Primeiro cod – criar pastas

#!/bin/bash

if [ ! -d "./tg1" ]; then //verifica se existe a pasta

mkdir ./tg1 //cria a pasta raiz e as próximas linhas cria as demais pastas

mkdir ./tg1/scripts

mkdir ./tg1/corpus

mkdir ./tg1/corpus\_txt

mkdir ./tg1/corpus\_info

mkdir ./tg1/words\_dic

mkdir ./tg1/sentences\_dic

echo "Directories created with success. //confirmação de conclusão.

Referência de uso – moodle primeiro exemplo de caso

Quinto cod – pares de palavras

#!/bin/bash

//usamos sed por ser um editor de fluxo, é uma ferramenta de linha de comando para manipular texto e arquivos. Mais eficaz que cat.

sed -e 's/[^[:alpha:]]/ /g' ./tg1/corpus\_txt/swmerged-ms98-a-trans.txt | sort -f | gawk -F, '{

//sed -e adicionamos o arquivo a ser trabalhado, 's/[^[:alpha:]]/ /g', substitui tudo que não for alfabeto por vazio, | sort -f, ordena e ignora sentenças em maiúsculas e minúsculas, gawk -F, inicia o programa awk com os determinador separadores.

for(i=0; i <=NF; i++){ //efetua a contagem de numero de campos atual, NF, embutido em parâmetro do wak.

j= i+1; // j ira receber o numero do parâmetro em i e soma-lo.

a[$i,$j]=$i" "$j; print a[$i,$j]; } //armazena i e j para escolha do que imprimir

}' | tr " " "#" | sort -d | uniq -c | tr "#" " " | gawk '{print $2 " " $3}' | sort -f | uniq -c -i >/tg1/words\_dic/words\_pairs.txt

// iniciamos traduzir os espaços para # usando o tr, em sort ordenamos com -d os alfanuméricos, uniq, deixar as únicas, sem repetições, -c, prefixa as linhas e efetua as contagens, voltamos em tr adicionar os espaços, imprimos as colunas sequenciais do armazenadas no (awk for), em sort ignorando a maiúsculas e minúsculas, dentro das palavras únicas também ignoramos as maiúsculas e minúsculas com -i e contamos, usando uniq -c -i, e enviamos o para o ficheiro words.

Referencia – vídeo no youtube relacionado em 6 – o documento wak também com modelos de caso.

Sexto cod – ficheiro de frases

sed -e 's/[^[:alpha:]]/ /g' ./tg1/corpus\_txt/swmerged-ms98-a-trans.txt | tr 'A-Z' 'a-z' | tr " " "|" | sort -d | uniq -c | gawk '

//sed -e adicionamos o arquivo a ser trabalhado, substituímos tudo que não é alfabético por vazio, usamos o tr para substituir tudo em maiúsculo para minúsculo, depois continuamos a substituir espaços por |, em sort ordenamos com -d os alfanuméricos, dentro das palavras únicas também ignoramos as maiúsculas e minúsculas e contamos, usando uniq -c.

{for(i = 0; i < NF; i++) { //efetua a contagem de numero de campos atual, NF, embutido em parâmetro do wak, enquanto houver sentença a ser contada.

print $i "";}}' >> ./tg1/sentences\_dic/sentences.txt

//imprimi a sentença awk armazenado anteriormente em i usando | de demais filtros e armazena a separação das frases no ficheiro sentences.

Referencia – vídeo youtube ralcionado do 4,5,7 – em documento [terminalroot.com.br/2015/07/30-exemplos-do-comando-sed-com-regex.html](https://terminalroot.com.br/2015/07/30-exemplos-do-comando-sed-com-regex.html)

Sétimo cod – pares de frases

#!/bin/bash

sed -e 's/[^[:alpha:]]/ /g' ./tg1/corpus\_txt/swmerged-ms98-a-trans.txt | sort -f | gawk -F,

//sed -e adicionamos o arquivo a ser trabalhado, substituímos tudo que não é alfabético por vazio, |, em sort ordenamos os alfanuméricos com -f ignoramos o que for maiúscula e minuscula, gawk -F, inicia o programa awk com os determinador separadores.

