Towards an Application Programming Interface (API) for Processing Kurdish Text

Dr. Abdul-Rahman Mawlood-Yunis
PhD from the School of Computer Science,
Carleton University,
Ottawa, Ont., Canada
armyunis@scs.carleton.ca
Outline

• Motivation
• Environment setup
• Character coding, read and write files
• Kurdish text processing operations
• Applications
• Conclusion
• Future work
• Promising Computer study trends for Kurdistan region
Motivation

- In order to use computers successfully in our daily life (e.g., business, government and research) we need an API for Kurdish text processing.

- An API for Kurdish text processing will open up doors for unlimited number of applications.

- Assists in standardizing Kurdish Language and Kurdish writing.
Outline

• Motivation
• Environment setup
  • Character coding, read and write files
  • Kurdish text processing operations
• Applications
• Conclusion
• Future work
• Promising Computer study trends for Kurdistan region
Eclipse setup for Kurdish text display

بۆ نەوەی بتوانیەن بە زمانتی کوردی بنووسەن پێویستمەن بە کۆدینگیکە
کە پیتێک کوردی پێ بنووسرە.
دبتوانریت بۆ نەوەیەم مامەستە بە کاربێت. UTF-8

C:\Users\Rahman\workspace>java Slaw
???? ????????? ?? (کوردی)
C:\Users\Rahman\workspace>java Slaw
Hello World (English)

Eclipse setup

1. Run → Run configuration → common tab → select utf-8 coding
2. Go to Eclipse -> Preferences -> General -> Appearance -> Colors and Fonts -> Debug -> Console font
3. Control Panel\System and Security\System → advance system settings → Environment variable → create new user variable
JAVA_TOOL_OPTIONS: -Dfile.encoding=UTF8
Eclipse javadoc with utf-8

- **JavaDoc setup** (to enter comments: shift-alt-J)
  - project ➔ generate javadoc in configuration choose javadoc.exe
    - for example:
      - C:\Program Files\Java\jdk1.7.0_04\bin\javadoc.exe

- project-> javadoc -> next -> in extra vm options write
  - -encoding UTF-8 -charset UTF-8 -docencoding UTF-8

  ```java
  //readFileToList("C:\\Users\\Rahman\\workspace\\goran.txt");
  // WriteListToFileToColumn("C:\\Users\\Rahman\\workspace\\goran_out.txt");
  ```
Redirect console output to separate frame

• PipedInputStream pin=new PipedInputStream()
• PipedOutputStream pout = new PipedOutputStream(this.pin)
• System.setOut(new PrintStream(pout, true))
• Catch Exceptions

// new RedirectConsoleOutput();
Redirect console output to file

• Run Configurations -> Common and in the Standard Input and Output choose File

• Other integration environments include, NetBean, jEdit

//KurdLangApi.count_words("C:\Users\Rahman\workspace\hawlati-24-6-2012\z1.txt");
Outline

• Motivation
• Environment setup
• **Character coding, read and write files**
• Kurdish text processing operations
• Applications
• Future work
• Promising Computer study trends for Kurdistan region
Kurdish character in UTF-8 representation

- **The extreme UTF-8 table**

- **Some special characters**
  
  \{ 33, 34, 40, 41, 44, 45, 46, 47, 58, 95, 1548, 1563, 1567, 1569, 1570, 1571, 1572, 1573, 1654, 8211, 8230, 61623, 65279 \}

- **Can be seen in the program debugging mode**

```cpp
//kurdishUnicodeCharValues();
```
1. Reader reader = new InputStreamReader(new FileInputStream("C:\Users\Rahman\workspace\h1.txt"), "UTF-8")

2. fin = new BufferedReader(reader)

3. Writer writer = new OutputStreamWriter(new FileOutputStream("C:\Users\Rahman\workspace\out1.txt"), "UTF-8")

4. BufferedWriter fout = new BufferedWriter(writer)

5. while ((s = fin.read()) != -1) {
    fout.write((char)s)
}

6. fin.close()
    fout.close()

