

UNIVERSIDADE FEDERAL DE PERNAMBUCO – UFPE
CENTRO DE INFORMÁTICA – CIn

Relatório do projeto da disciplina:
Processamento de Cadeias de Caracteres (2015.2)

Equipe:
João Guilherme Farias Duda
Paulo de Barros e Silva Filho
Raul Maia Falcão

Recife, 10 de Janeiro de 2016

Conteúdo

| | | |
|----------|---|-----------|
| 1 | Introdução | 3 |
| 2 | Descrição de uso da ferramenta <i>ipmt</i> | 3 |
| 3 | Implementação | 4 |
| 3.1 | Descrição dos algoritmos implementados | 4 |
| 3.1.1 | Linear Suffix Array | 4 |
| 3.1.2 | Linear Suffix Tree | 4 |
| 3.1.3 | LZ78 | 5 |
| 3.2 | Detalhes de implementação | 5 |
| 3.2.1 | Linear Suffix Array | 5 |
| 3.2.2 | Linear Suffix Tree | 6 |
| 3.2.3 | LZ78 | 6 |
| 3.3 | Descrição do formato .idx | 7 |
| 4 | Experimentos | 7 |
| 4.1 | Como a nossa implementação do LZ78 se compara ao gzip? | 8 |
| 4.1.1 | Tempo | 8 |
| 4.1.2 | Taxa de compressão | 12 |
| 4.2 | Como a busca de padrão do <i>ipmt</i> se compara ao grep? | 14 |
| 4.3 | Como a ferramenta <i>ipmt</i> se compara ao codesearch? | 15 |
| 5 | Conclusão | 16 |

1 Introdução

Este documento é sobre a ferramenta *ipmt*. Essa ferramenta é capaz de pré-processar um arquivo de texto, gerando um índice. Sucessivas buscas podem ser feitas através desse índice, sem a necessidade de percorrer o texto novamente.

ipmt primeiro gera um índice para o texto usando o algoritmo LSA (Linear Suffix Array), depois esse índice é comprimido em um arquivo juntamente com o texto usando o algoritmo LZ78. Para realizar buscas, primeiramente o arquivo é descomprimido usando o LZ78-decode e depois o casamento de padrões é realizada de acordo com o LSA.

Os integrantes da equipe foram responsáveis pelas seguintes tarefas:

- João:
- Paulo: Realização de experimentos; implementação do algoritmo LZ78 e da interface de comunicação entre os algoritmos.
- Raul: Implementação das estruturas de indexação Linear Suffix Array (LSA) e Linear Suffix Tree (ST).

2 Descrição de uso da ferramenta *ipmt*

O projeto contém um arquivo Makefile. Após a execução do comando make, o executável *ipmt* será gerado no diretório bin. A ferramenta *ipmt* possui 4 modos de execução:

- Modo de indexação - *ipmt* index file.txt

O comando acima irá criar o arquivo file.idx, um arquivo comprimido que contém o conteúdo de file.txt e um índice para esse arquivo que possibilita a realização de buscas.

- Modo de busca - *ipmt* search -c herself file.idx

O comando acima irá listar a quantidade de ocorrência do padrão "herself" encontradas no arquivo indexado file.idx. O argumento -c é opcional, caso não informado todas as linhas contendo ocorrências serão impressas. O argumento -p pode ser utilizado para passar como parâmetro um arquivo contendo uma lista de padrões (um padrão por linha).

- Modo de compressão - *ipmt* compress file.txt
O comando acima irá comprimir o arquivo file.txt em um arquivo file.comp.
- Modo de descompressão - *ipmt* decompress file.comp
O comando acima irá descomprimir o arquivo file.comp em um arquivo file.comp.decomp.

Os dois últimos modos não foram pedidos na especificação do projeto, mas nós os criamos para facilitar a comparação do nosso algoritmo de compressão e descompressão com algoritmos existentes.

3 Implementação

Todos os algoritmos foram implementados em C++ por questões de eficiência. A implementação em Python feita em sala de aula foi usada como uma base inicial.

3.1 Descrição dos algoritmos implementados

3.1.1 Linear Suffix Array

O algoritmo de indexação implementado teve como base [2] para a construção em tempo linear de um array de sufixos. Em suma, o array de sufixos é um array de inteiros que armazena a permutação de n índices ordenados lexicograficamente, onde n é o tamanho do texto. Uma vez construído o array de sufixos, a complexidade da busca passa a ser linear com relação ao tamanho do padrão.

3.1.2 Linear Suffix Tree

O algoritmo de indexação implementado¹ teve como base [4] para a construção em tempo linear de uma árvore de sufixos. A estrutura implementada representa todos os sufixos de uma cadeia. A implementação contém alguns truques para que a construção seja feita em tempo linear. Um desses truques é adicionar aos nós suffix links, também chamados como transições de falha ou fronteiras. Devido ao alto consumo de memória ao gerar a árvore de sufixo,

¹Apesar de termos implementado o ST, não conseguimos o fazer de maneira eficiente, portanto não o integramos à interface *ipmt*. A implementação ST.cpp está disponível para que sejam realizadas buscas.

resolvemos deixar a feature de melhorar o gerenciamento de memória para o futuro. Por consequência não foi gerado os índices dos sufixos, mas existe a opção de busca exata retornando o número de ocorrências de um dado padrão. Segundo [1] se o núcleo da implementação for orientada a objeto, a árvore de sufixo apresenta efeitos indesejáveis de memória fragmentada.

3.1.3 LZ78

O algoritmo de compressão LZ78 teve como base a implementação vista em sala de aula e descrita em [3]. O LZ78 utiliza um dicionário dinâmico explícito, onde a referência compreende um par composto pelo índice no dicionário e o caractere de mismatch.

Durante a compressão (LZ78-encode), o dicionário é criado dinamicamente a cada mismatch. Junto com o dicionário, também é criado um código que representa a string que está sendo comprimida. O LZ78-encode é linear de acordo com o tamanho da string que está sendo comprimida.

O processo de descompressão (LZ78-decode) recebe somente o código gerado durante a compressão, e é capaz de gerar o dicionário dinamicamente, bem como a string original que foi comprimida. O LZ78-decode é linear de acordo com o tamanho do código recebido na entrada.

3.2 Detalhes de implementação

Abaixo descrevemos algumas decisões e peculiaridades de cada algoritmo.

3.2.1 Linear Suffix Array

Na construção do Linear Suffix Array há uma etapa de criação de dois arrays de sufixos, S1 e S2. Seja $index$ a posição de um caracter em um texto: A função `buildS1andS2` constrói o array de sufixo S1 que contém sufixos tal que $index \% 3 = 0$ e também constrói o array de sufixo S2 que contém sufixos tal que $index \% 3 \neq 0$. Após a construção de S1 e S2, estes são ordenados através de uma implementação do Radix Sort com o objetivo de otimizar essa etapa. Como o Radix Sort não faz comparações entre valores, nesse contexto, o seu desempenho é superior a um algoritmo de ordenação por comparação. A ordenação de S1 e S2 foi necessária para a etapa de merge ($S1 \cup S2 = SA$) de tal forma que o custo do merge é realizado em tempo linear. Após o merge, obtemos os índices devidamente ordenados.

3.2.2 Linear Suffix Tree

Inicialmente na construção do Linear Suffix Tree foi necessário criar um nó auxiliar (\perp) o qual possui transições de todas as letras do alfabeto para o nó inicial (root) que corresponde a uma cadeia vazia (ε). Após essa etapa, uma construção on-line é feita adicionando caracter por caracter a árvore através das funções *update* e *canonize*. A função *update* transforma a árvore na iteração anterior em uma árvore na iteração corrente inserindo transições do caracter corrente a ser adicionado. A função *update* utiliza a função *canonize* e a função *test_and_split* que testa se há ou não referência a um nó terminador. Ao final da função *update* é retornado a referência do par do nó terminador. Após adicionar todos os caracteres a árvore de sufixo está devidamente montada e pronta para realizar buscas de padrões exatos.

3.2.3 LZ78

O dicionário dinâmico possui uma estrutura de Trie: Cada nó mapeia um índice a somente um char, e possui nós descendentes de forma que o nó original e cada um de seus descendes forma uma sequência diferente de caracteres encontrada no texto.

Usamos como alfabeto do código de saída o sistema binário. Como cada elemento do código de saída tem somente um bit, usamos como estrutura de dados para guardar o código um vector de bool². Optamos por essa estrutura de dados pois ela possui uma otimização de espaço: Um bool em C++ ocupa 8 bits (1 byte) de memória, porém um vector de bool usa somente 1 bit para cada elemento.

Outra peculiaridade desse algoritmo é que pode ocorrer do arquivo descomprimido conter algum "lixo" no último byte. Isso acontece porque a escrita em arquivo só pode ser feita de byte em byte, porém o código gerado pelo LZ78-encode é uma sequência de bits. Caso o número de bits não seja um múltiplo de 8, o último byte precisa ser preenchido com uma sequência de 0s, o que pode alterar o último byte na hora da descompressão. Isso poderia ser contornado com alguma flag no início do arquivo comprimido, informando quantos bits devem ser descartados do último byte. Deixamos isso como trabalho futuro, pois verificamos através de experimentos que isso pouco afeta o desempenho e resultados do algoritmo.

²http://en.cppreference.com/w/cpp/container/vector_bool

3.3 Descrição do formato .idx

A ferramenta *ipmt* gera e lê arquivos no formato .idx. Esse arquivo é gerado da seguinte forma para um arquivo de entrada file.txt:

1. Primeiramente é gerado o Linear Suffix Array a partir do conteúdo de file.txt.
2. Após isso, conta-se o número de linhas de file.txt.
3. É criado um novo arquivo com o seguinte conteúdo:

```
[Número de linhas contidas em file.txt]
[Conteúdo de file.txt]
[Elementos do LSA separados por um espaço]
```

(Note que quebras de linha separam os elementos acima.)

4. Esse novo arquivo é então comprimido usando o LZ78, gerando o arquivo file.idx.

Na hora de ler o arquivo .idx, primeiro é realizada a descompressão. Através do resultado, a ferramenta sabe que a primeira linha contém o número de linhas do texto. Logo, as linhas seguintes são referentes ao LSA, que é utilizado pela busca.

4 Experimentos

Comparamos nossas implementações com o gzip³, grep⁴ e com o codesearch⁵. Realizamos experimentos para responder às seguintes perguntas:

1. Como a nossa implementação do LZ78 se compara ao gzip?
2. Como a busca de padrão do *ipmt* se compara ao grep?
3. Como a ferramenta *ipmt* se compara ao codesearch?

³<http://www.gzip.org/>

⁴<https://www.gnu.org/software/grep/>

⁵<https://github.com/google/codesearch>

Foram implementados scripts em BASH para controlar os experimentos e fazer medições, gerando arquivos .raw de saída contendo resultados. Foram também implementados scripts em R que desenham gráficos de acordo com os arquivos .raw gerados pelos scripts BASH.

Todos os experimentos foram realizados em uma máquina com processador Intel Core i5 2.6Ghz e 8Gb de RAM. Cada medição de tempo nos experimentos foi realizada 10 vezes, e somente a média foi considerada e reportada nos resultados. Todos os scripts e resultados estão disponíveis no diretório *experiments*.

4.1 Como a nossa implementação do LZ78 se compara ao gzip?

Nós comparamos a nossa implementação do LZ78 com o gzip em dois aspectos: Tempo e taxa de compressão. Para realizar essa comparação, primeiramente nós dividimos um arquivo⁶ que contém 1GB de texto em inglês em arquivos de tamanhos distintos: 100KB, 200KB, 300KB, 700KB, 1MB, 2MB, 3MB, 5MB, 50MB, 100MB, 200MB, 300MB e 500MB. Cada arquivo desse contém os primeiros n bytes do arquivo original, onde n é o tamanho do arquivo. Fizemos medições em relação ao tempo e à taxa de compressão para cada operação de compressão e descompressão levando em conta os algoritmos LZ78 e gzip.

Separadamente, repetimos também o mesmo experimento acima mas usando como entrada um arquivo⁷ de 200MB contendo sequências de proteínas. Esse arquivo também foi dividido em arquivos de tamanhos distintos: 100KB, 200KB, 300KB, 700KB, 1MB, 2MB, 3MB, 5MB, 50MB, 100MB, 200MB.

4.1.1 Tempo

Abaixo estão dois gráficos que relacionam o tempo que leva para comprimir um arquivo contendo com o tamanho dele, para ambos o nosso LZ78 e o gzip. O primeiro gráfico é referente aos arquivos contendo texto em inglês, enquanto o segundo aos arquivos contendo sequências de proteínas.

Note que os resultados foram muito parecidos apesar do conteúdo dos arquivos serem diferentes. Ambas as funções de cada gráfico são (ou se aproximam muito de) retas, o que comprova que a nossa implementação do LZ78, bem como o gzip, comprimem em tempo linear de acordo com

⁶ <http://pizzachili.dcc.uchile.cl/texts/nlang/english.1024MB.gz>

⁷ <http://pizzachili.dcc.uchile.cl/texts/protein/proteins.200MB.gz>

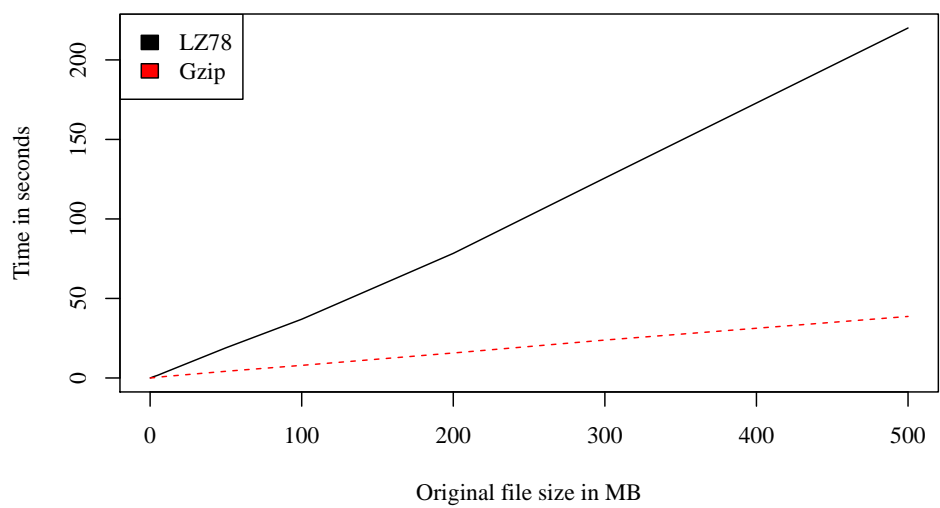


Figura 1: Compressão: Arquivos contendo textos em inglês

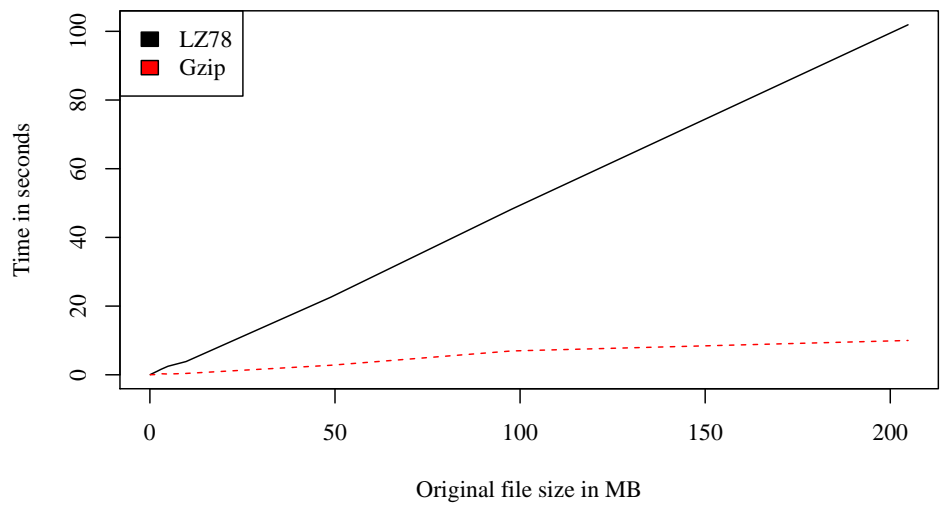


Figura 2: Compressão: Arquivos contendo sequência de proteínas

o tamanho do arquivo. Porém, a constante que multiplica a função do gzip é bem menor do que a nossa. Isso acontece porque o gzip está em desenvolvimento a mais de 23 anos, onde experts estão sempre otimizando o algoritmo, fazendo com que essa constante da função linear seja cada vez menor.

Os arquivos comprimidos acima foram descomprimidos com ambas as ferramentas LZ78 e gzip.

Abaixo estão dois gráficos que relacionam o tempo que leva para descomprimir um arquivo e o tamanho dele, um para os arquivos contendo texto em inglês e outro para arquivos contendo sequência de proteínas.

Obtivemos novamente um resultado consistente com o anterior: ambos possuem complexidade linear, porém a constante da descompressão do gzip é muito superior à da nossa implementação. Isso ficou ainda mais acentuado na descompressão do que na compressão.

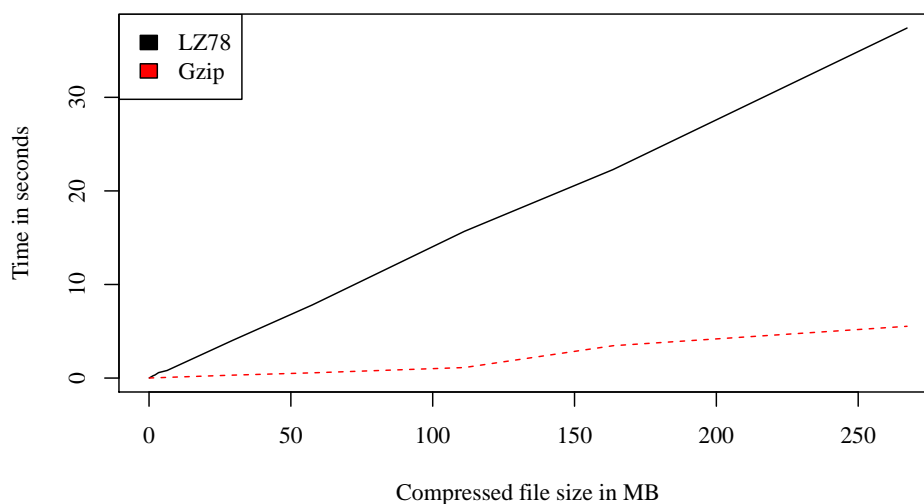


Figura 3: Descompressão: Arquivos contendo textos em inglês

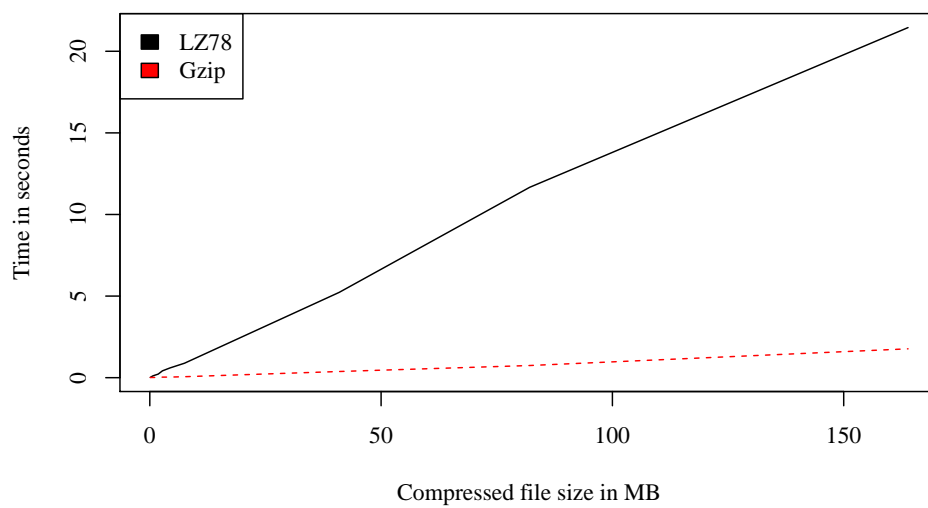


Figura 4: Descompressão: Arquivos contendo sequência de proteínas

4.1.2 Taxa de compressão

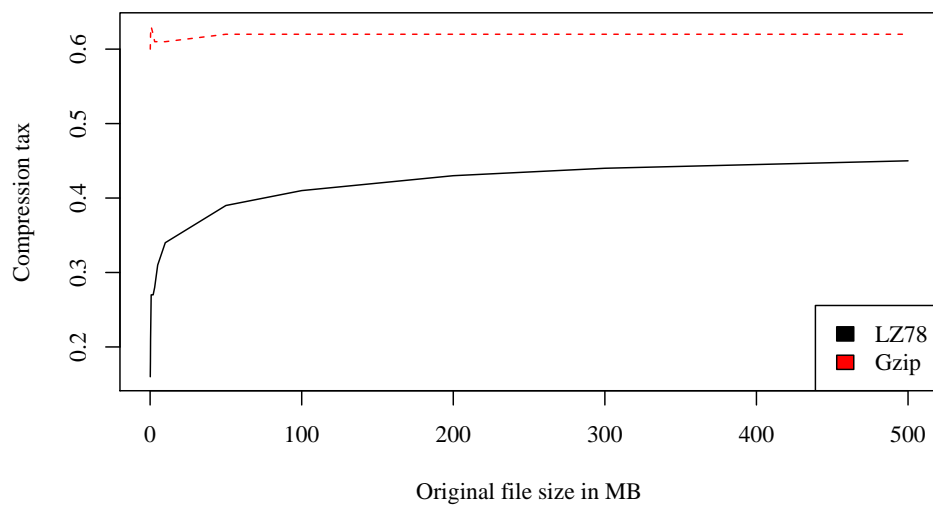


Figura 5: Taxa de compressão: Arquivos contendo textos em inglês