'{for(i=0; i <NF; i++) //efetua a contagem de numero de campos atual, NF, embutido em parâmetro do wak, enquanto houver sentença a ser contada.

{ j= i+1; // j ira receber o numero do parâmetro em i e soma-lo.

a[$i,$j]=$i" "$j; //armazena i e j conforme separação dos parâmetros

print a[$i,$j];} //armazena i e j e imprimir

}' | sort -d | uniq -c | tr " " "|" | sed -e 's/||/ /g' > ./tg1/sentences\_dic/sentence\_pairs.txt

// em sort ordenamos com -d os alfanuméricos, dentro das palavras únicas também ignoramos as maiúsculas e minúsculas e contamos, usando uniq -c. tr para reescrever os espaços, e terminamos com o sed -e que faz a captura do arquivo novamente substituindo o que estiver || para a devida ordenação em pares e armazena no arquivo sentences pairs.

Referencia – vídeo 2,7,8,9 – documento [rexegg.com/regex-quickstart.html](https://www.rexegg.com/regex-quickstart.html) - livrosdelinux.com.br/respostas-livro-shell-linux/

Oitavo cod – palavras 250.000

#!/bin/bash

cat ./tg1/corpus\_txt/swmerged-ms98-a-trans.txt | egrep -o '\w+' | sort -f | uniq -c -i >

//utilizamos o cat para exibir o arquivo unificado, filtramos o texto com egrep, -o, mostramos as partes da linhas //coincidentes, ‘w’ corresponde a zero ou mais caracteres, em sort ordenamos os alfanuméricos com -f ignoramos o //que for maiúscula, dentro das palavras únicas também ignoramos as maiúsculas e minúsculas com -i e contamos, //usando uniq -c -i, enviamos para o ficheiro.

./tg1/words\_dic/words\_8.txt sed -n '3,250002p' ./tg1/words\_dic/words\_8.txt >

//após o envio da primeira diretriz para o ficheiro usamos o sed para capturar e -n para suprimir a impressão automática, conta e separa as palavras iniciando dentro do intervalo, adiciona para o ficheiro words\_8.

./tg1/words\_dic/words\_8\_temp.txt mv ./tg1/words\_dic/words\_8\_temp.txt ./tg1/words\_dic/words.txt rm ./tg1/words\_dic/words\_8.txt

//mv, movemos o conteúdo do arquivo para words e removemos o ficheiro temporario

sed -e 's/[^[:alpha:]]/ /g' ./tg1/corpus\_txt/swmerged-ms98-a-trans.txt | sort -f | gawk -F,

//sed -e adicionamos o arquivo a ser trabalhado, substituímos tudo que não é alfabético por vazio, |, em sort ordenamos os alfanuméricos com -f ignoramos o //que for maiúscula, gawk -F, inicia o programa awk com os determinador separadores.

'{ for(i=0; i <=NF; i++) //efetua a contagem de numero de campos atual, NF, embutido em parâmetro do wak, enquanto houver sentença a ser contada.

{ j= i+1; // j ira receber o numero do parâmetro em i e soma-lo.

a[$i,$j]=$i" "$j; //armazena i e j conforme separação dos parâmetros

print a[$i,$j]; }//armazena i e j e imprimir

}' | tr " " "#" | sort -d | uniq -c | tr "#" " " | gawk '{print $2 " " $3;}' | sort -f | uniq -c -i >

./tg1/words\_dic/words\_pairs\_8.txt

//traduzimos os espaços para # usando o tr, ordenamos utilizando os alfanuméricos, dentro das palavras únicas contamos, usando uniq -c, voltamos os espaços com o tr, usamos o padrão do wak, divide suas linhas de entrada em espaço em branco e salva cada campo como $1, $2 e imprimos os valores quenos interessa, estão no 2 e3, em sort ordenamos os alfanuméricos com -f ignoramos o //que for maiúscula, dentro das palavras únicas também ignoramos as maiúsculas e minúsculas com -i e contamos, //usando uniq -c -i.

sed -n '3,250002p' ./tg1/words\_dic/words\_pairs\_8.txt > ./tg1/words\_dic/words\_pairs\_8\_temp.txt

