ESCOLA POLITÉCNICA
DE PERNAMBUCO

Departamento de
Sistemas
Computacionais

ENTROPIA E AUTOMATIZAÇÃO DE SÉRIES DE APROXIMAÇÃO DA LÍNGUA PORTUGUESA

Trabalho de Conclusão de Curso

Engenharia da Computação

Leopoldo Gomes Nogueira Rabelo
Orientador: Prof. Renato Mariz de Moraes

Recife, novembro de 2007

UNIVERSIDADE
DE PERNAMBUCO

ESCOLA POLITÉCNICA
DE PERNAMBUCO

Departamento de
Sistemas
Computacionais

ENTROPIA E AUTOMATIZAÇÃO DE SÉRIES DE APROXIMAÇÃO DA LÍNGUA PORTUGUESA

Trabalho de Conclusão de Curso

Engenharia da Computação

Este Projeto é apresentado como requisito parcial para obtenção do diploma de Bacharel em Engenharia da Computação pela Escola Politécnica de Pernambuco – Universidade de Pernambuco.

Leopoldo Gomes Nogueira Rabelo
Orientador: Prof. Renato Mariz de Moraes

Recife, novembro de 2007

UNIVERSIDADE
DE PERNAMBUCO

Leopoldo Gomes Nogueira Rabelo

ENTROPIA E AUTOMATIZAÇÃO DE SÉRIES DE APROXIMAÇÃO DA LÍNGUA PORTUGUESA

Resumo

À medida que aumentam a geração e a transmissão de informação, surge a preocupação com o armazenamento dos dados, com o uso dos canais de comunicação e com a segurança das informações transmitidas. Para esses casos, faz-se necessário conhecer as características da informação gerada. Um estudo realizado por Claude Shannon com a língua inglesa demonstrou a importância de saber como a informação é gerada e como esse conhecimento pode ser aplicado para permitir o uso mais eficiente da capacidade do canal de comunicação e em técnicas que visem garantir a proteção das informações, bem como a compressão dos dados.

Este trabalho se propôs a desenvolver uma ferramenta capaz de realizar de forma automatizada estudos similares aos realizados por Shannon para cálculo de entropia e geração de séries de aproximação. Com a ferramenta desenvolvida, foram replicados os estudos de Shannon para a língua inglesa, a título de comparação e validação. Em seguida, foram realizados estudos com a língua portuguesa.

Abstract

As the generation and transmission of information increases, it raises the concern with the storage of data, the use of communication channels, and also with the security of sent information. In all of these cases, it is necessary to know the characteristics of the generated information. A study performed by Claude Shannon with the English language demonstrated the relevance of knowing how information is generated, and how this knowledge can be applied to enable a better use of the channel's capacity, and in techniques that aim to ensure the protection of information, as well as data compression.

This work proposes the development of a tool able to perform, in an automated way, studies similar to those performed by Shannon to compute entropy and approximations of language series. With the developed tool, Shannon's studies for the English language were replicated for the purpose of comparison and validation. Additionally, studies with the Portuguese language were performed.

Sumário

Índice de Figuras

Índice de Tabelas

Índice de Equações

Tabela de Símbolos e Siglas

GNA – Gerador de números aleatórios
DNA – *Desoxyribonucleic acid* (Ácido desoxirribonucleico)
RNA – *Ribonucleic acid* (Ácido ribonucleico)

Agradecimentos

Primeiramente eu gostaria de agradecer a minha família por ter me dado a vida e pela minha formação ética e moral.
Gostaria de agradecer ao corpo docente do Departamento de Sistemas Computacionais pela formação intelectual e social. Em especial agradeço ao Professor Renato Mariz de Moraes pela orientação, pela paciência e pela colaboração na realização deste trabalho.
Agradeço aos meus colegas de turma, sem os quais, eu não teria concluído esta graduação. Agradeço especialmente a Thiago Ramos Trigo e Leopoldo Motta pela contribuição para a realização deste projeto.
Finalmente agradeço aos contribuintes pernambucanos por terem custeados meus estudos.

Capítulo 1

Introdução

Aumenta a cada dia a preocupação com a transmissão de dados, tanto pela quantidade de informação gerada cada vez maior, quanto pela segurança da informação gerada [1]. Dessa forma são precisos técnicas e ferramentas que permitam fazer um melhor uso do canal de transmissão. Técnicas de codificação ou de compactação surgiram para representar a informação de modo a conseguir uma economia na banda de transmissão. Já técnicas de proteção da informação, como a criptografia, garantem a segurança das informações. Em ambos os casos é preciso ter um conhecimento sobre como essa informação é gerada e como ela pode ser medida.

Um dos primeiros estudos realizados na área da teoria de informação foi realizado por *Claude Elwood Shannon* em 1948, na obra intitulada *The Mathematical Theory of Communication* [2]. Nessa obra é introduzida uma visão estatística da comunicação e é apresentado o conceito de medida da informação, onde a quantidade de informação contida em uma mensagem pode ser medida de forma bem definida por um valor matemático. A quantidade de informação não se refere ao volume de dados, mas a probabilidade de que uma mensagem ocorra dentro de um conjunto de possibilidades. A obra na verdade lança as bases teóricas do que hoje é entendido como Teoria da Informação. A visão matemática estabelecida sobre informação e comunicação permitiu responder perguntas como: **(i)** qual a maior taxa de transmissão possível em um canal para que não ocorram erros? **(ii)** até que limite é possível comprimir um conjunto de dados?

A primeira pergunta pode ser respondida pelo primeiro teorema de *Shannon* [2]. Para a segunda pergunta, a resposta é dada pelo valor da entropia do conjunto de dados que determina o limite de compressão de um conjunto de dados, como uma mensagem por exemplo.

Nas últimas décadas, um dos modelos de comunicação mais influentes é o apresentado por Shannon, que concebe a comunicação como uma transmissão de sinais. Conceitos como os de emissor, destinatário, código, sinal, informação, codificação e decodificação, utilizados de modo recorrente nas discussões sobre comunicação, são derivados desse modelo [3]. Trata-se de um modelo linear da comunicação visto como um processo de transporte da informação de um ponto A (o emissor) para um ponto B (o receptor).

O processo de comunicação é reduzido a uma questão de transporte, no qual as mensagens e significados, são tratados como meros sinais a serem identificados e decodificados por um receptor [3].

Shannon apresentou também um estudo sobre a fonte geradora de informação e em como as probabilidades de emissão dos sinais têm influência na capacidade de desenvolver uma codificação capaz de reduzir o uso do canal de comunicação. O estudo teve como exemplo a

língua inglesa, considerando como sinais emitidos, os símbolos contidos no alfabeto. Os símbolos gerados de forma aleatória considerando a probabilidade de emissão do sinal, ou no caso, as letras do alfabeto.

As informações geradas pelo estudo sobre a língua inglesa permitem desenvolver técnicas e aplicações que lidam com a codificação, compactação e transmissão de mensagens escritas em inglês.

O estudo de *Shannon* restringiu-se a língua inglesa, mas as técnicas e o modelo desenvolvidos por Shannon podem ser aplicadas para qualquer linguagem natural, inclusive as línguas que possuem alfabetos diferentes do ocidental, como o árabe ou japonês e ainda para linguagens artificiais, desde que elas possuam um conjunto básico de símbolos, ou alfabeto, e que cada símbolo tenha uma probabilidade associada a sua ocorrência.

1.1 Objetivos

Conforme visto é essencial o conhecimento da linguagem original quando se está envolvido em atividades como a compactação, a codificação e a proteção de dados. Um estudo sobre fontes geradoras de caracteres seguindo as regras gramaticais da língua inglesa foi realizado por Shannon, que estabeleceu uma série de técnicas para a realização desse estudo. A carência de estudos similares para a língua portuguesa é a principal motivação desse projeto, que se propõe a realizar uma análise similar à feita por Shannon, realizando as seguintes atividades:

- Processar um texto em português contendo milhares de palavras, provavelmente um livro;
- Calcular a partir desse texto, a freqüência relativa de ocorrência de cada letra, definida no alfabeto de símbolos válidos;
- Determinar a probabilidade de cada letra e de seqüências de letras, como a ocorrência de duplas de letras;
- Calcular a partir das probabilidades encontradas, a entropia da língua portuguesa;
- Gerar séries de aproximação para a língua portuguesa conforme os métodos descritos por *Shannon*;
- Comparar os resultados obtidos com trabalhos similares.

O resultado final deste projeto é desenvolver uma ferramenta que automatize o processo definido por *Shannon*, podendo realizar essas atividades em um único ambiente que permita receber como entrada um texto em português e gerar, a partir dele, o cálculo das probabilidades, o cálculo da entropia e as series de aproximação para a língua portuguesa, seguindo todos os passos descritos acima.

