Projeto de LI3 - Wikipedia
Grupo 69

Sérgio Jorge (A77730) Vítor Castro (A77870) Marcos Pereira (A79116)

20 de Abril de 2017

Resumo

Neste relatório vamos fazer uma análise acerca do trabalho realizado na primeira fase do projecto no âmbito da UC de Laboratórios de Informática III no qual surgiu o objetivo de desenvolver um programa tendo por base a linguagem de programação C. Além disso, o relatório apresenta também a solução encontrada pelo nosso grupo para o problema.

Conteúdo

1 Introdução

Este projeto foi realizado com o objetivo de construir um aplicativo que faça uma análise dos artigos que estão nos backups da Wikipedia, fornecidos pelos professores e que foram realizados em diferentes meses, para que seja possível determinar e descobrir informações úteis acerca desses mesmos backups. Assim, foram propostas pelos professores, 10 interrogações e tarefas computacionais, realizadas na linguagem de programação C, às quais o nosso programa e o nosso trabalho devem responder com sucesso. A realização das 10 tarefas permitiram por um lado, melhorar e consolidar os conhecimentos adquiridos nas UCs de Programação Imperativa, Algoritmos e Complexidade e Arquitetura de Computadores. Por outro, permitiram também a aprendizagem de processos para a correta resolução de problemas tal como a divisão do projeto em alguns módulos

afim de ser possível que todos os membros do grupo conseguissem ajudar à concretização do projeto de igual forma. Assim, de modo a facilitar a compreensão do projeto, o relatório está dividido da seguinte forma:

Secção 2 : Problema;

Secção 3 : Solução;

Secção 4 : Conclusão.

2 Problema

Neste projeto de LI3, é-nos pedido para a partir de backups da Wikipedia, fornecidos pelos professores, fazermos a leitura dos dados e a extração de informação que a equipa docente considera relevante. Assim, a informação que devemos gerar é:

1 - All articles
devolve o número de artigos encontrados nos backups analisados. Para esta interrogação, artigos duplicados em backups sucessivos e novas revisões de artigos também contam.

2 - Unique articles
devolve o número de artigos únicos (com id único) encontrados nos vários backups analisados. Artigos duplicados ou revisões dos mesmos que estejam presentes em backups distintos não são contabilizados.

3 - All revisions
pretende saber quantas revisoões foram efetuadas naqueles backups (o numero total de versoes diferentes encontradas nos backups). O valor retornado deve incluir quer a versai base do artigo bem como as revisões feitas ao mesmo.

4 - Top 10 contributors
devolve um array com os identificadores dos 10 autores que contribuíram para um maior numero de revisoes de artigos (i.e., contar contribuições para artigos e respetivas revisões). O resultado deve ser ordenado pelos autores com mais contribuições. Se existirem autores com o mesmo número de contribuições, o resultado deve apresentar primeiro os autores com um identificador menor.

5 - Contributor name
devolve o nome do autor com um determinado identificador.

6 - Top 20 largest articles
devolve um array com os identificadores dos 20 artigos que possuem textos com um maior tamanho em bytes. Para cada artigo deve ser contabilizado o maior texto encontrado nas diversas versões(revisões) do mesmo. O resultado deve ser ordenado pelos artigos com maior tamanho. Se existirem artigos com o mesmo tamanho, o resultado deve apresentar primeiro os artigos com um identificador menor

7 - Article title
devolve o título do artigo com um determinado identificador.

8 - Top N articles with more words
devolve um array com os identificadoeres dos N (passado como argumento) artigos que possuem textos com o maior numero de palavras. Resultado ordenado pelos artigos com maior numero de palavras.

9 - Titles with prefix
devolve um array de títulos de artigos que começam com um prefixo passado como argumento e o resultado deve ser ordenado por ordem alfabética.

10 - Article timestamp
devolve o timestamp para uma certa revisão de um artigo.

3 Solução

3.1 Estrutura de Dados

As estruturas de dados são bastante importantes no desenvolvimento do programa pois permite a transformação e armazenamento dos dados. Destaca-se neste trabalho as seguintes estruturas:

Hashtables : Essenciais ao longo de todo o projeto;

3.2 Implementação

3.2.1 All articles

3.2.2 Unique articles

3.2.3 All revisions

3.2.4 Top 10 contributors

3.2.5 Contributor name

3.2.6 Top 20 largest articles

3.2.7 Article title

3.2.8 Top N articles with more words

3.2.9 Titles with prefix

3.2.10 Article timestamp

3.3 Resultado Final

IMAGEM DO TERMINAAAAAAAAAAAAAAAAAAAAAAAAAAAAL

4 Conclusões

Este projeto serviu para aprofundarmos o conhecimento da linguagem C, assim como as bibliotecas que lhe estão associadas. Achámos que, com a realização de um trabalho deste tipo permite uma consolidação proveitosa da linguagem, não só em termos teóricos como também em termos práticos e acaba por ser uma forma diferente de cimentar os conteúdos de algumas das UCs que nos surgiram neste curso. Permite também melhorar as habilidades na resolução de problemas. Concluímos também que:

- Hoje em dia, é possível ler e manipular grandes quantidades de informação em poucos segundos;
- A separação do trabalho em módulos ajuda a dividir o trabalho pela equipa e também, neste contexto, ressalta-se também a utilidade que mais uma veja encontramos no "git";