Branch: master ▾ Find file Copy path

cs-35l / assignment2 / lab2.log


prithvikannan ran script again

d05e243 on Oct 14

1 contributor

Raw Blame History

107 lines (91 sloc) 4.51 KB

```
Laboratory:

export LC_ALL='C'
    I started by running this command to get to the standard C locale

sort /usr/share/dict/words > words
    This grabs the words file from the usr folder and makes a sorted version in
    my current directory

wget https://web.cs.ucla.edu/classes/fall19/cs35L/assign/assign2.html
    This command gets the html of the assign2 page

tr -c 'A-Za-z' '[\n*]' < assign2.html
    This command takes everything that is not an alphabetical character (a-z
    or A-Z) and replaces it with the new line character
tr -cs 'A-Za-z' '[\n*]'< assign2.html
    This command does the same thing as above, but series of new line
    characters are collapsed into one new line. This means that there are no
    longer large spaces.
tr -cs 'A-Za-z' '[\n*]' < assign2.html | sort
    This command orders the output of the above in alphabetical order.
tr -cs 'A-Za-z' '[\n*]' < assign2.html | sort -u
    The -u flag means to sort unique, meaning each word will only show up once
    even if it exists multiple times in the assign2 document
tr -cs 'A-Za-z' '[\n*]' < assign2.html | sort -u | comm - words
    This command compares the unique sorted list of words to the words file.
    The output is in three columns: column 1 is only in assign2.html, column 2
    is those only in words, and column 3 is those in both
tr -cs 'A-Za-z' '[\n*]' < assign2.html | sort -u | comm -23 - words
    This command uses the -23 flag of comm, which supress column 2 and 3,
    meaning it only displays the lines that are in assign2.html and not in the
    words file

wget https://www.mauimapp.com/moolelo/hwnwdshw.htm
    This command gets the html file for the hawaiian words

touch buildwords
    I created a file called buildwords which I am going to use to hold my
    script
grep -E '<td>.+</td>' $@
    I use this grep command with extended regex to remove the <td> tags
sed 's/<[^>]*>//g'
    I use the sed command to remove everything within html tags (< and >)
tr [:lower:] [:upper:]
    Then I use tr to make everything lowercase, since we want our dictionary
    to be only lowercase
sed "s/\`/\'/g"
    Now I replace the grave accent with apostrophe. I needed to use the double
    quotes since the since single quote (apostrophe) is part of my string
tr ',' '\n'| tr ' ' '\n'
    This command splices words on the space and comma characters by adding in
```

```
    the newline character
sed "/[^pk/'mnwlhaeiou]/d"
    This command removes all of non-hawaiian letters. As per the piazza post,
    the way for us to tell that a word is hawaiian is if it only contains
    hawaiian characters. Originally I tried to implement a system where I
    removed alternate words, but the piazza post confirmed that this method
    would not work for this file.
sed "/^$/d"
    This removes the whitespaces introduced in the earlier steps
sort -u
    This command alphabetizes the list of hawaiian words and removes duplicates
    with the -u flag

cat hwnwdshw.htm | ./buildwords | less > hwords
    I piped the output to a file hwords which will hold all of my hawaiian
    words

tr [:upper:] [:lower:] <assign2.html | tr -cs "A-Za-z'" '[\n*]' | sort -u |
comm -23 - hwords > hmispelled
    This command finds all of the maximal nonempty sequences of ASCII letters
    or apostrophes and checks if they are not in the hawaiian dictionary

tr [:upper:] [:lower:] <assign2.html | tr -cs "A-Za-z'" '[\n*]' | sort -u |
comm -23 - hwords | wc -w
    The number of non hawaiian words is 554.

tr [:upper:] [:lower:] <assign2.html | tr -cs "A-Za-z'" '[\n*]' | sort -u |
comm -23 - words > emispelled
    This command does the same as above, but checks if they are not in the
    english dictionary

tr [:upper:] [:lower:] <assign2.html | tr -cs "A-Za-z'" '[\n*]' | sort -u |
comm -23 - words | wc -w
    The number of non english words is 65.

comm emispelled hmispelled
    This command compares the words that are marked mispelled in english with
    those mispelled in hawaiian

comm -13 emispelled hmispelled
    This command finds all of the words that are exclusively wrong by the
    hawaiian checker
    Two examples are: web and were

comm -13 emispelled hmispelled | wc -w
    This command counts the output of the above
    There are 492 words exclusively mispelled in hawaiian

comm -23 emispelled hmispelled
    This command finds all of the words that are exclusively wrong by the
    english checker
    Two examples are: lau and wiki

comm -23 emispelled hmispelled | wc -w
    This command counts the output of the above
    There are 3 words exclusively mispelled in english
```