O projeto tem também como objetivo repetir o estudo para a língua inglesa e comparar os resultados obtidos com os apresentados por *Shannon* e por trabalhos similares.

1.2 Estrutura do trabalho

Esta monografia está organizada em Capítulos e Apêndices. A seguir está detalhado o conteúdo de cada parte:

O Capítulo 2 discute a fundamentação teórica necessária para o entendimento do trabalho. Detalha e explica os estudos realizados por Shannon envolvendo a influência do conhecimento estatístico na fonte de informação

O Capítulo 3 introduz o programa desenvolvido e explica como cada função apresentada no trabalho de Shannon foi implementada.

O Capítulo 4 descreve como foram realizados os testes e quais os resultados obtidos quando comparados com os resultados disponíveis para o inglês e para o português.

O Capítulo 5 conclui o trabalho, detalhando os resultados, as principais contribuições referentes a este trabalho, as dificuldades encontradas para a sua realização e possíveis trabalhos futuros.

Capítulo 2

Fundamentação Teórica

Neste capitulo é apresentado o trabalho realizado por Shannon sobre uma fonte discreta de informação envolvendo questões sobre como descrever matematicamente uma fonte de informação e sobre qual o efeito do conhecimento das probabilidades de emissão de um símbolo pela fonte visando à redução do uso do canal de transmissão disponível.

Os conceitos apresentados neste capitulo são fundamentais para o entendimento do nosso trabalho. As informações descritas aqui serviram como ponto de partida para a realização desse projeto, que compõe em implementar as técnicas utilizadas e desenvolvidas por Shannon.

Além das técnicas e da experiência, é preciso relatar as razões para utilização de cada uma delas e que se esperar com isso.

2.1 Fonte de informação

A descrição matemática de uma fonte discreta geradora de informação considera uma fonte como um gerador aleatório de símbolos [3]. Esses símbolos estão contidos no conjunto de possíveis símbolos emitidos pela fonte. A emissão de cada símbolo ocorre de acordo com as probabilidades associadas a cada símbolo e a relação entre os símbolos, já que a saída de um símbolo pode estar condicionada a saída de um ou mais antecessores. Uma mensagem emitida pela fonte pode ser considerada como o conjunto de símbolos emitidos durante certo tempo.

A forma como os símbolos se relacionam e a probabilidade individual de ocorrência de cada símbolo permite criar uma codificação que permita aproveitar melhor o uso do canal de comunicação [2], um exemplo disso é o código Morse que foi adotado pelo telégrafo, que emite apenas dois possíveis símbolos: o ponto e o traço, e gera mensagens através da combinação dos dois e do intervalo, (período de não emissão de símbolos). O código Morse foi baseado na língua inglesa, cujo símbolo mais freqüente é a letra “e”, que é codificado na forma mais simples possível com o ponto em contrapartida letras menos freqüentes como “z” ou “q” são codificados com uma sentença extensa de pontos e traços. A figura 1 mostra a codificação Morse utilizada para transmissão de símbolos pelo telégrafo.

A	.-	M	--	Y	-.--	6	-....
B	-...	N	-.	Z	--..	7	--...
C	-.-.	O	---	Ä	.-.-	8	---..
D	-..	P	.--.	Ö	---.	9	----.
E	.	Q	--.-	Ü	..--	.	.-.-.-
F	..-.	R	.-.	Ch	----	,	--..--
G	--.	S	...	0	-----	?	..--..
H	....	T	-	1	.----	!	..--.
I	..	U	..-	2	..---	:	---...
J	.---	V	...-	3	...--	“	.-..-.
K	-.-	W	.--	4	....-	‘	.----.
L	.-..	X	-..-	5	.....	=	-...-

Figura 1. Código Morse.

A fonte de informação emite símbolos de forma aleatória, o conjunto dos símbolos pode ser considerado como um conjunto de variáveis aleatórias, onde cada símbolo pode ser representado na forma X(t), tendo como variável o tempo, assim a fonte pode ser vista como um processo estocástico [2].

2.2 Entropia

Definido como ocorre a geração de informação é possível medir a quantidade de informação produzida através de uma equação estabelecida por *Shannon*. A equação (1) para medição da informação para um evento X é dada por:

$$I(X) = \log 1/p_i \,. \qquad \textbf{(1)}$$

A equação define a quantidade de informação associada à ocorrência de um determinado evento X, com probabilidade positiva P(X) de acontecer [4]. Essa definição comprova o que pode ser intuitivamente percebido, pois um evento com probabilidade um de ocorrer não traz nenhuma informação. A escolha da base dois para o logaritmo é arbitraria e usada apenas para representar a quantidade de informação na forma de *bits*, para o caso de utilizar o logaritmo natural o resultado seria medido em *nats*.

A partir da definição da quantidade de informação é possível medir a quantidade de informação média associada a uma variável aleatória, essa medida é denominada entropia ou incerteza da variável aleatória. A entropia é a esperança matemática para a quantidade de informação contida em um evento possível qualquer. A equação da entropia foi definida por *Shannon* da seguinte forma:

$$H(x) = -\sum_i P(x_i) \log_2 P(x_i) \text{ bits.} \qquad \textbf{(2)}$$

Para facilitar o entendimento considere, por exemplo, que o resultado do lançamento de uma moeda é menos incerto que o lançamento de um dado e que por sua vez o lançamento de um dado é menos incerto que a rodada de uma roleta. A informação associada ao evento rodar a roleta é maior que a informação associada ao lançamento de uma moeda.

A entropia é uma função apenas das probabilidades [4] e isso pode ser percebido pela seguinte situação, considerando duas variáveis aleatórias distintas X e Y, onde X pode assumir os valores X={00,01} e Y pode assumir Y={100,1000}. Considerando também que as probabilidades dos eventos X= 00 e Y=100 são iguais a P e as probabilidades dos eventos X=01 e Y=1000 é igual a 1-P. As incertezas dos eventos X e Y são iguais mesmo gerando valores diferentes.

Com base no conhecimento intuitivo, é possível estabelecer as seguintes propriedades, conforme apresentado por Silva [4].

1. $H(p_1,p_2,...,p_n)$ é máxima quando $p_1= p_2=...=p_n= 1/n$. A entropia é máxima quando os eventos forem equiprováveis;
2. Para qualquer permutação na ordem do conjunto $\{p_1,p_2,..p_n\}$ a entropia não se altera;
3. $H(p_1,p_2,..p_n) \geq 0$; sendo zero apenas se P_i for igual a um para algum i . Estabelece a entropia como função positiva sendo nula quando não houver incerteza;
4. $H(p_1,p_2,...p_n,0) = H(p_1,p_2,...p_n)$. Garante que apenas eventos com probabilidade de ocorrer são significativos;
5. $H(1/n,1/n,\ldots,1/n) \leq H(1/(n+1), 1/(n+1),\ldots,1/(n+1))$. Estabelece que em espaços equiprováveis, terá maior entropia aquele que tiver um número maior de elementos;
6. $H(p_1,p_{2,...,}p_n)$ deve ser função contínua dos argumentos, ou seja, uma pequena variação em um dos argumentos não deve provocar uma alteração significativa no valor da entropia;
7. $H(1/mn,1/Mn,...,1/mn) = H(1/m,1/m,...,1/m) + H(1/n,1/n,...,1/n)$. Estabelece a condição de linearidade, por exemplo, a incerteza no lançamento de dois dados deve ser mesma da soma das incertezas de cada dado, separadamente;
8. Seja $p = p_1+p_2+...+p_m$ e $q = q_1+q_2+...+q_n$, onde $p_i,q_j \geq 0$ $p+q = 1$. Então $H(p_1,p_2,...p_n,q_1,q_2,...,q_m) = H(p,q) + pH(p_1/p,p_2/p,...,p_m/p) + qH(q_1/q,q_2/q,...,q_n/q)$. Indica que a incerteza pode ser dividida em sub-espaços, por exemplo, em uma corrida entre *m* carros brancos e *n* carros pretos, com p_i a probabilidade de ganhar do *i*-ésimo carro branco e q_j a probabilidade de ganhar do *j*-ésimo carro preto, então a incerteza global será a incerteza de ganhar um carro branco ou preto, mais a soma ponderada das incertezas dentro de cada grupo individualmente.