//ReadAndWriteFile();
Outline

- Motivation
- Environment setup
- Character coding, read and write files
- Kurdish text processing operations
- Applications
- Future work
- Promising Computer study trends for Kurdistan region
Kurdish text processing operations

- Counting words
  - `isSpace, isNumeric`
- Sorting words
  - `System.getProperty("line.separator")`
- Cleaning words from noise
- The frequency use of ئ in Kurdish writing

`org.apache.commons.lang3.StringUtils` jar file

```java
// 1. KurdLangApi.count_words("C:\\Users\\Rahman\\workspace\\hawlati-24-6-2012\\z2.txt"); // isSpace
// 2. readFileToList("C:\\Users\\Rahman\\workspace\\goran.txt");
// WriteListToFileToColumn("C:\\Users\\Rahman\\workspace\\goran_out.txt"); // line separator
// 3. KurdLangApi.remove_two_letter_words(fin, fout)
```
Outline

• Motivation
• Environment setup
• Character coding, read and write files
• Kurdish text processing operations
• Applications
• Future work
• Promising Computer study trends for Kurdistan region
### Application
#### Most common words in Kurdish

Ex: English common words

<table>
<thead>
<tr>
<th>Rank</th>
<th>Word</th>
<th>Rank</th>
<th>Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>the</td>
<td>11</td>
<td>it</td>
</tr>
<tr>
<td>2</td>
<td>be</td>
<td>12</td>
<td>for</td>
</tr>
<tr>
<td>3</td>
<td>to</td>
<td>13</td>
<td>not</td>
</tr>
<tr>
<td>4</td>
<td>of</td>
<td>14</td>
<td>on</td>
</tr>
<tr>
<td>5</td>
<td>and</td>
<td>15</td>
<td>with</td>
</tr>
<tr>
<td>6</td>
<td>a</td>
<td>16</td>
<td>he</td>
</tr>
<tr>
<td>7</td>
<td>in</td>
<td>17</td>
<td>as</td>
</tr>
<tr>
<td>8</td>
<td>that</td>
<td>18</td>
<td>dd</td>
</tr>
<tr>
<td>9</td>
<td>have</td>
<td>19</td>
<td>do</td>
</tr>
<tr>
<td>10</td>
<td>l</td>
<td>20</td>
<td>at</td>
</tr>
</tbody>
</table>
The **Teacher's Word Book** is an alphabetical list of the 10,000 words which are found to occur most widely in:

- 625,000 words from literature for children
- 3,000,000 words from the Bible and English classics
- 300,000 words from elementary-school text books
- 50,000 words from books about cooking, sewing, farming, the trades, and the like;
- 90,000 words from the daily newspapers

( Forty-one different sources were used)
نمودارهای لی به‌کاربردنی ووشکان له ووتاری کوردی

له

بی

غ

ووشکان

پرورش

پروردگار

کورس

که

فرشته‌هاگان

پنیا

پتر

روشکان

0

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17
نموداری که به کار بردنی ووشکان له ووتاری کوردی به کار بردنی ووشکان
Other application

- Spell checker
- Thesauri (e.g. word web)
- Crossword
- Unlimited application
Future work

• Extend the current work to a comprehensive API
  1. Number of lines in a text
  2. Number of paragraphs
  3. The longest and the shortest line or paragraph
  4. The average length
  5. Remove double space,
A course on natural language processing and Computational Linguistic

- Phonetics and Phonology — knowledge about linguistic sounds
- Morphology — knowledge of the meaningful components of words
- Syntax — knowledge of the structural relationships between words
- Semantics — knowledge of meaning
- Pragmatics — knowledge of the relationship of meaning to the goals and intentions of the speaker
- Discourse — knowledge about linguistic units larger than a single utterance
Thanks