Feedback — Text Processing and Edit Distance

Help

You submitted this quiz on **Tue 20 Mar 2012 6:14 AM PDT**. You got a score of **5.00** out of **5.00**.

Question 1

Which of the following strings doesn't match the regex:

`/[a-z]+'[a-z]+/`

Your Answer		Score	Explanation
cat's			
rock 'n' roll	✔	1.00	Correct
won't			
we're			
Total		1.00 / 1.00	

Question Explanation

Question explanation

Question 2

Download either the Java or Python implementations of the porter stemmer from:

http://tartarus.org/martin/PorterStemmer/

Any implementation should work, but only Python and Java have been tested. Stem the following excerpt from Wikipedia:

```
In linguistic morphology and information retrieval, stemming is the process for re
ducing inflected words to their stem, base or root form-generally a written word f
orm. The stem need not be identical to the morphological root of the word; it is u
```

```
sually sufficient that related words map to the same stem, even if this stem is no
t in itself a valid root. Algorithms for stemming have been studied in computer sc
ience since 1968. Many search engines treat words with the same stem as synonyms a
s a kind of query broadening, a process called conflation.
```

Feel free to download a raw text file of the excerpt here.

You entered:

in linguist morpholog and inform retriev, stem is the process for reduc inflect word to their stem, base or root form-gener a

Your Answer		Score	Explanation
in linguist morpholog and inform retriev, stem is the process for reduc inflect word to their stem, base or root form-gener a written word form. the stem need not be ident to the morpholog root of the word; it is usual suffici that relat word map to the same stem, even if thi stem is not in itself a valid root. algorithm for stem have been studi in comput scienc sinc 1968. mani search engin treat word with the same stem as synonym as a kind of queri broaden, a process call conflat.	✔	1.00	Correct!
Total		1.00 / 1.00	

Question Explanation

To use the java file:

```
$ wget http://spark-university.s3.amazonaws.com/stanford-lang2info/quiz/w1_redo/s
tem.txt
$ wget http://tartarus.org/martin/PorterStemmer/java.txt
$ mv java.txt Porter.java
$ javac Porter.java
$ java Porter stem.txt > stem.stemmed.txt
$ cat stem.stemmed.txt
```

to use the python version:

```
$ wget http://spark-university.s3.amazonaws.com/stanford-lang2info/quiz/w1_redo/s
tem.txt
$ wget http://tartarus.org/martin/PorterStemmer/python.txt
$ mv python.txt porter.py
$ python porter.py stem.txt > stem.stemmed.txt
$ cat stem.stemmed.txt
```

Question 3

First tokenize the following text on all non-alphabetical characters and count the number of unique word types. How many word types disappear if we do case normalization?

```
The AU was originally defined as the length of the semi-major axis of the Earth's
elliptical orbit around the Sun. In 1976 the International Astronomical Union revi
sed the definition of the AU for greater precision, defining it as that length for
 which the Gaussian gravitational constant (k) takes the value 0.017 202 098 95 wh
en the units of measurement are the astronomical units of length, mass and time.[5
][6][7] An equivalent definition is the radius of an unperturbed circular Newtonia
n orbit about the Sun of a particle having infinitesimal mass, moving with an angu
lar frequency of 0.017 202 098 95 radians per day,[2] or that length for which the
 heliocentric gravitational constant (the product GM) is equal to (0.017 202 098 9
5)2 AU3/d2. It is approximately equal to the mean Earth-Sun distance.
```

Feel free to download a raw text file of the excerpt here.

Take special care to avoid treating two consecutive non-alphabetical characters as delimiting an empty token (see the `-s` flag for `tr`)

You entered:

4

Your Answer		Score	Explanation
4	✔	1.00	Correct!
Total		1.00 / 1.00	

Question Explanation

First, download or copy the file:

```
$ wget http://spark-university.s3.amazonaws.com/stanford-lang2info/quiz/w1_redo/au.txt
```

Then, for the BSD version of `tr` the following code should work:

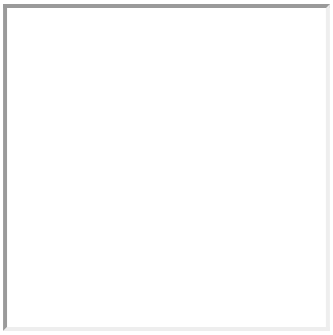
```
$ tr -sc '[:alpha:]' '\n'
-

$ tr -sc '[:alpha:]' '\n'
In general the manual page is your best friend for problems like these. At a unix-like
```

command-line type:

```
$ man NAME-OF-PROGRAM
```

Question 4

What is the value in the marked cell of the following Levenshtein distance (substitutions cost 2) table below:



You entered:

2

Your Answer		Score	Explanation
2	✔	1.00	
Total		1.00 / 1.00	

Question Explanation

Question 5

Consider the Levenshtein distance (substitutions cost 2) table below. What are the backpointers for

the highlighted cell?

[the highlighted cell corresponds to the letter 'a' in 'cave' and the 'a' in 'vase']

Your Answer		Score	Explanation
○ diagonal, left (↙,←)			
◉ diagonal (↙)	✔	1.00	Correct
○ left, down (←,↓)			
○ diagonal, down (↙,↓)			
○ down (↓)			
○ diagonal, left, down (↙,←,↓)			
○ left (←)			
Total		1.00 / 1.00	

Question Explanation