É possível visualizar graficamente a variação da quantidade de informação associada a um evento através do exemplo do lançamento de uma moeda descrito em [4]. Considerando os valores das probabilidades P(cara) = p e P(coroa) = 1-p e que para H(p=0) = H(p=1) = 0 de modo que H é continuo no intervalo [0,1], aplicando os valores na equação da entropia e variando os valores de p de 0 a 1, é possível visualizar o comportamento de H(p), que tem seu ponto máximo quando as duas situações possíveis têm mesma probabilidade de ocorrer. A figura 2 ilustra o exemplo:

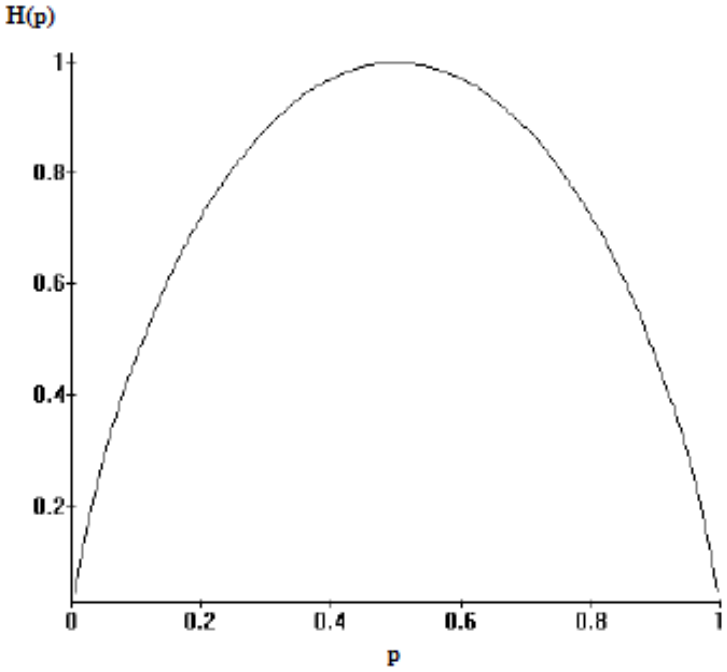

Figura 2. Variação da entropia pela variação do valor de p.

Outros conceitos referentes à entropia dizem respeito à entropia conjunta e a entropia condicional. A prova das equações seguintes pode ser obtida em [4], sendo aqui enunciadas.

- Entropia conjunta: seja X um vetor formado por X = {A,B} e A e B variáveis aleatórias assumindo valores finitos possíveis a_i e b_j, a entropia do vetor X ou do conjunto {A,B} é dada pela seguinte equação (3):

$$H(X) = H(A,B) = - \sum p(A = a_i, B = b_j) \log_2 p(A = a_i, B = b_j). \quad \textbf{(3)}$$

- Entropia condicional: Considerando que duas variáveis aleatórias X e Y podem assumir um número finito de valores. A entropia condicional de X em relação a Y, ou seja, incerteza sobre X uma vez que aconteceu Y, é definida pela equação:

$$H(X/Y) = - \sum p\,(x,y) \log_2 p\,(x/y). \quad \textbf{(4)}$$

2.3 A experiência de *Shannon*

No estudo sobre fontes geradoras de informação, *Shannon* definiu uma fonte como um processo estocástico tal qual uma linguagem natural como o inglês, o chinês ou o português ou uma linguagem artificial criada a partir de um conjunto de símbolos possíveis.

Um conjunto finito de símbolos é definido como alfabeto [4]. Alguns exemplos de alfabetos são:

- Conjunto das letras minúsculas;

- Conjunto das letras maiúsculas e minúsculas;
- Seqüência de numerais de zero a nove;
- Conjunto de palavras de um dicionário.

Uma seqüência de n-letras desse alfabeto é chamada n-grama sobre esse alfabeto. A partir de um alfabeto é possível gerar outro a partir da construção de n-gramas do alfabeto original. O novo alfabeto é uma extensão do original. Para o caso de digramas, ou seja, uma seqüência formada por dois elementos do alfabeto, é possível estender o alfabeto original $\mathcal{A}$ e gerar um alfabeto $\mathcal{A}^2$ contendo n^2 elementos, onde n é o número de elementos do alfabeto $\mathcal{A}$. Generalizando, o valor do n-grama determina o grau de extensão do alfabeto original e número de elementos do novo alfabeto.

Shannon definiu um alfabeto para uma linguagem artificial contendo as vinte seis letras do alfabeto inglês e o espaço, ou seja a não emissão de uma letra é considerada como um símbolo válido. A partir da definição desse alfabeto ele iniciou um processo de aproximação dessa linguagem artificial para a língua inglesa.

Inicialmente a linguagem não apresentava nenhuma regra sobre como seria a geração de símbolos pela fonte de informação. Nesse estágio *Shannon* estabeleceu que os símbolos do alfabeto fossem equiprováveis e independentes, desse modo todos os símbolos tinham a mesma probabilidade de serem emitidos e a ocorrência de um símbolo não influenciava na emissão do próximo. Uma seqüência gerada nessas condições, utilizando-se o alfabeto $\mathcal{A}$ = {A,B,C,D,...,Z,_}, onde o símbolo "_" representa o espaço, teria o seguinte aspecto:

XFOML_RXKHRJFFJUJ_ZLPWCFWKCYJ_FFJEYVKCQSGHYD_QPAAMKBZAACIBZLHJQD

A seqüência gerada é chamada de grau zero de aproximação e é possível perceber a presença de letras que normalmente são vistas com pouca freqüência em textos escritos em inglês como é o caso do "q" e do "z". A seqüência não pode ser identificada como gerada por nenhuma língua que se baseie nesse alfabeto. A série de ordem zero fornece a entropia máxima possível para uma linguagem qualquer que utilize esse alfabeto. O valor é calculado pela equação (2) definida na seção 2.2 e tem o valor de 4,75 bits /letra [5].

O próximo passo estabelecido por *Shannon* foi introduzir as probabilidades presentes na língua inglesa. Os símbolos agora não têm mais a mesma probabilidade de emissão, mas ainda sim ocorrem de forma independente. As probabilidades de cada letra na língua inglesa foi disponibilizada por *Fletcher Pratt* em [6]. Novamente foi apresentada uma seqüência de símbolos emitida pela fonte de informação:

OCRO_HLI_RGWR_NMIELWIS_EU_LL_NBNESEBYA_TH_EEI_ALHENHTTPA_OOBTTVA_NAH_BRL

Ainda não é possível identificar a língua de procedência da seqüência emitida, mas pode-se perceber que os símbolos com menor freqüência na língua inglesa, emitidos na série de grau zero não apareceram mais e por outro lado a letra "e" que possui maior probabilidade das letras na língua inglesa foi emitida 7 vezes e o espaço foi emitido 12 vezes.

No estágio seguinte *Shannon* faz uso das freqüências dos digramas presentes na língua inglesa e aumenta a aproximação da língua inglesa, os símbolos emitidos continuam independentes mas a relação entre dois símbolos está estabelecida, dessa forma tende a diminuir a seqüência de símbolos não existentes na língua inglesa[1]. *Shannon* disponibilizou uma série formada nessas condições.

1 Por não existentes entenda-se com probabilidade zero de ocorrer.

ON_IE_ANTSOUTINYS_ARE_T_INCTORE_ST_BE_S_DEAMY_ACHIN_D_ILONASIVE_TUCOOWE_AT_ TEASONARE_FUSO_TIZIN_ANDY_TOBE_SEACE_CTISBE

Pela primeira vez surgem palavras identificadas e mesmo as sem sentido, considerando a língua inglesa, começam a ter um aspecto semelhante às palavras existentes e com uma relação entre consoantes e vogais similar a presente na língua inglesa.

A aproximação aumenta com as probabilidades dos trigramas, gerando a série chamada de aproximação de terceira ordem. A série apresentada por Shannon apresenta um número maior de palavras identificadas.

IN_NO_IST_LAT_WHEY_CRATICT_FROURE_BIRS_GROCID_PONDENOME_OF_DEMONSTURES_ OF_THE_REPTAGIN_IS_REGOACTIONA_OF_CRE.

A partir desse ponto torna-se mais fácil utilizar palavras do que seguir com o processo de n-gramas. As palavras são emitidas de forma independente a partir de suas probabilidades individuais. Essa série gerada é chamada de primeira ordem de aproximação por palavras. A seqüência de palavras não apresenta sentido.

REPRESENTING_AND_SPEEDILY_IS_AN_GOOD_APT_OR_COME_CAN_DIFFERENT_NATURAL_HERE_ HE_THE_A_IN_CAME_THE_TOOF_TO_EXPERT_GRAY_COME_TO_FURNISHES_THE_LINE_MESSAGE_ HAD_BE_THESE

A segunda ordem de aproximação por palavras apresenta a probabilidade de ocorrência de duplas de palavras, nesse nível algumas frases já começam a apresentar certo sentido.