// sed -n adicionamos o arquivo a ser trabalhado e suprime a impressão automática para ser usado posteriormente, considera o intervalo a ser salvo, com as devidas separações, adiciona para arquivo temporario

mv ./tg1/words\_dic/words\_pairs\_8\_temp.txt ./tg1/words\_dic/words\_pairs.txt

//move o conteúdo para o ficheiro que deve estar o ficheiro final

rm ./tg1/words\_dic/words\_pairs\_8.txt

//remove o ficheiro temporario

sed -e 's/[^[:alpha:]]/ /g' ./tg1/corpus\_txt/swmerged-ms98-a-trans.txt | tr 'A-Z' 'a-z' | tr " " "|" | sort -d | uniq -c | gawk '

//sed -e adicionamos o arquivo a ser trabalhado, substituímos tudo que não é alfabético por vazio, usamos o tr para substituir tudo em maiúsculo para minúsculo, depois continuamos a substituir espaços por |, em sort ordenamos com -d os alfanuméricos, dentro das palavras únicas contamos, usando uniq -c, com o gwak capturamos o resultado anterior e fazemos a sequencia abaixo.

{ for(i = 0; i < NF; i++)//efetua a contagem de numero de campos atual, NF, embutido em parâmetro do wak, enquanto houver sentença a ser contada.

{ print $i ""; }}' >> ./tg1/sentences\_dic/sentences\_8.txt //imprime o resultado e armazena em sentences.

sed -n '18,2500017p' ./tg1/sentences\_dic/sentences\_8.txt > ./tg1/sentences\_dic/sentences\_8\_temp.txt

// sed -n adicionamos o arquivo a ser trabalhado e suprime a impressão automática para ser usado posteriormente, considera o intervalo a ser salvo 18 a 2500017,paragrafo, com as devidas separações, adiciona para arquivo temporário.

mv ./tg1/sentences\_dic/sentences\_8\_temp.txt ./tg1/sentences\_dic/sentences.txt //move o conteúdo para ficheiro final

rm ./tg1/sentences\_dic/sentences\_8.txt //remove o ficheiro temporario

sed -e 's/[^[:alpha:]]/ /g' ./tg1/corpus\_txt/swmerged-ms98-a-trans.txt | sort -f | gawk -F, '

//sed -e adicionamos o arquivo a ser trabalhado, substituímos tudo que não é alfabético por vazio, |, em sort ordenamos os alfanuméricos com -f ignoramos o que for maiúscula, gawk -F, inicia o programa awk com os determinador separadores.

{ for(i=0; i <NF; i++) //efetua a contagem de numero de campos atual, NF, embutido em parâmetro do wak, enquanto houver sentença a ser contada.

{ j= i+1; // j ira receber o numero do parâmetro em i e soma-lo.

a[$i,$j]=$i" "$j; //armazena i e j conforme separação dos parâmetros

print a[$i,$j]; } //armazena i e j e imprimir

}' | sort -d | uniq -c | tr " " "|" | sed -e 's/||/ /g' > ./tg1/sentences\_dic/sentence\_pairs\_8.txt

// ordenamos utilizando os alfanuméricos, tiramos as repetições com uniq e contamos, substituímos os espaços por |, capturamos novamente o conteúdo do documento, com -e adicionamos os scripts aos comandos, substuimos os espações duplos com || para vazio.

sed -n '10,250009p' ./tg1/sentences\_dic/sentence\_pairs\_8.txt > ./tg1/sentences\_dic/sentence\_pairs\_8\_temp.txt

// sed -n adicionamos o arquivo a ser trabalhado e suprime a impressão automática para ser usado posteriormente, considera o intervalo a ser salvo 10 em diante, ,paragrafo, com as devidas separações, adiciona para arquivo temporário.

mv ./tg1/sentences\_dic/sentence\_pairs\_8\_temp.txt ./tg1/sentences\_dic/sentence\_pairs.txt

//movemos o conteúdo do arquivo para o ficheiro final.

rm ./tg1/sentences\_dic/sentence\_pairs\_8.txt

//removemos o ficheiro temporário.

Referencia - livrosdelinux.com.br/respostas-livro-shell-linux/ - [livrosdelinux.com.br/tabela-de-comandos/](https://www.livrosdelinux.com.br/tabela-de-comandos/) - vivaolinux.com.br