

Escola Superior de Tecnologia e Gestão de Beja  
Curso de Engenharia Informática

## **Sistemas Operativos**

### **Trabalho de Grupo N.º 1**

*Scripts para a Geração de Dicionários do Eugénio para a  
Língua Italiana*

Luís Garcia

## Scripts para Geração de Dicionários do Eugénio

Neste trabalho pretende-se que os alunos desenvolvam um conjunto de *scripts* que permitam a criação de dicionários do Eugénio V3 para a língua italiana. Além disso também se pretende que os alunos desenvolvam scripts que forneçam informações sobre os textos utilizados (corpus), e os dicionários criados, como por exemplo o seu número de palavras, o tamanho médio das palavras, entre outras estatísticas. O corpus utilizado neste trabalho será o corpus de italiano criado com textos da web (paisa.raw.utf8.gz) que pode ser obtido [nesta página](#). Na Figura 1 apresenta-se esta página e o link onde pode obter este corpus. Depois de descarregar e descompactar o corpus, o ficheiro que lhe interessará para este trabalho é o ficheiro paisa.raw.utf8.

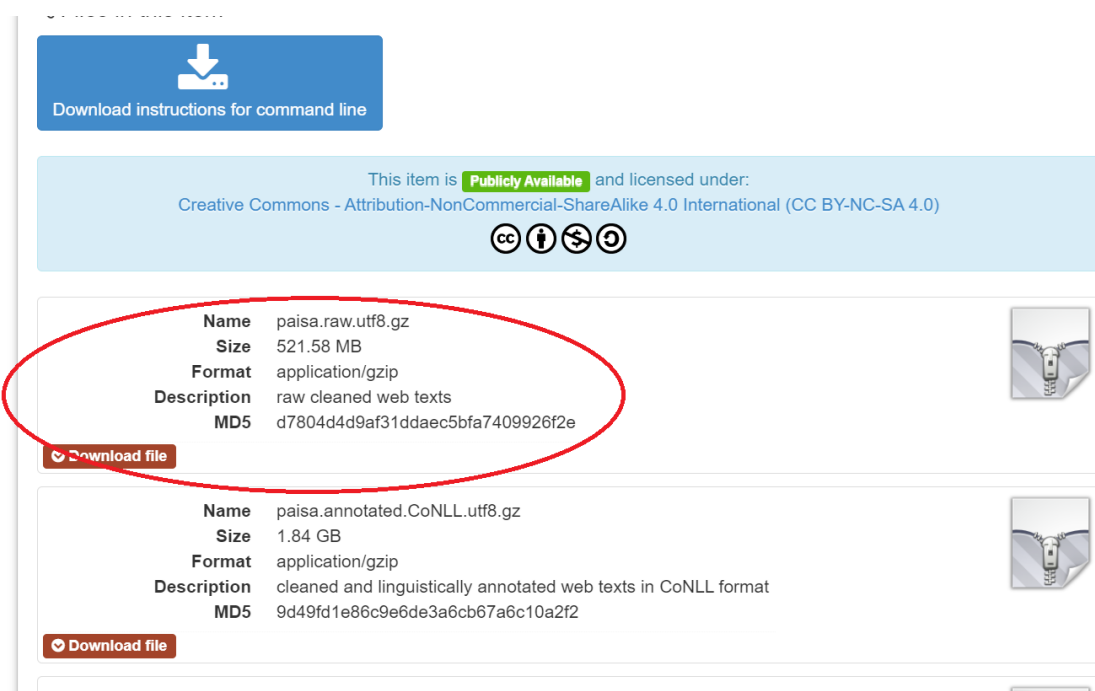


Figura 1 -Página do Corpus Italiano.

O Eugénio V3 é uma evolução do nosso sistema anterior, o Eugénio V2. Além de predição de palavras o Eugénio V3 também realiza predição de frases. O Eugénio V3 está disponível [neste sítio web](#). Para nos ajudar

a testar o sistema, gostaríamos que descarregasse e os instaladores do Eugénio (versões 32 bits e 64 bits para o Português e o Inglês), e as instalasse no seu computador, para nos indicar se este ficou a funcionar corretamente.

Para a realização do trabalho pode utilizar qualquer uma das versões.

Na Figura 3 apresentamos o Eugénio V3 com um teclado de ecrã que oferece predição de palavras e predição de frases.