THE_HEAD_AND_IN_FRONTAL_ATTACK_ON_AN_ENGLISH_WRITER_THAT_THE_CHARACTER_OF_ THIS_POINT_IS_THEREFOREANOTHER_METHOD_FOR_THE_LETTERS_THAT_THE_TIME_OF_WHO_E VER_TOLD_THE_PROBLEM_FOR_AN_UNEXPECTED

Para a realização dessas séries *Shannon* utilizou tabelas contendo as freqüências das letras, dos digramas e trigramas, mas caso essas informações não estejam disponíveis ele criou um método que permite realizar de forma similar essa atividade. Este método é conhecido como é conhecido como método *Markoviano* e funciona da seguinte forma:

1. Abrir um livro em uma página qualquer;
2. Escolher aleatoriamente uma letra nessa pagina;
3. Essa letra é gravada;
4. Abrir o livro novamente em outra página qualquer;
5. Ler a página até encontrar a letra gravada;
6. Encontrada a letra, a seguinte é gravada;
7. Repetir o processo para essa nova letra gravada.

2.4 Aplicações

A Teoria da Informação está presente em diversas áreas como comunicações, codificação de dados, compressão de dados, recuperação de dados, criptografia e criptoanalise. Alguns exemplos de aplicações envolvem os conceitos descritos na obra de *Shannon* e servem para demonstrar suas aplicabilidades.

Braga, no seu trabalho intitulado “Análise de freqüência de línguas” [7], calcula o histograma da freqüência do português e do inglês e aplica essas informações em mensagens cifradas e conclui que é possível descobrir a cifra processando o histograma da mensagem cifrada cujo conteúdo seja desconhecido, e ordenando-o da mesma forma que foi processado o histograma da língua, assim é possível determinar a correspondência entre os símbolos e descobrir a cifra, seja ela de transposição ou substituição. As cifras de transposição são mais simples de quebrar, e ainda por cima é um caso especial das cifras de substituição [7]. Para o caso da cifra ser de substituição, o autor afirma que pode haver uma confusão na decisão entre letras que possuem naturalmente a freqüência parecida, neste caso ele faz uso do histograma de freqüência dos digramas repete, o processo realizado para as letras tentando verificar a correspondência entre os histograma de freqüência do texto cifrado e o histograma da língua.

Em outro trabalho, realizado por Fernandes, intitulado “Teoria da Informação e suas Aplicações em Compressão e Aleatoriedade” [8], o conceito de entropia é usado para aplicação na compactação de textos em português, além disso ele descreve formalmente um gerador de números aleatórios (GNA), cita exemplos e calcula a entropia de um GNA. Com relação à aplicação de compactação o autor conclui que é possível obter uma compactação de 15% do texto original considerando as probabilidades de primeira ordem, ou seja apenas a probabilidades das letras, e uma compactação de 23% considerando as probabilidades dos digramas.

Outra publicação demonstra a aplicação dos conceitos definidos por Shannon em áreas além da comunicação já que o conceito de incerteza associada a probabilidades se aplica em diversas outras áreas, é o caso do artigo intitulado *Shannon entropy applied to productivity of organizations* [9], em que a noção de entropia de *Shannon* é aplicada a organizações e indica que existe uma capacidade máxima por pessoa de tomada de decisão e que essa capacidade diminui à medida que a organização cresce, a menos que aconteça uma reestruturação podendo haver perda de produtividade dos gerentes da organização.

A Teoria da Informação está cada vez mais presente na área das ciências biológicas em aplicações por exemplo no estudo de padrões de controle genético para DNA e RNA [10].

Capítulo 3

O projeto

Neste capítulo é apresentado o projeto em si, tendo suas funcionalidades e características explicadas. Será demonstrado o processo que permite, a partir de um texto escrito em português, gerar séries de aproximação e calcular a entropia. Conforme visto no capítulo 2, a geração das séries e o cálculo da entropia requerem o conhecimento das probabilidades e estas são calculadas a partir do texto. De forma geral o programa permite realizar:

- O cálculo das probabilidades dos n-gramas de um texto até o grau cinco;
- O cálculo das probabilidades das palavras de um texto até o grau três;
- O cálculo da entropia a partir do texto;
- A geração de séries de aproximação para a língua em que o texto foi escrito.

3.1 O tratamento do texto

O texto é o ponto de partida do processo de geração de séries e do cálculo da entropia. Um texto em português contém uma variedade de símbolos maior que um texto em inglês e isso se deve a existência no português de palavras acentuadas e do "ç". Um alfabeto é o conjunto finito de símbolos utilizados em determinada linguagem para a geração de textos [4]. A fim de realizar uma comparação com o estudo feito para a língua inglesa, foi considerado apenas o alfabeto de vinte e seis letras e o espaço, dessa forma foi preciso realizar algumas alterações nos textos, uma vez que eles apresentam símbolos diferentes dos contidos no alfabeto adotado:

{a,b,c,d,e,f,g,h,i,j,k,l,m,n,o,p,q,r,s,t,u,v,w,y,x,z,espaço}

As letras que em um texto em português se apresentam acentuadas foram convertidas e contabilizadas sem o acento, ou seja, um "á" por exemplo, é considerado simplesmente como um "a", o "ç" é transformado em "c" e o hífen é considerado como um espaço. O programa ainda converte letras maiúsculas em minúsculas com o objetivo de manter um padrão na geração das séries de aproximação.Assim portanto, para fins de cálculo de probabilidade, a última atividade necessária é a eliminação dos elementos estranhos ao alfabeto definido e para tanto o programa desconsidera quaisquer outros símbolos que não estão contidos no alfabeto, como o ponto, a

vírgula, a exclamação e demais símbolos de pontuação. O programa realiza esse tratamento conforme descrito abaixo:

1. O texto em formato .txt é formado por uma seqüência de caracteres que são inseridos em uma matriz;
2. Cada posição da matriz tem seu conteúdo comparado com os símbolos definidos como o alfabeto;
3. Se o símbolo for válido a comparação segue para a próxima posição;
4. Se o símbolo for válido, mas estiver na forma maiúscula o valor daquela posição é substituído pelo similar minúsculo;
5. Se o símbolo não for pertencente ao alfabeto mas for igual a alguns dos símbolos a serem convertidos como “á” ou “ç”, o valor contido nessa posição é substituído pelo similar sem o acento ou em “c” no caso do “ç”. E a comparação passa para a posição seguinte;
6. Se o símbolo não se enquadrar em nenhuma das situações anteriores seu valor é apagado da matriz, que passa a ter o tamanho reduzido.
7. O processo segue até toda as posições da matriz serem analisadas.

O texto após esse processo se apresenta pronto para ser utilizado. A Figura 3 mostra a tela inicial do programa antes do texto ser carregado, todas as funções nesse ponto se encontram indisponíveis.

Figura 3. Tela inicial do Programa antes do texto ser carregado.

O tamanho do texto e o assunto abordado influenciam nos resultados obtidos. Foi observado em [7] que um texto precisa ter um tamanho mínimo para que se possa estimar as freqüências relativas das letras de um idioma de forma precisa, a partir desse limite mínimo a precisão obtida é praticamente irrelevante. Um texto abordando apenas um determinado assunto,

como um manual de uso de um produto ou um tutorial de uma ferramenta podem não exprimir necessariamente o comportamento do idioma analisado.

3.2 Cálculo das probabilidades

A fase de cálculo das probabilidades ocorre após o texto ser carregado no programa e ter passado pelo processo de tratamento, em que os caracteres indesejados foram eliminados. O usuário então tem duas opções de cálculo das probabilidades: **(i)** cálculo das probabilidades das letras; **(ii)** cálculo das probabilidades das palavras.

3.2.1 Cálculo das probabilidades das letras

A opção de cálculo por letras inicia ao mesmo tempo o cálculo das probabilidades das letras e dos n-gramas até o grau cinco, ou seja, uma vez que se inicia o cálculo, as probabilidades serão disponibilizadas após o termino de todas as n-gramas.

O cálculo das probabilidades das letras ocorre pela divisão do número de ocorrências de uma determinada letra pelo total de letras contido no texto. O processo obedece a seguinte lógica:

1. Cada linha do texto é inserido em uma matriz de caracteres;
2. A matriz é percorrida;
3. O primeiro elemento é posto em uma estrutura contento o próprio elemento e um valor, inicialmente zero, associado a ele;
4. O segundo elemento é comparado com o primeiro;
5. Se for igual o valor associado ao elemento é acrescido em uma unidade;
6. Se não for, o novo elemento é inserido na estrutura com o valor associado;
7. O processo segue até o valor do tamanho da matriz que contém o texto;
8. O valor associado a cada letra é dividido pelo valor do tamanho da matriz do texto;
9. A estrutura contendo os elementos e os novos valores associados é então exibida.

