Hausaufgabe

Abgabe bis 15.12. um 14 Uhr (diesmal bitte elektronisch) (BN: nur Aufgabe 3 und 4)

Aufgabe 1:

ELIZA: Das Skript "`eliza.pl`" ist ein simples Eliza-Programm. Überlegen Sie sich, wie man das Programm verbessern könnte (z.B. durch die Hinzunahme weiterer Antwortmöglichkeiten), und nehmen Sie mindestens eine Verbesserung vor. Bitte dokumentieren Sie in ihrem Programm, worin Ihre Verbesserung besteht und wodurch Sie sie erreichen.

`eliza.pl`

```
#!perl
# Mini Eliza

print "Hello, how are you? \n";
my $text=<STDIN>;
chomp($text);

while($text=~/.+/){
   $text =~ s/\bI('m| am)\b/you are/gi;
   $text =~ s/\bmy\b/your/gi;
   $text =~ s/\bme\b/you/gi;
   $text =~ s/\bmine\b /yours/gi;
   $text =~ s/(.*).$/$1/;
   print "MY_ELIZA: $text?\n";
   if($text =~ /.*you are (depressed|sad).*/) {
      my $random = int(rand(10));
      if($random < 5){
         $text =~ s/.*you are (depressed|sad).*/MY_ELIZA: I am sorry to hear you are $1./i;
      } else {
         $text =~ s/.*you are (depressed|sad).*/MY_ELIZA: What makes you feel $1?/i;
      }
   } else {
      $text =~ s/.*you are (.*)/MY_ELIZA: Why do you think you are $1\?/i;
   }
   $text =~ s/.*all.*/MY_ELIZA: In what way\?/i;
   $text =~ s/.*always.*/MY_ELIZA: Can you think of a specific example\?/i;
   if($text !~ /^MY_ELIZA/){
      $text= "MY_ELIZA: Please tell me more.";
   }
   print "$text\n\n";
   $text=<STDIN>;
   chomp($text);
}
```

Aufgabe 2:

Tokenizer: Das Skript "`tokenizer.pl`" ist ein simpler Tokenizer. Erstellen Sie eine beliebige Textdatei mit Endung `.txt` (unter `http://www.gutenberg.org/catalog/` finden Sie zahlreiche literarische Texte, die Sie verwenden können). Starten Sie "`tokenizer.pl`" und geben Sie den Namen Ihrer Textdatei ein, wenn Sie nach dem Dateinamen gefragt werden. Betrachten Sie den generierten Output "`token_$name.txt`". Wie könnte der Tokenizer verbessert werden (z.B. bessere Erkennung von Punkten am Satzende)? Nehmen Sie mindestens eine Verbesserung vor. Bitte dokumentieren Sie in ihrem Programm, worin Ihre Verbesserung besteht und wodurch Sie sie erreichen.

tokenizer.pl

```
#!perl -w
# Tokenizer

print "Dateiname:";
my $name = <>;
chomp($name);

open(INPUT,"<$name.txt");
open(OUTPUT,">token_$name.txt");

while(my $line=<INPUT>){
  $line =~s/ü/ue/g; # Ersetzen von Umlauten und ß
  $line =~s/ö/oe/g;
  $line =~s/ä/ae/g;
  $line =~s/Ü/Ue/g;
  $line =~s/Ö/Oe/g;
  $line =~s/Ä/Ae/g;
  $line =~s/ß/ss/g;
  $line =~s/[^A-Za-z0-9\.\!\?\s]//g; # Löschen aller "Sonderzeichen"
  $line =~s/\./\n\[SATZENDE\]\n/g;  # Tag für Satzende
  $line =~s/\!/\n\[SATZENDE\]\n/g;
  $line =~s/\?/\n\[SATZENDE\]\n/g;
  $line =~s/[0-9]+([\.|,][0-9]+)*/\n\[ZAHL\]\n/g; # Tag für Zahlen (Datum, Dezimalzahl,...)
  $line =~s/([A-Za-z]+)/\L$1/g;  # Kleinschreibung aller Wörter
  $line =~s/[\s]+/\n/g;     # Leerzeichen werden zu Zeilenumbrüche
  print OUTPUT $line;         # schreibt Tokenliste in token_$name.txt
}

close(OUTPUT);
close(INPUT);
```

Aufgabe 3:

Password: Da diese Aufgabe zuvor leider irreführend gestellt war, hier noch einmal:
Erstellen Sie Perl-Regexe für die folgenden Aufgaben:

1. Überpfüfung, ob ein Password aus mindestens 6 Zeichen besteht.
2. Überpfüfung, ob ein Password mindestens 2 Buchstaben enthält.
3. Überpfüfung, ob ein Password mindestens einen Großbuchstaben enthält.
4. Überpfüfung, ob ein Password mindestens einen mindestens 1 Ziffer enthält.

Aufgabe 4:

Wir haben gesehen, daß in Texten zumeist wenige Wörter häufig und viele Wörter selten vorkommen. Wählen Sie einen Text von `http://www.gutenberg.org/catalog/` und speichern Sie ihn mit der Endung `.txt`. Laden Sie das Skript `textstat.pl` von der geschützen Bereich der Kursseite herunter. Starten Sie das Skript und lesen Sie ihren Text ein. Betrachten Sie die Dateien `stat_$name.txt` und `frequ_$name.txt`. Welche für den Text zentralen Wörter finden sich unter den 20 häufigsten? Wieviel Prozent aller Typen kommen nur einmal vor? Welchen Anteil haben die drei häufigsten Typen am Gesamttext? Erläutern Sie, welche Auswirkung diese Verteilung für computerlinguistische Aufgaben hat. (Insgesamt nicht mehr als 300 Wörter)

Aufgrund des Umfangs der Aufgaben ist der Abgabetermin auf den 15.12. verschoben worden. Beachten Sie aber bitte, dass in der kommenden Sitzung eine kleinere Hausaufgabe hinzukommen wird, fangen Sie daher mit der Bearbeitung frühzeitig an