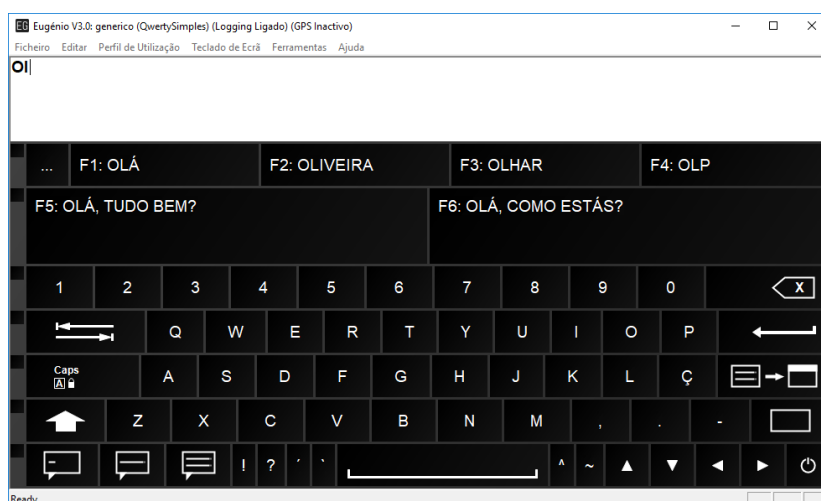


Figura 2 - Eugénio V3 com predição de palavras e predição de frases.

Para poder realizar a predição de palavras, e predição de frases, o Eugénio V3 recorre a um conjunto de dicionários com informação estatística sobre a utilização de palavras e frases no português.

Os dicionários das versões portuguesa e inglesa foram produzidos com base no processamento de uma coleção de textos do português (corpus de frases do Público) e uma coleção de textos do inglês (Brown corpus). Pretende-se agora que os alunos realizem o mesmo processo para a língua italiana.

Os dicionários utilizados na predição de palavras consistem em ficheiros de texto com os seguintes dados:

- Ficheiro de palavras – cada linha do ficheiro contém uma palavra e as suas ocorrências no corpus.
- Ficheiro de pares de palavras – cada linha do ficheiro contém um par de palavras e as suas ocorrências no corpus.

Os dicionários utilizados na predição de frases são semelhantes, mas em vez de palavras contêm frases:

- Ficheiro de frases – cada linha do ficheiro contém uma frase e as suas ocorrências no corpus.
- Ficheiro de pares de frases – cada linha do ficheiro contém um par de frases e as suas ocorrências no corpus.

Neste trabalho pretende-se que os alunos desenvolvam um conjunto de *scripts* que consigam gerar estes dicionários para a língua italiana. Também será solicitado aos alunos que desenvolvam um conjunto de *scripts* que forneçam informação sobre o corpus utilizado. Os requisitos mínimos a desenvolver estão assinalados neste enunciado.

*(Requisito Mínimo)*

1. Crie uma diretoria *tg1*, e dentro desta, crie as seguintes sub-diretorias: *scripts*, *corpus*, *corpus\_txt*, *corpus\_info*, *words\_dict*, *sentences\_dict*.

*(Requisito Mínimo)*

2. Obtenha o corpus italiano na página web fornecida. Este corpus tem mais de 8 milhões de linhas. Para este trabalho utilize apenas as primeiras 200.000 linhas do corpus. Deve guardar este sub-corpus na sub-diretoria *corpus\_txt*.

*(Requisito Mínimo)*

3. Para caracterizar o corpus utilizado desenvolva um script que calcula as seguintes métricas: número de caracteres, quantidade de linhas não vazias, número total de palavras, número total de palavras diferentes, o quociente entre o total de palavras diferentes e o total de palavras, número total de frases, o número total de frases diferentes, o quociente entre o total de frases diferentes e o total de frases. O script que realiza esta operação deve ser armazenado na diretoria scripts. O resultado do script deve ser armazenado na sub-diretoria *corpus\_info*, num ficheiro denominado *corpus\_info.txt*.

*(Requisito Mínimo)*

4. Desenvolva um script que cria o ficheiro de palavras. Este ficheiro deverá conter em cada linha, uma palavra e as suas ocorrências no corpus. O ficheiro deverá ficar ordenado de forma alfabética. O script que realiza esta operação deve ser armazenado na diretoria scripts. O resultado do script deve ser armazenado na sub-diretoria *words\_dict*, num ficheiro denominado *words.txt*.

*(Requisito Mínimo)*

5. Desenvolva um script que cria o ficheiro com os pares de palavras. Este ficheiro deverá conter em cada linha um par de palavras e as suas ocorrências no corpus. O ficheiro deverá ficar ordenado de forma alfabética. O script que realiza esta operação deve ser armazenado na diretoria scripts. O resultado do script deve ser armazenado na sub-diretoria *words\_dict*, num ficheiro denominado *words\_pairs.txt*.