Um processo similar ao descrito acima ocorre para os n-gramas seguintes contendo apenas algumas alterações. A primeira alteração decorre da decisão do que considerar como n-grama no texto, sendo possíveis duas interpretações:

A primeira interpretação é formar n-gramas através da combinação dos caracteres de uma palavra, ou seja é possível formar (x-n+1) n-gramas a partir de uma palavra, onde x é o número de caracteres da palavra e n é o grau do n-grama e $x \geq n$, por exemplo na seqüência de caracteres “cabana” é possível formar cinco digramas (ca,ab,ba,an,na) ou quatro trigramas (cab,aba,ban,ana) e não é possível formar um 7-grama.

A outra interpretação é dividir as palavras de modo a formar os n-gramas, ou seja uma palavra com seis caracteres formaria três digramas e dois trigramas.

Nos estudos realizados não foi possível definir qual foi a interpretação adotada por cada autor, dessa forma o projeto adotou a primeira interpretação por considerá-la mais abrangente já que permite a formação de mais n-gramas. Para essa interpretação o cálculo das probabilidades ocorre contabilizando a ocorrência dos n-gramas e dividindo pelo total de n-gramas gerados. O programa efetua o cálculo até o pentagrama. A Figura 4 apresenta a tela com as probabilidades calculadas.

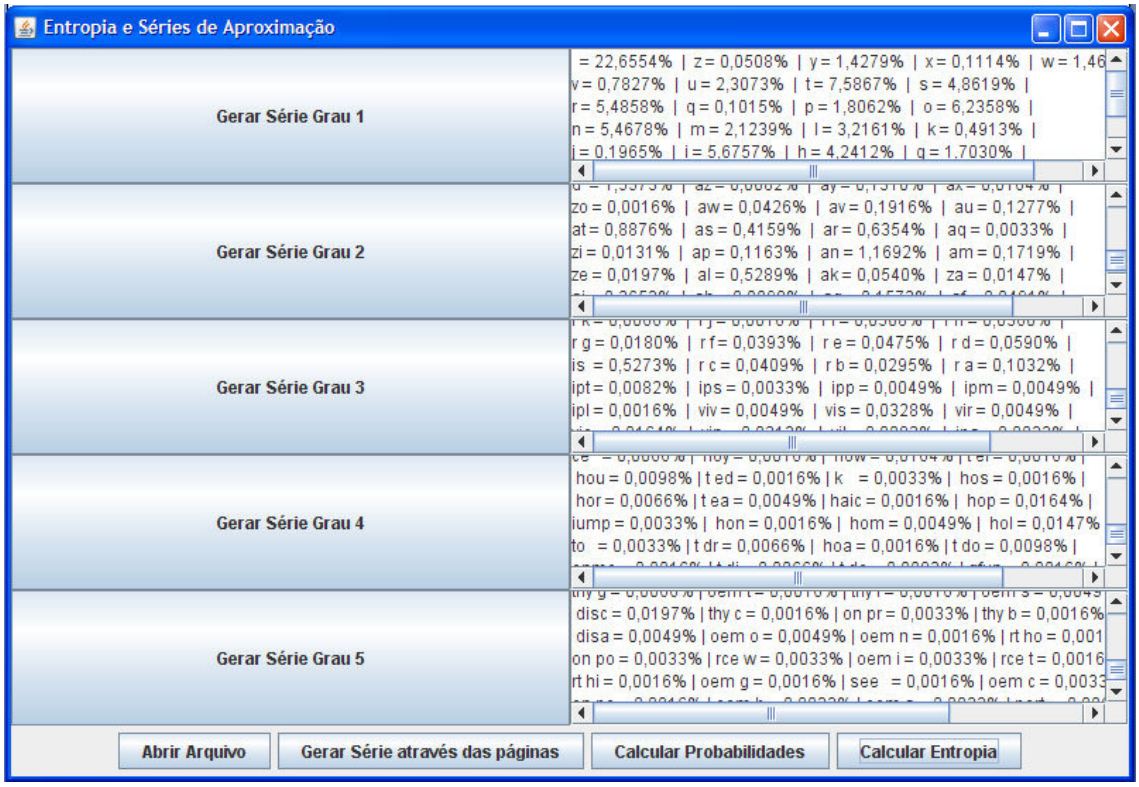

Figura 4. Tela do programa com as probabilidades calculadas para livro *The Raven*.

3.2.2 Cálculo das probabilidades das palavras

A outra opção é o cálculo das probabilidades das palavras contidas no texto. O programa realiza esse cálculo da mesma maneira que o realizado para as letras, considerando como palavra uma seqüência de caracteres delimitada pela presença de espaços, contudo os espaços são exibidos junto com as palavras, uma vez que estes estão incluídos no alfabeto e servirão para separar as palavras no momento da geração das séries.

A opção de cálculo por palavras calcula ao mesmo tempo a ocorrência das palavras isoladamente e a ocorrência de duas palavras e três palavras seguidas, de forma análoga aos n-gramas, assim a seqüência formada de uma palavra, espaço e outra palavra é armazenada com um valor associado, esse valor indica a ocorrência dessa seqüência e é incrementado a cada seqüência igual. O cálculo das probabilidades das palavras não ocorre simultaneamente com o cálculo das letras.

O método utilizado no programa forma todas as seqüências possíveis de acordo com o grau em questão, por exemplo, para o segundo grau, uma seqüência como “casa de madeira” gera duas seqüências possíveis, uma formada por “casa de” e outra formada por “de madeira”, a idéia se estende para o terceiro grau.

3.3 Geração das séries de aproximação

A geração de séries de aproximação é uma das principais funcionalidades do programa, onde essas séries são seqüências de símbolos pertencentes a um determinado alfabeto e geradas por uma fonte de texto [4] e que se aproximam de uma determinada linguagem [2].

Fonte de texto é um modelo matemático capaz de reproduzir, o mais fielmente possível, textos aceitáveis como pertencente a uma determinada linguagem [4], ou seja, fonte de texto é um mecanismo gerador de textos onde é possível se identificar em que linguagem o texto foi escrito.

O resultado da geração da fonte de texto é a série de aproximação. Cada símbolo dessa série é gerado de forma aleatória levando em consideração a probabilidade de ocorrência de cada símbolo. Esse método matemático de geração de símbolos considerando um conjunto de probabilidades é um processo estocástico [2].

Conforme visto no capitulo 2, as séries de aproximação são geradas de diferentes formas, tendo como objetivo se aproximar o máximo possível da linguagem em questão. Os três métodos observados são reproduzidos pelo programa e suas implementações explicadas nos próximos tópicos, sendo que os dois primeiros tópicos referem-se à geração de séries tendo como ponto de partida as letras pertencentes ao alfabeto. A geração de séries por letras tem duas diferentes abordagens: **(i)** A geração por n-gramas; **(ii)** A geração pelo método *Markoviano*. O terceiro método é a geração de séries partindo da freqüência das palavras.

3.3.1 Geração de séries de aproximação por n-gramas

O método de geração por n-gramas, considerando as letras como 1-grama, consiste em simular uma fonte de texto que emite símbolos de acordo com as probabilidades calculadas de cada letra e de cada n-grama, no caso desse programa o cálculo é feito até 5-grama e o processo é descrito na seção 3.2.1.

O caso inicial gera séries, emitindo um símbolo de cada vez, considerando a probabilidade de cada letra calculada pelo programa a partir do texto utilizado. A opção de gerar as séries de aproximação por letras é disponibilizada pelo programa após o cálculo das probabilidades de todos os n-gramas. A saída de cada símbolo ocorre independente da saída anterior. O processo de geração de séries segue a seguinte lógica:

1. O usuário determina qual série quer gerar dentre as cinco opções. As cinco opções de series são disponibilizadas ao mesmo tempo;
2. O usuário determina o tamanho da série a ser gerada pelo número de caracteres;
3. O programa cria uma matriz cujo tamanho é o somatório das freqüências obtidas pelo cálculo de probabilidade e preenche essa matriz com a quantidade relativa a cada probabilidade dos n-gramas. Por exemplo, no caso de cinco elementos, {a,b,c,d,e} cada um com probabilidade 20%, a matriz terá cem posições e será preenchida com vinte elementos “a”, vinte elementos “b” e assim sucessivamente;
4. Uma função geradora de números aleatórios uniforme, determina o símbolo a ser gerado pela fonte de texto. A função gera números contidos em um intervalo igual ao tamanho da matriz. O número na verdade indica a posição da matriz. O valor contido nesta posição será a letra gerada pela fonte;
5. A letra escolhida aleatoriamente é exibida;
6. O processo se repete até o valor determinado pelo usuário.

A idéia descrita acima é utilizada em todos os casos de geração por letras, tendo como diferença o fato que em cada posição da matriz estão os símbolos obtidos no cálculo das probabilidades e que esses símbolos têm tamanhos definidos pelo valor do n-grama, dessa forma quando o usuário determina o número de caracteres que serão gerados, na verdade esse valor se refere ao número de saídas, de fato será exibido uma seqüência com (x*n) valores saídas da fonte de texto, onde x representa o valor da série determinado pelo usuário e n o valor do n-grama. A figura 5 apresenta um exemplo de série gerada pelo programa com x = 30 e n = 5.

