IPBeja Escola Superior de Tecnologia e Gestão

Escola Superior de Tecnologia e Gestão

Curso de Engenharia Informática

Sistemas Operativos

Teste 1 – Parte Prática III

(30 de novembro de 2023)

Informação sobre a estrutura das respostas e entrega do teste

Antes de iniciar o desenvolvimento dos scripts, crie uma diretoria denominada t1. Dentro desta diretoria, crie as subdiretorias alineaA, alineaB e alineaC. Desenvolva a solução para cada uma das alíneas na respetiva sub-diretoria. ***Comente o código para explicar a lógica da solução.*** *Em cada uma das subdiretorias deve incluir um ficheiro de texto denominado resposta.txt onde indica como deve ser chamado o script, ou scripts, que respondem ao solicitado naquela alínea. No final crie um ficheiro zip da diretoria t1, que se pode denominar t1.zip, e submeta o ficheiro através da ligação disponibilizado no moodle.*

(10 valores)

No desenvolvimento de sistemas de linguagem natural, como é o caso da predição de palavras no Eugénio, é comum dividir o *corpus* em duas partes, a parte de treino e a parte de teste. A parte de treino do *corpus* é utilizada para o treino dos modelos de língua, que por exemplo no caso do trabalho de grupo 1 são as ocorrências das palavras e as ocorrências dos pares de palavras. A parte de teste do *corpus* é utilizada para avaliar o desempenho do sistema. No caso do Eugénio o corpus de teste poderia ser utilizado para avaliar a capacidade de predição sistema.

Nesta questão pretende-se desenvolver um utilitário que divida um *corpus* em duas partes, a parte de treino e a parte de teste. A percentagem do *corpus* que deverá ser utilizada para treino e para teste deve ser definida pelo utilizador. Percentagens comuns para esta divisão são respetivamente 80% e 20%. Para desenvolver este

utilitário responda às seguintes questões utilizando como *corpus* o ficheiro fornecido nas aulas *frasesPublico10000.txt*.

(4 valores)

a) Nesta alínea deve desenvolver três *scripts awk*. O primeiro script deve calcular o total de linhas do *corpus (get_number_lines.awk)*. O segundo *script* deve apresentar no ecrã uma determinada percentagem de linhas da parte inicial do *corpus (get_first_lines.awk)*. O terceiro *script* deve apresentar no ecrã uma determinada percentagem de linhas da parte final do *corpus (get_last_lines.awk)*. Estes dois últimos scripts devem receber como parâmetros o total de linhas do ficheiro e a percentagem de linhas a apresentar no ecrã. Como exemplo, apresenta-se a seguir a chamada do segundo script. O script é chamado com a indicação que o corpus tem *10000* linhas (*total_lines*) e que se pretende mostrar no ecrã *80%* (percentage) da parte inicial do corpus.

```
gawk -v total_lines=10000 -v percentage=0.8 -f
get_first_lines.awk ../frasesPublico10000.txt
```

(3 valores)

b) Desenvolva um *script da shell* que recorre aos *scripts awk* anteriores para dividir o *corpus* em duas partes, definidas pelas percentagens definidas. A seguir apresenta-se um exemplo de chamada deste *script*. Com esta chamada do *script* pretende-se dividir o *corpus (frasesPublico10000.txt)* em duas partes, uma com *80%* das linhas iniciais do corpus e a outra com *20%* das últimas linhas do corpus, deverão ser armazenadas nos ficheiros *treino.txt* e *teste.txt*, respetivamente.

```
split_corpus.sh ../frasesPublico10000.txt 0.8 0.2 treino.txt
teste.txt
```

(3 valores)

c) Reprograme o *script* desenvolvido na alínea b utilizando apenas a programação da *shell*, e por conseguinte, sem utilizar os *scripts awk* desenvolvidos na alínea a). O *awk*

apenas poderá ser utilizado para operações simples, como por exemplo a escrita de uma determinada coluna de dados. Para obter a parte inicial e parte final do corpus sugere-se a utilização dos comandos *head* e *tail*, respetivamente.

Luís Garcia