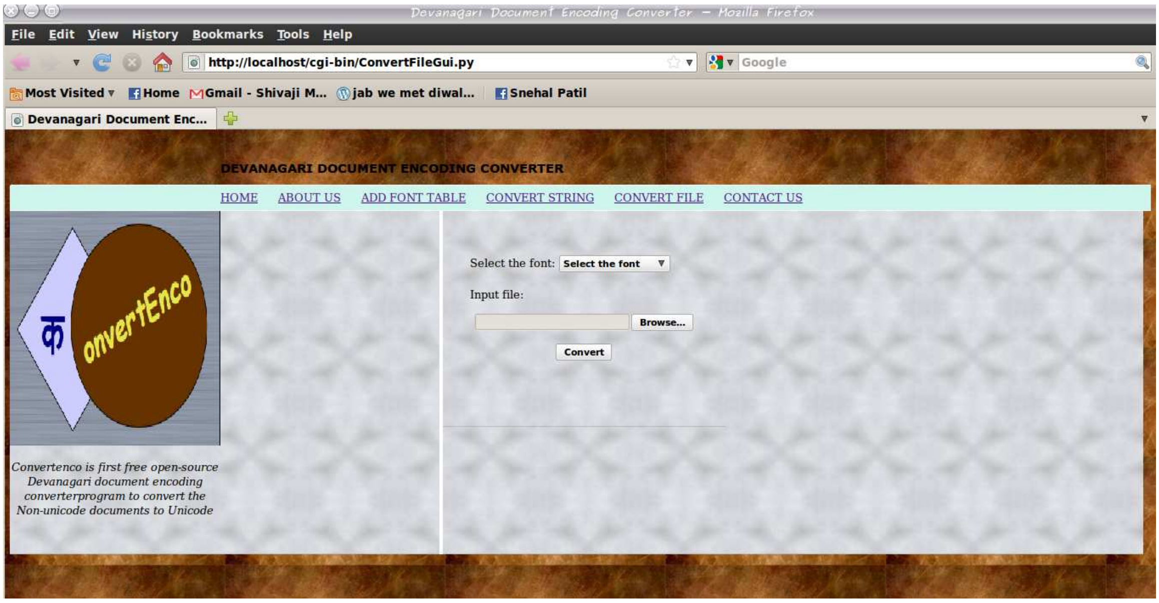
Figure 5.10: Adding new fonts to table



Devanagari Document Encoding Converter - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost/cgi-bin/ConvertStringGui.py
Google
Most Visited Home Gmail - Shivaji M... jab we met diwal... Snehal Patil
Devanagari Document Enc...
DEVANAGARI DOCUMENT ENCODING CONVERTER
HOME ABOUT US ADD FONT TABLE CONVERT STRING CONVERT FILE CONTACT US
क onvertEnco
Select the font:
Select the font
Input string:
Convert

Convertenco is first free open-source Devanagari document encoding converterprogram to convert the Non-unicode documents to Unicode

Figure 5.11: GUI for accepting file as a input


Devanagari Document Encoding Converter - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost/cgi-bin/ConvertFileGui.py
Google
Most Visited
Home
Gmail - Shivaji M...
jab we met diwal...
Snehal Patil
Devanagari Document Enc...
DEVANAGARI DOCUMENT ENCODING CONVERTER
HOME
ABOUT US
ADD FONT TABLE
CONVERT STRING
CONVERT FILE
CONTACT US
क onvertEnco
Browse...
Convert

Select the font: Select the font

Input file:

Convertenco is first free open-source Devanagari document encoding converterprogram to convert the Non-unicode documents to Unicode

Figure 5.12: GUI for converting srting

name.rpm and removed by command rpm -e package-name.

After installation, directly command *convertenco* and *convertergui* can be used. convertenco is a command to run converter through commandline and its format is *convertenco -i input-file -o output-file -f font-name*. For example, convertenco -i project/DDEC/input/shivajiin -o output.txt -f shivaji. convertergui is a command for converter gui. After running this command, gui will appear.

Bibliography

[1] http://en.wikipedia.org/wiki/utf-16.

[2] http://en.wikipedia.org/wiki/utf-32.

[3] http://en.wikipedia.org/wiki/utf-8.

[4] http://python.org.

[5] http://unicode.org/standard/whatisunicode.html.

[6] http://www.akruti.com.

[7] http://www.baraha.com/about.html.

[8] http://www.devanagarifonts.net.

[9] http://www.google.com/transliterate/about.html.

[10] http://www.modular-infotech.com/html/shreelipi.html.

[11] http://www.unicode.org/faq/utf-bom.html.

[12] http://www.wxformbuilder.org.

[13] http://www.wxpython.org/.

[14] Mark Lutz and David Ascher. *Learning Python*. O'Reilly and Associates.