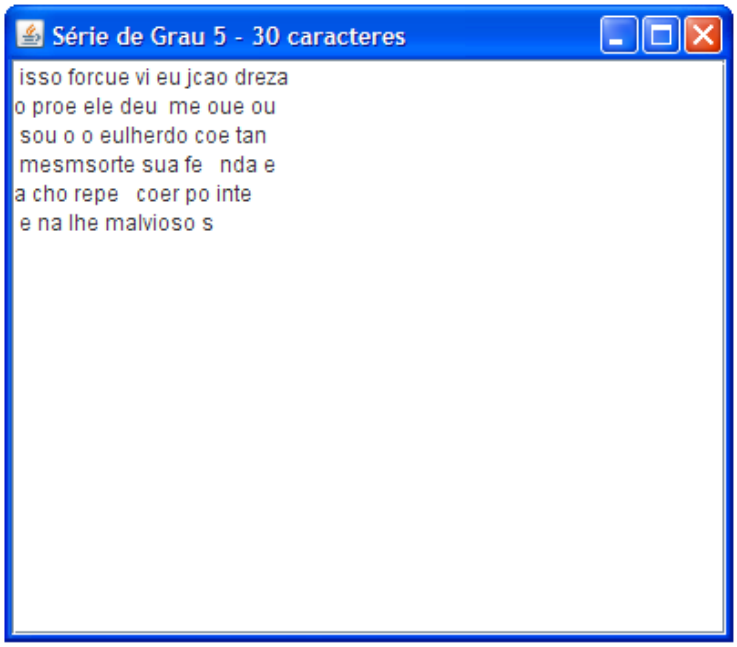

Figura 5. Exemplo de série gerada pelas probabilidades de 5-gramas.

3.3.2 Geração de séries pelo método *Markoviano*

Este método simula um gerador *Markoviano* descrito na seção 2.3. Na experiência de *Shannon* foi feito uso deste expediente com o intuito de gerar séries que fossem capazes de se aproximar cada vez mais da linguagem desejada. Este método está disponível ao usuário do programa imediatamente após o texto ser carregado, uma vez que não faz uso das probabilidades calculadas. Os processo é descrito abaixo:

1. O usuário escolhe a opção de gerar séries através das páginas do livro;
2. É dado ao usuário a opção gerar séries de grau um até quatro, mas ao contrário da opção de gerar séries pelos n-gramas, apenas uma opção é disponibilizada de cada vez;
3. Se o usuário escolher outra opção que não as permitidas, uma mensagem de erro é exibida e a tela com as opções é reexibida ao usuário;
4. O usuário após escolher corretamente dentre as opções válidas é apresentado a uma nova tela para a inserção do tamanho da seqüência gerada;
5. O programa então simula o processo conforme descrito na seção 2.3. Uma função geradora de números aleatórios determina a página a ser aberta. Como o texto na prática é uma seqüência de caracteres, foi preciso definir o que seria uma página.

A página foi definida como uma seqüência de cinqüenta linhas e com um total de dois mil caracteres;

6. Uma vez escolhida a página, tem-se novamente o processo de escolha aleatória, sendo agora para a letra. A letra é escolhida dentre as duas mil por uma função que gera números aleatórios uniformemente contidos em um intervalo, que no caso é número de caracteres da página;
7. O caractere escolhido é exibido e seu valor armazenado para posterior comparação;
8. A mesma função do item 5 escolhe a nova página e a percorre a partir da posição inicial comparando os valores com o valor armazenado.
9. Se o valor for igual ocorre um incremento na posição e o valor dessa posição é exibido e ocupa o lugar do caractere que estava sendo usado para comparação;
10. O processo continua até que seja gerado uma série de tamanho igual ao determinado pelo usuário;
11. Se o caractere não for encontrado na página a busca continua na página seguinte até o final do livro;
12. Se a busca não obtiver resultado até o final, o processo retorna ao inicio do livro e recomeça da posição inicial do texto.

O processo simula a dependência da emissão de um símbolo, pela fonte, com o símbolo anterior e esse processo é estendido aumentando a dependência de um símbolo com um número de símbolos anteriores cada vez maior. Para o caso de grau dois, por exemplo, a função escolhe aleatoriamente um símbolo, o armazena, exibe e usa nas comparações o símbolo e seu sucessor, se a dupla for novamente encontrada no texto o primeiro símbolo da seqüência é descartado. Por exemplo, se a dupla inicial “ca” for novamente encontrada no texto e seu sucessor for a letra “s” a série gerada até então será “cas” e a nova busca ocorrerá agora com dupla “as”.

A idéia segue para os casos seguintes onde a dependência do símbolo emitido aumenta para três e posteriormente para os quatro símbolos anteriores.

3.3.3 Geração de séries por palavras

Os métodos de geração de séries de aproximação descritos anteriormente geram séries baseadas em letras, podendo ser emitidos mais de uma de cada vez e de forma independente ou não. O passo seguinte na aproximação da linguagem desejada é a geração de séries baseadas nas probabilidades de ocorrência das palavras em uma determinada linguagem. O método descrito no capitulo 2 é realizado pelo programa partindo das probabilidades calculadas pelo processo descrito na seção 3.2.2.

Para esse caso a fonte de texto passa a emitir palavras inteiras e com tamanhos variáveis. A opção de gerar séries por palavras é disponibilizada somente após o cálculo das probabilidades por palavras. O processo de geração ocorre de maneira similar ao processo de geração a partir das probabilidades das letras e é descrito abaixo:

1. O usuário determina qual série quer gerar dentre as três opções. As três opções de séries são disponibilizadas ao mesmo tempo;
2. O usuário determina o tamanho da série a ser gerada pelo número de palavras;
3. O programa cria uma matriz cujo tamanho é o somatório das freqüências obtidas pelo cálculo de probabilidade e preenche essa matriz com a quantidade relativa a

cada probabilidade de cada palavra. A idéia é a mesma para o caso das letras descrito na seção 3.3.1;

4. Uma função geradora de números aleatórios, determina a palavra a ser gerada pela fonte de texto. A função gera números contidos em um intervalo igual ao tamanho da matriz. O número na verdade indica a posição da matriz. O valor contido nesta posição será a palavra gerada pela fonte;
5. A palavra escolhida aleatoriamente é exibida;
6. O processo se repete até o valor determinado pelo usuário.

O processo é o mesmo para a geração de mais de uma palavra de cada vez, sendo que nos casos com mais de uma palavra a fonte estará na verdade emitindo (x/n) saídas, onde x é igual número de palavras determinada pelo usuário e n o grau de aproximação determinado. Assim como na seção 3.3.1 cada emissão da fonte de texto simulada pelo programa ocorre independente da saída anterior.

3.4 Cálculo da entropia

O cálculo da entropia depende do texto de entrada. A entropia como foi explicado na seção 2.2 é obtida em função das probabilidades.

O programa calcula a entropia do texto através da implementação da equação descrita na seção 2.2, contudo foi usado um fator de conversão já que a linguagem de programação utilizada no desenvolvimento do programa não realiza cálculos de logaritmos na base dois e sim na base dez ou natural. Para que os resultados fossem obtidos pela equação (2), então foi preciso multiplicar a equação (2) por 3,32 que converte o cálculo da base dez para a base dois.

A entropia pode ser obtida utilizando as probabilidades das letras ou das palavras, dessa forma a opção de calcular a entropia somente será disponibilizada pelo programa após o cálculo das probabilidades. O processo de cálculo da entropia é descrito abaixo:

1. Após o cálculo das probabilidades a opção de calcular a entropia estará disponível e somente para a opção condizente com o cálculo, ou seja, se as probabilidades calculadas foram referente às letras apenas será possível obter a entropia a partir dessas probabilidades;
2. A geração ocorre de forma automática sem que seja necessário a inserção de nenhum valor;
3. A função desenvolvida realiza os cálculos tendo como entrada os valores das probabilidades. A função implementada no programa para o cálculo da série de primeira ordem está descrita na seção 2.2. Para os demais casos foi considerado a extensão do alfabeto, sendo então utilizada a equação (2) multiplicada por 1/n, onde n indica o grau de extensão do alfabeto [4];
4. A entropia é apresentada com os valores referentes a cada grau, sendo todos exibidos ao mesmo tempo.

A Figura 6 exibe a seqüência de valores da entropia obtidas a partir de um texto qualquer. É possível perceber a redução no valor da entropia calculada a medida em que aumenta a aproximação com a linguagem estudada. As regras presentes em cada língua aumentam a organização das letras e palavras e isso provoca uma redução na incerteza e conseqüentemente no valor da entropia.