*(Requisito Mínimo)*

6. Desenvolva um script que verifica se as palavras que formam os pares de palavras existem no ficheiro de palavras (*words.txt*).
7. Desenvolva um script que cria o ficheiro de frases. Este ficheiro deverá conter em cada linha uma frase e as suas ocorrências no corpus. Em cada frase, os espaços entre as palavras deverão ser substituídos pelo carater '|'. O ficheiro deverá ficar ordenado de forma alfabética. O script que realiza esta operação deve ser armazenado na diretoria scripts. O resultado do script deve ser armazenado na sub-diretoria *sentences\_dict*, num ficheiro denominado *sentences.txt*.
8. Desenvolva um script que cria o ficheiro com os pares de frases. Este ficheiro deverá conter em cada linha um par de frases e as suas ocorrências no corpus. Em cada frase, os espaços entre as palavras deverão ser substituídos pelo carater '|'. O ficheiro deverá ficar ordenado de forma alfabética. O script que realiza esta operação deve ser armazenado na diretoria scripts. O resultado do script deve ser armazenado na sub-diretoria *sentences\_dict*, num ficheiro denominado *sentences\_pairs.txt*.
9. Desenvolva um script que verifica se as frases que formam os pares de frases existem no ficheiro de frases (*sentences.txt*).
10. O Eugénio suporta apenas um máximo de 250.000 palavras e 250.000 frases. Através da modificação dos scripts anteriores, ou através de novos scripts, garanta que não existem nos ficheiros mais de 250.000 palavras e frases.
11. Desenvolva um script Windows que copie os ficheiros gerados para a diretoria do Eugénio. Se tiver instalado a versão 64 bits, a diretoria do Eugénio será C:\Program Files\Eugénio. No caso de ter instalado a versão de 32 bits a diretoria será C:\Program

Files (x86)\ Eugénio. Para a instalação dos dicionários em Italiano deve realizar as seguintes copias de ficheiros:

- words.txt -> geral.pal
- words\_pairs.txt -> geral.par
- sentences.txt -> geral.frs
- sentences\_pairs.txt -> geral.paf

Para experimentar o script deve desinstalar o Eugénio, e voltar a correr a instalação. Antes de chamar o Eugénio deve correr o script que altera os dicionários. Só depois deve chamar o Eugénio.

*(Requisito Mínimo)*

10. Desenvolva um relatório descrevendo as tarefas realizadas em cada uma das etapas do trabalho.

## **Avaliação**

Os trabalhos serão classificados com uma das seguintes notas (esta nota será adicionada à média dos dois testes):

- -0,5 – o trabalho não cumpre os requisitos mínimos, ou o aluno não conseguiu demonstrar que desenvolveu o trabalho
- 0 – o trabalho cumpre os requisitos mínimos mas não a totalidade dos requisitos, e o aluno conseguiu demonstrar que desenvolveu o trabalho
- +0,5 - o trabalho cumpre a totalidade dos requisitos e o aluno conseguiu demonstrar que desenvolveu o trabalho

## Grupos de Trabalho

O trabalho deve ser desenvolvido por grupos com um máximo de dois elementos. Não são permitidas alterações nos elementos do grupo salvo em situações extraordinárias e devidamente justificadas.

## Entrega do Trabalho

Os alunos devem entregar os scripts e um relatório que apresente a solução encontrada. No relatório devem ser apresentadas e explicadas as principais partes do código. Num anexo deste relatório devem ser fornecidos todos os scripts desenvolvidos. Para a entrega do trabalho os alunos devem compactar os vários ficheiros num único ficheiro (zip ou equivalente). O nome deste ficheiro deve conter a indicação TG1 (Trabalho de Grupo 1) e o número dos alunos que compõem o grupo. Por exemplo TG1\_2000\_3000.zip seria o Trabalho de Grupo 1 do grupo formado pelos alunos com os números 2000 e 3000. O trabalho deve ser entregue através do sistema *Moodle*. **Não serão consideradas soluções entregues por e-mail.**

## Apresentação do Trabalho

Na apresentação os alunos devem demonstrar conhecer a totalidade do trabalho e estar aptos a realizar modificações ao código apresentado de acordo com as indicações do docente. Esta apresentação será realizada num horário especialmente marcado para o efeito.

*Bom Trabalho*

*Luís Garcia*