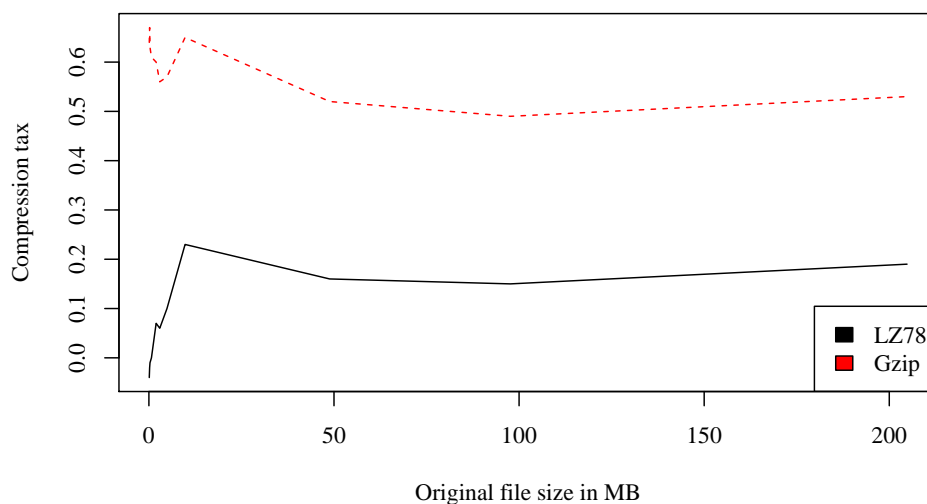


Figura 6: Taxa de compressão: Arquivos contendo sequência de proteínas

Acima estão dois gráficos, um considerando arquivos de textos em inglês e outro com arquivos contendo sequências de proteínas, que relacionam a taxa de compressão de um arquivo com o tamanho original dele, para ambos o nosso LZ78 e o gzip.

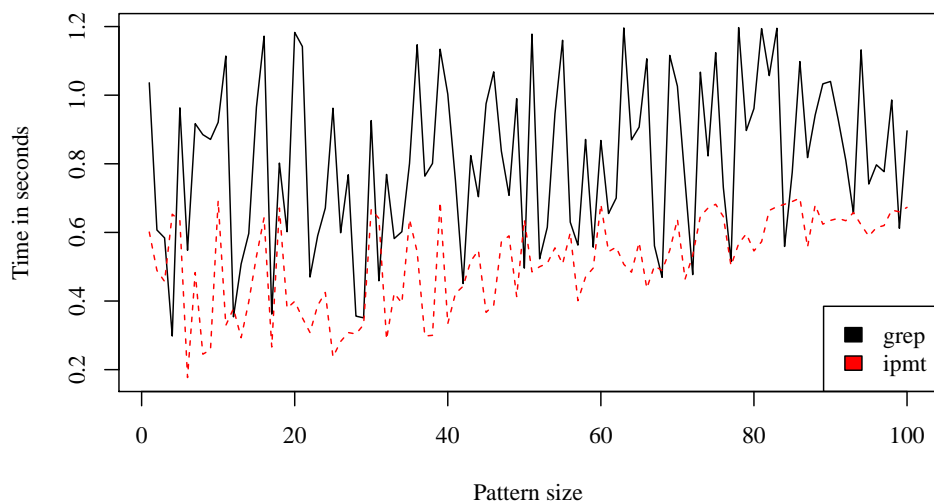
Novamente, os resultados se assemelham quando analisados separadamente: Ambos possuem uma variação inicial (apesar do gzip ter uma bem menor) e se estabilizam para arquivos maiores. Porém, para textos em inglês, o gzip se estabiliza com uma taxa de compressão próxima de 60%, enquanto que o nosso LZ78 se estabiliza com uma taxa próxima de 42%. Já para sequências de proteínas o gzip se estabiliza com taxa de 53%, enquanto o LZ78 se estabiliza com taxa de 20%. Recorremos novamente ao argumento de que o gzip está em desenvolvimento a muito mais tempo e por isso possui mais otimizações. Também podemos perceber que a taxa de compressão varia com o conteúdo do arquivo, diferente da pequena diferença encontrada nos tempos de compressão e descompressão.

4.2 Como a busca de padrão do *ipmt* se compara ao *grep*?

Para responder à essa pergunta nós utilizamos a ferramenta *ipmt* para criar um índice de um arquivo de texto. Após isso comparamos o tempo de realizar buscas de diversos padrões utilizando *ipmt* no arquivo indexado e comparando com o tempo para pesquisar os padrões utilizando *grep* no texto original. O tempo que leva para criar o Linear Suffix Array foi desconsiderado.

Nós consideramos somente o arquivo de 50MB contendo texto inglês pelo seguinte motivo: O arquivo `.idx` gerado pela ferramenta *ipmt* possui ambos o texto original e o índice comprimidos. Acontece que o array do nosso LSA tem, em média, quase 10 vezes o tamanho do arquivo original. Ou seja, indexar e comprimir um arquivo de 50MB acaba se tornando uma tarefa de comprimir um arquivo de aproximadamente 500MB, que dura entre 3 e 4 minutos no nosso ambiente de experimentos. Nesse caso, nossa compressão criou um arquivo `.idx` de 341MB, onde as buscas serão realizadas. Indexar arquivos maiores do que 50MB é uma tarefa factível, mas custosa e demorada nas nossas implementações.

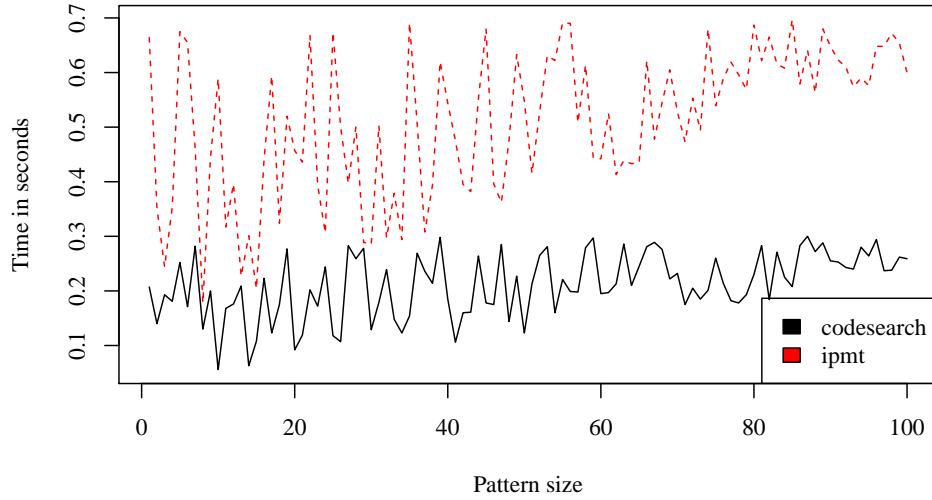
Geramos randomicamente 1000 padrões (frases ou palavras em inglês): 10 padrões com tamanho de cada valor entre 1 e 100 incluso. Após isso, realizamos a medição de tempo que o *grep* e que o *ipmt* levaram para buscar por esses padrões no arquivo de 50MB. Consideramos a média entre 10 execuções de cada ferramenta. O resultado está na figura abaixo:



Podemos ver que *ipmt* se saiu melhor na maioria dos casos. Isso era esperado devido ao pré-processamento da indexação. Podemos também observar que o tempo de resposta do *grep* variou independente do tamanho do padrão. Já no *ipmt* pode-se observar que quanto maior o padrão mais demorada tende a ser a resposta. Dado que o arquivo observado foi de 50MB, para um arquivo maior do que esse, *ipmt* tende a se sair melhor ainda em relação ao *grep*, visto que ele varia de acordo com o tamanho do padrão, enquanto que o *grep* varia de acordo com o tamanho do texto.

4.3 Como a ferramenta *ipmt* se compara ao *codesearch*?

Realizamos o mesmo experimento da seção anterior, porém comparando a ferramenta *ipmt* ao *codesearch*. Os resultados estão abaixo:



Visto que as médias que foram consideradas, o *codesearch* se saiu melhor que o *ipmt* em todos os casos.

5 Conclusão

Nesse relatório foi apresentada a ferramenta *ipmt*, capaz de indexar e comprimir um arquivo, possibilitando buscas cujo tempo varia linearmente com o tamanho do padrão.

Como esperado, conseguimos obter resultados melhores do que a ferramenta *grep*, porém ainda existem pontos onde a ferramenta pode melhorar para ficar tão eficiente quanto o *codesearch*.

Por exemplo, no método de contar os matches poderia ser implementado uma busca binária em dois estágios, um que busca a primeira ocorrência e outra que busca a última ocorrência. Logo o número de ocorrências seria a subtração entre a última e a primeira ocorrência. Além disso, atualmente o array do LSA é transformado em uma sequência de caracteres separados por um espaço. Ou seja, o array com objetos `[1, 5, 10]` será transformado na string `"1 5 10"`. Note que cada algarismo está utilizando um byte para ser representado, porém sabemos que cada algarismo vai variar entre 0 e 9, então na verdade precisamos somente de 4 bits para representar cada um,

possivelmente melhorando a taxa de compressão após a indexação.

Referências

- [1] B. Dorohonceanu and C. Nevill. A practical suffix-tree implementation for string searches. In *Algorithmica*, 2000.
- [2] J. Kärkkäinen and P. Sanders. Simple linear work suffix array construction. In *Automata, Languages and Programming, 30th International Colloquium, ICALP 2003, Eindhoven, The Netherlands, June 30 - July 4, 2003. Proceedings*, pages 943–955, 2003.
- [3] J. A. Storer. *Data Compression: Methods and Theory*. Computer Science Press, Inc., New York, NY, USA, 1988.
- [4] E. Ukkonen. On-line construction of suffix trees. In *Algorithmica*, pages 249–260, 1995.