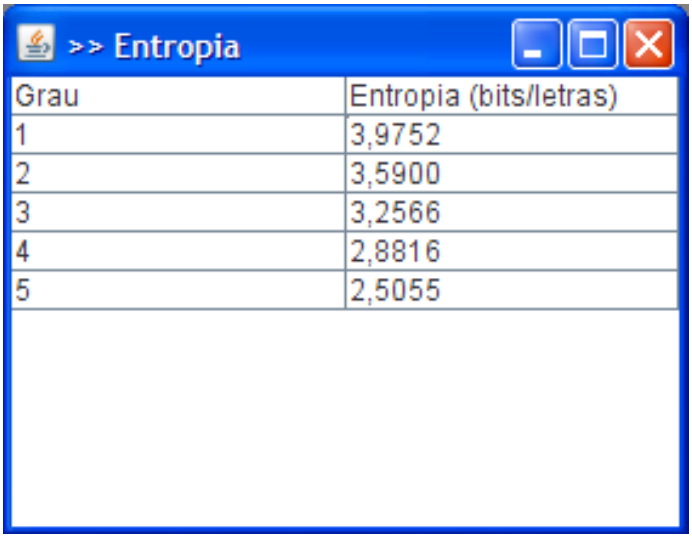

Figura 6. Tela de exibição das entropias.

Capítulo 4

Experimentos e Resultados

Os experimentos foram realizados de forma a verificar as funcionalidades do programa desenvolvido, em outras palavras avaliar se as séries de aproximação geradas e o cálculo da entropia são realizados corretamente. Foram utilizados diversos livros de diferentes autores, tanto para o português como para o inglês e a fim de realizar um estudo mais apurado. Foram escolhidos autores brasileiros e portugueses bem como autores americanos e ingleses. Os experimentos foram realizados primeiramente para a língua inglesa, para fins de comparação e validação do nosso trabalho tendo em vista que os estudos iniciais realizados por *Shannon* deram-se nessa língua. Em seguida, visando cumprir o objetivo deste projeto de conclusão de curso, foram realizados os experimentos para o português.

4.1 Experimentos para a língua inglesa

O objetivo dos experimentos com a língua inglesa é validar as funcionalidades do programa, para tal foi realizado uma comparação entre os resultados obtidos pela execução do programa com textos em língua inglesa e os valores apresentados em publicações por *Shannon* [2] e *Abransom* [5]. As principais funções a serem avaliadas são as de geração de séries de aproximação e o cálculo da entropia, contudo é necessário verificar também o cálculo das probabilidades pois este influencia de forma determinante as séries geradas e a entropia calculada.

Como fora dito anteriormente foram utilizados textos escritos por autores ingleses e americanos. Abaixo a lista dos livros utilizados nos experimentos, obtidos já em formato . txt, na biblioteca digital *Project Gutenberg* [11]:

- *Hamlet – William Shakespeare*
- *A Christmas Carol – Charles Dickens*
- *Alice's Adventures in Wonderland – Lewis Carroll*
- *The Adventures of Huckleberry Finn – Mark Twain*
- *The Raven – Edgar Allan Poe*
- *The Wonderful Wizard of Oz – L. Frank Baum*

Os livros utilizados foram escolhidos aleatoriamente entre autores ingleses e americanos, o objetivo é atenuar ou mesmo eliminar as particularidades de escrita ou qualquer diferença entre

o inglês britânico e o inglês americano. O texto foi processado pelo programa e seus resultados comparados com os artigos de *Shannon* [2] e *Abransom* [5].

A comparação das probabilidades será limitada ao caso da ocorrência de 1-grama, ou seja da ocorrência das letras na língua inglesa, isso se deve a dois fatores, primeiro ao tamanho ocupado pelas tabelas das probabilidades e segundo pela falta de tabelas acima de 2-gramas disponíveis para comparação. Para realizar a comparação, foi considerado a diferença média entre os valores obtidos nos dois casos. A equação é apresentada abaixo:

$$\text{Média} = \sum(X_i - Y_j)/N. \qquad \textbf{(5)}$$

onde X é o valor calculado pelo programa, Y o valor apresentado por *Shannon* [2], i e j os índices referentes às letras e N o tamanho do alfabeto.

A geração de séries por n-gramas foi realizada por Shannon até 3-gramas e apresentada por Abransom para o 1-grama, ou seja as letras. A comparação ocorreu contabilizando o número de palavras identificadas como escritas em inglês em séries com o mesmo número de caracteres. Apesar de não ter sido encontradas publicações com séries maiores que 3-grama, serão apresentadas também as séries de ordem quatro e cinco.

O mesmo método comparativo foi aplicado para a geração de séries pelo método *Markoviano*. Este método criado por *Shannon* foi disponibilizado na obra de *Abransom* [2] para os graus um e dois. O nosso programa apresenta também as séries geradas por esse processo para os graus três e quatro.

O método de geração por palavras apresenta séries formadas por palavras e tem como critério de avaliação o sentido proporcionado pela seqüência de palavras, ou seja, se uma palavra combinada com suas sucessoras estão em concordância gramatical e formam uma frase ou se são apenas uma coleção de palavras sem nexo.

A última função analisada é o cálculo da entropia, as tabelas disponíveis contendo os valores das probabilidades calculadas são apresentadas e a diferença média é comparada. Os valores comparados restringem-se aos valores apresentados em [5].

4.1.1 Resultados para a língua inglesa

Os resultados e as séries geradas são dispostas na seguinte ordem:

1. Tabelas com probabilidades das letras
2. Séries geradas por n-gramas
3. Séries geradas pelo método *Markoviano*
4. Séries geradas por palavras
5. Tabelas com os valores da entropia

CÁLCULO DAS PROBABILIDADES

A tabela 1 apresenta a comparação dos resultados calculados pelo nosso programa, presentes nas colunas 2 e 6, e os resultados fornecidos por *Shannon* [5], presentes nas colunas 3 e 7. O único símbolo com valor destoante é o espaço, isso deve-se ao fato do programa contabilizar um espaço a cada vez que verifica o final de uma linha.

Tabela 1. Comparação entre os valores calculados pelo programa e os valores fornecidos por [5].

Símbolo	Probabilidade		Diferença	Símbolo	Probabilidade		Diferença
	Leopoldo	Shannon			Leopoldo	Shannon	
a	6,41%	6,42%	-0,01%	o	6,51%	6,32%	0,19%
b	1,24%	1,27%	-0,03%	p	1,25%	1,52%	-0,27%
c	1,81%	2,18%	-0,37%	q	0,08%	0,08%	0,00%
d	3,86%	3,17%	0,69%	r	4,54%	4,84%	-0,30%
e	9,93%	10,31%	-0,38%	s	4,96%	5,14%	-0,18%
f	1,66%	2,08%	-0,42%	t	7,64%	7,96%	-0,32%
g	1,75%	1,52%	0,23%	u	2,37%	2,28%	0,09%
h	5,51%	4,67%	0,84%	v	0,67%	0,83%	-0,16%
i	5,20%	5,75%	-0,55%	w	2,17%	1,75%	0,42%
j	0,11%	0,08%	0,03%	x	0,09%	0,13%	-0,04%
k	0,80%	0,49%	0,31%	y	1,80%	1,64%	0,16%
l	3,39%	3,21%	0,18%	z	0,07%	0,05%	0,02%
m	2,03%	1,98%	0,05%	espaço	21,40%	18,59%	2,81%
n	5,32%	5,74%	-0,42%	Média			0,35%

Resultados: A diferença média em módulo calculada entre as probabilidades foi de 0,35%, indicando uma boa aproximação entres os valores.

SÉRIES DE APROXIMAÇÃO GERADAS POR N-GRAMAS

Primeiro caso: 1-grama. Os símbolos são geradas de forma aleatória e independentes mas a probabilidade de cada símbolo é considerada

A série gerada abaixo contendo 71 caracteres foi apresentada por Shannon:

```
OCRO_HLI_RGWR_NMIELWIS_
EU_LL_NBNESEB_YA_TH_
EEI_ALHENHTTPA_OOBTTVA_NAH_BRL
```

A série seguinte contém 112 caracteres e foi gerada por Abransom:

```
URTESHETHING_AD_E_AT_FOULE_ITHALIORT_W
ACT_D_STE_MINTSAN_OLINS_TWID_OULY_TE_T
HIGHE_CO_YS_TH_HR_UPAVIDE_PAD_CTAVED
```

Série gerada pelo nosso programa, a partir do livro *The Raven*, para comparação contendo aproximadamente o mesmo número de caracteres que os anteriores:

```
LNOT_PO_R_OWV_GOOD_TO_
N__BATLTHEWGFE_MH_RIHOIREAVTIET_
O_OTF_MT_L_S_RRW_HUG_MNOI_OL_
YI_M_URTIMTP_U_TEC_AND
```

Resultados: Os resultados observados demonstram a dificuldade em gerar séries em que se possa identificar palavras na língua inglesa. Os resultados podem ser visualizados na tabela 2.

Tabela 2. Comparação entre as séries geradas para 1-grama.

Autor	Número de palavras	Palavras
Shannon	0	-
Abransom	1	at
Leopoldo	3	good, to, and

Segundo caso : 2-grama: Neste caso as séries geradas a partir da probabilidade dos digramas, os símbolos emitidos, no caso a dupla de letras é gerada de forma independente.

A série gerada por *Shannon* contendo 117 caracteres foi esta:

ON_IE_ANTSOUTINYS_ARE_T_INCTORE_ST_BE_
S_DEAMY_ACHIN_D_ILONASIVE_TUCOOWE_AT_
TEASONARE_FUSO_TIZIN_ANDY_TOBE_SEACE_
CTISBE

Um exemplo de série gerada pelo nosso programa com 116 caracteres, a partir do livro *The Raven*, é apresentado abaixo:

NERSSS_OVEO_TIY__OR_TELE_Y_RMIUL_
TOUND_R_WA_O_N_TIR_R_RI_TWIONPHONMA_
ON_OR_PM_S_WERM_URNWTHNDXEMILEMYTSO
UR_N_NSTHS_TTOD

Resultados: Em ambos os casos já se torna possível identificar palavras de pequeno tamanho, com duas ou três letras. Neste grau o número de séries em que é possível identificar palavras aumenta bem como a quantidade de palavras. A tabela 3 apresenta os resultados dos dois autores

Tabela 3. Comparação entre as séries geradas para 2-grama.

Autor	Número de palavras	Palavras
Shannon	4	on, are, be, at
Leopoldo	4	or, tele, on, or

Terceiro caso: 3-grama: As séries são geradas a parir da probabilidade dos trigramas e os símbolos emitidos de forma independente.

A série apresentada por Shannon para o caso composto por trigramas contendo 107 caracteres:

IN_NO_IST_LAT_WHEY_CRATICT_FROURE_BIRS_
GROCID_PONDENOME_OF_DEMONSTURES_
OF_THE_REPTAGIN_IS_REGOACTIONA_OF_CRE.

Para o caso gerado por trigramas o nosso programa, a partir do livro *the Raven*, gerou a série abaixo com 107 caracteres:

BUT_COLD_D_EFO_BEWOOY_HUT_IN_HINAGEMB
E_DIES_NE_AS_LE_ROLOW_
BOUT_T__OY_SUCF_BES_E__RGIAY_
N_B_UR_ED__LOW_TGHT

Resultados: O incremento do processo causado pela geração a partir dos trigramas pode ser observado pelo aumento das palavras identificadas, aumenta também a freqüência de séries geradas em que se é possível identificar alguma palavra. A tabela 4 apresenta o resultado.

Tabela 4. Comparação entre as séries geradas para 3-grama.

Autor	Número de palavras	Palavras
Shannon	8	in, no, whey, of, of, the, is, of
Leopoldo	8	but, cold, hut, in, dies, as, bout, low

Quarto caso: 4-grama: Séries geradas a partir das probabilidades dos tetragramas, não foi encontrado outros autores que geraram séries até esse grau não havendo portanto comparação de resultados.

Série gerada pelo nosso programa, a partir da obra "*The Raven*" contendo 106 caracteres é apresentada abaixo:

_MAKNEVEERE_HEN_NT_T_AS_FOR_Y__I_OF_AY
ERAND_AT_I_A_TTER_E_WAS_WIRE_ITOOD_HAN
KITCTOMH_A_OTHEULD_DNT__THE_AND

Resultados: Como era esperado é possível identificar um número maior de palavras, aumenta também a quantidade de séries com palavras que são identificadas. A tabela 5 apresenta os resultados gerados na série.

Tabela 5. Resultados referentes à série gerada pelas probabilidades dos tetragramas.

Autor	Número de palavras	Palavras
Leopoldo	11	hen, as, for, of, at, a, was, wire, a, the, and

Quinto caso: 5-grama

Um exemplo de série gerada pelo programa, a partir do livro "*The Adventures of Huckleberry Finn*" contendo 110 caracteres segue abaixo:

IT_TH_ONCE_OUT_ND_SPL_BE_DS_OF
TIED_CLOTHTURE__YOU__IT_H
U_WASY_AIS_HOMER_NO_E_IT
ND_SA_YOU__WE_CY_TO_D_AND_LL_OF

Resultados: Nesse nível as séries contêm um grande número de palavras identificadas, palavras com quatro letras ou mais aparecem com maior freqüência. Na tabela 6 está disposto o resultado.

Tabela 6. Resultado referente à série gerada pelo programa para o pentagrama.

Autor	Número de palavras	Palavras
Leopoldo	16	It, once, out, be, of, tied, you, it, homer, no, it, you, to, and, of

SÉRIES DE APROXIMAÇÃO GERADAS PELO MÉTODO *MARKOVIANO*

Primeiro caso: Grau 1: As séries são geradas pelo método descrito na seção 2.3

A série gerada por Abransom [5], com 110 caracteres, segue abaixo :

URTESHETHING_AD_E_AT_FOULE_ITHALIORT_W
ACT_D_STE_MINTSAN_OLINS_TWID_OULY_TE_T
HIGHE_CO_YS_TH_HR_UPAVIDE_PAD_CTAVED

A série que foi apresentada por Silva [4] contém 90 caracteres:

SCULDOVESCOR_F_SUS_RED_WASIVECUSOR_S
CTHAMBLLOURMERN_G_BURIOTHESY_SCHEMAL
LANEEVISE_ETHE_OREN

A série disponibilizada, a partir do livro "*The Adventures of Huckleberry Finn*" pelo nosso programa é apresentada contendo 71 caracteres :

MINED_ATHOURELOOS_THE_SPI__
LDOOWOWOUTOMLLLAPAY_F_TE_THEO
_BE_ID_AT_THAUC

Resultados: O método neste nível apresenta um número ainda reduzido de palavras identificadas a série gerada pelo programa apesar de ser menor contém um número de palavras maior que o apresentado pelos estudos similares. A tabela 7 apresenta os resultados obtidos.

Tabela 7. Comparação entre os resultados dos três autores.

Autor	Número de palavras	Palavras
Abransom	3	at, act, pad
Silva	1	red
Leopoldo	4	mined, the, be, at

Segundo caso: Grau 2: A saída de um símbolo depende da saída dos dois símbolos anteriores conforme descrito na seção 3.3.1, a série gerada por Abransom, contendo 112 caracteres:

IANKS_CAN_OU_ANG_RLER_THATTED_OF_TO_S
HOR_OF_TO_HAVEMEM_A_I_MAND_AND_BUT_
WHISSITABLY_THERVEREER_EIGHTS_TAKILLIS_TA

A série gerada por Silva contém 71 caracteres:

```
EXT_OR_SYMAD_RES_FUN_PLUDEFOLY_
FOR_ORY_THER_CURSOR_STEME
_TO_OU_ANNINPUT
```

A série gerada pelo nosso programa, a partir do livro "*The Adventures of Huckleberry Finn*" está apresentada abaixo contendo também 71 caracteres :

```
ANK_ING_ALLAZED_OF_THE_DON_AND_
WAY_SH_ITHE_HIS_BECK_OVER_
NE_NOOD_HE_A_F
```

Resultados: Houve um aumento esperado na quantidade de palavras identificadas, a tabela 8 apresenta os resultados.

Tabela 8. Comparação entre os resultados dos três autores.

Autor	Número de palavras	Palavras
Abransom	9	can, of, to, of, to, a, mand, and, but
Silva	5	or, fun, for, cursor, to
Leopoldo	8	of, the, and, way, his, over, he, a

Terceiro caso: Grau 3: A saída de um símbolo depende da saída dos três símbolos anteriores

Abaixo a série gerada por Silva com 102 caracteres :

```
MAJORIZONTAL_IS_ZEROUS_COMPUT_
COPY_OF_AND_THE_LIMILABLESPOT_
LEVE_AVE_MODE_IS_DEALINE_OF_
PERMINORMATION
```

A série gerada pelo programa com 99 caracteres, a partir do livro *A Christimas Carol*:

```
ING_BIRD_YOU_CHAMBER_DEFERIOR_
THE_LORE_AGAZINER_FEATING_
WITH_THING_WITHE_PERFULL_IS_
OF_THE_LEFT_THE
```

Resultados: Em ambos os casos ocorre um aumento no número de palavras identificadas. A tabela 9 apresenta os resultados comparativos entre os dois autores que geraram séries no terceiro grau.

Tabela 9. