**Universidade de Aveiro**

**Departamento de Eletrónica, Telecomunicações e Informática**

Mestrado em Engenharia Informática

***Teoria Algorítmica da Informação***

Trabalho Prático nº1

Rafael da Fonseca Fernandes | 95319

Gonçalo Junqueira | 95314

João Pedro Pereira | 106346

Aveiro | 13 de julho de 2021

Índice Geral

[1. Desenvolvimento do Programa 3](#_Toc87127724)

[1.1. Estrutura do Programa 3](#_Toc87127725)

[1.2. Criação do Alfabeto 3](#_Toc87127726)

[1.3. Criação do Modelo 3](#_Toc87127727)

[1.4. Cálculos das Probabilidades e Entropia 4](#_Toc87127728)

[1.5. Geração de Texto 4](#_Toc87127729)

[2. Análise de Resultados 4](#_Toc87127730)

[2.1. Variação dos Parâmetros k e alpha 4](#_Toc87127731)

[2.2. Variação da Entropia em Diferentes Textos 5](#_Toc87127732)

[3. Conclusões 6](#_Toc87127733)

1. Desenvolvimento do Programa

O objetivo do programa é criar um modelo com o objetivo de recolher informação estatística sobre textos utilizando modelos de contexto finito.

O modelo de contexto finito, de ordem k, produz a distribuição de probabilidade do próximo símbolo numa sequência de símbolos.

1.1. Estrutura do Programa

O modelo contém um dicionário, que é utilizado para guardar o número de vezes que cada símbolo aparece, além disso, guarda a probabilidade de um determinado símbolo ser o próximo.

O dicionário utiliza uma estrutura chave-valor, mais conhecida como hash map, na qual a chave é a sequência de n caracteres e o valor é um novo dicionário, onde a chave é o próximo caratere e o valor é a probabilidade do mesmo ser o próximo símbolo. Com esta abordagem foi possível armazenar vários dados cujos valores estão interligados, armazenar os dados de forma organizada e devido a utilizar menos memória.

Cada um dos programas criados recebem um conjunto de parâmetros no qual o programa irá se basear, sendo eles: a sequência, quantos símbolos deseja gerar, o tamanho da ordem e o parâmetro do alpha. Depois de os dados serem obtidos é criada a tabela consoante o tamanho do texto, isto é, caso o texto tenha um tamanho inferior a 250 é criada uma tabela bidimensional com todas as combinações e carateres possíveis a seguir, caso contrário utiliza um dicionário bidimensional.

1.2. Criação do Alfabeto

O alfabeto é produzido automaticamente com base nos caracteres existentes no texto fornecido, caso o tamanho seja inferior a 250, caso contrário é produzido automaticamente com todas as letras do abecedário. Essa criação é feita numa primeira leitura ao ficheiro que contém o texto, depois verifica o tamanho do texto e por fim todos os caracteres que não sejam duplicados são adicionados a uma estrutura de dados responsável por armazenar o alfabeto.

1.3. Criação do Modelo

Tal como referido na estrutura do programa, o modelo consiste num dicionário que guarda informação estatística. Nesse sentido, é necessário ler um documento de texto para ser possível recolher informação. Antes de começarmos a treinar o modelo é feita uma medição ao tamanho do documento para, caso tenha mais de 250 bits criar um dicionário bidimensional, caso contrário um array bidimensional. O processo para treinar o modelo consiste em recolher o número de vezes que cada símbolo ocorre num determinado contexto.

Para determinar o tamanho da sequência dos caracteres o grupo utilizou a ordem do modelo k. Esta operação vai percorrer todos os caracteres do ficheiro, para cada sequência de caracteres de tamanho k o que vem a seguir, além disso, guarda o número de vezes em que o mesmo aparece nesse contexto.

No final do processo é possível utilizar o modelo para o cálculo das probabilidades, obter a entropia do modelo e gerar um conjunto de caracteres.

1.4. Cálculos das Probabilidades e Entropia

Para o cálculo das probabilidades foi utilizado o parâmetro alpha.

Caso o parâmetro alpha seja menor que 0, o programa está construído para alterar esse valor para 1 e de seguida faz o cálculo de entropia para todos os valores diferentes de zero.

Por outro lado, se o parâmetro alpha possuir um valor maior que zero e os valores da tabela sejam diferentes de zero faz o cálculo de entropia.

1.5. Geração de Texto

Na geração do texto inicialmente é obtido uma sequência de N símbolos, e esses N símbolos são o valor do modelo. De seguida, esse valor é utilizado para verificar se existe na tabela, caso não exista é gerado um símbolo aleatório da tabela ASCII. Caso contrário, é escolhido uma das opções disponíveis segundo a percentagem, ou seja, um valor com 70% tem mais chances de ser escolhido do que um valor com 10%.

2. Análise de Resultados

Utilizando o programa desenvolvido para treinar o modelo e gerar texto, o grupo procedeu à análise de resultados relevantes.

O grupo analisou o impacto da variação dos parâmetros no cálculo da entropia e no texto gerado. Além disso, foi feita uma análise e comparação dos valores da entropia entre diferentes tipos de texto.

2.1. Variação dos Parâmetros k e alpha

Esta análise tem como objetivo analisar o impacto da variação dos parâmetros k (ordem do modelo) e alpha (parâmetro) no cálculo da entropia e no impacto da variação do parâmetro k.

Esta análise foi feita ao texto de exemplo fornecido pelo professor “example.txt”.

A tabela 1 permite analisar como varia a entropia em relação à ordem do modelo e do alpha.

Tabela 1 - Análise da entropia com base na variação dos parâmetros k e alpha

|  |  |
| --- | --- |
|  | Entropia |
| K=1, alpha=0 | 3.318718515445381 |
| K=4, alpha=0 | 1.6616451387330122 |
| K=7, alpha=0 | 1.0786532902441823 |
| K=1, alpha=0.5 | 3.3192605947817735 |
| K=4, alpha=0.5 | 1.6986357808274835 |
| K=7, alpha=0.5 | 1.222885684024801 |

Analisando o impacto do parâmetro k consoante a existência de um valor alpha podemos reparar que existe dois comportamentos distintos.

Quando o alpha é igual a zero verifica-se que o valor da entropia é menor do que quando o alpha é diferente de zero. Além disso, é possível verificar que quando maior o k menor é o valor da entropia.

A tabela 2 permite compreender como é que o valor do parâmetro alpha influencia o valor da entropia.

Tabela 2 - análise da entropia com base na variação do parâmetro alpha

|  |  |
| --- | --- |
|  | Entropia |
| K=5, alpha=0 | 1.4407113170887231 |
| K=5, alpha=0.6 | 1.5149462199269839 |
| K=5, alpha=1 | 1.5603730469373427 |
| K=5, alpha=1.5 | 1.6239821078458248 |

É possível verificar que quando maior for o valor do parâmetro alpha maior será o valor da entropia.

2.2. Variação da Entropia em Diferentes Textos

Para testar a variação do valor da entropia, consoante o tipo de texto utilizado para criar o modelo, o grupo definiu um conjunto de textos em diferentes linguagens. Os textos utilizados estão presentes na tabela 1.

Os seguintes testes foram efetuados para um modelo de ordem k=2 e para um parâmetro de suavização alpha=0.2

Tabela - Variação da entropia nos diferentes tipos de texto

|  |  |
| --- | --- |
| Exemplo | Entropia |
| A cidade e as Serras (Português) | 2.911592485134473 |
| Exemple (Inglês) | 2.549149351531155 |

A tabela acima permite estabelecer comparações entre a entropia de diferentes tipos de texto.

Nos exemplos referidos acima podemos verificar que as entropias das obras variam entre aproximadamente 2 e 3. Podemos verificar que o texto em inglês apresenta uma entropia mais baixa do que o português.

3. Conclusões

Podemos afirmar que os quanto maior a ordem do modelo menor é o resultado da entropia e quanto maior o alpha maior é o valor da entropia. Por fim, é de referir que quantas mais sequências forem necessárias mais espaços será preciso, isto é, quanto mais sequências maiores será a tabela.

Concluindo, o grupo reparou que um texto pequeno teve um valor de entropia maior do que um texto mais longo, devido a ser necessário gravar na tabela todas as sequências de texto possíveis segundo a ordem k e nesse caso o texto pequeno teve mais sequências do que o texto maior.