Corpus Linguistics UNIX Assignment 2

Hand in on 14-11-2012

Name: _____

For each question, provide the appropriate command ($) and response (if needed).

1. Using the *echo* command and redirection, create a file *week* with one day per line: *monday*, *tuesday*,...,*sunday*.

$

$

$

$

$

$

$

2. Sort the *week* file in reverse order.

$

Response:

3. Using the *wget* command, download this file: '*http://alfclul.clul.ul.pt/crpc/curso/CL2012/files/ciencia.txt*'. This is a Portuguese file about science. Create a token/word list and save it to a file called 'tokens'. Create a type list and save it to a file called 'types'.

$

$

$

4. Extract lines 12-50 from *ciencia.txt* and save it into a file called 'middle'. All this in only one line!

$

5. Create a list of words sorted by the most frequent first.

$

What are the three most frequent words?

6. Create a list of words according to their rhyming order.

$

Give all words rhyming in 'gico'.

7. Create a list of bigrams and save it to 'bigrams.sorted'.

$

$

$

What is the most frequent bigram?

8. Find bigrams that appear more than 6 times.

$

Which bigrams?

9. Find the words that appear more than 30 times.

$

Which words?

10. Find the words which are considered 'hapax' in the corpus.

$

Give one such 'hapax' which has a letter 'z'.

11. Find all palindromes.

$

$

Give a palindrome which has three letters.

12. Count the number of words in 'ciencia.txt'.

$

How many words?

13. Create a list of words with only one syllable.

$

Give three such words with at least three letters.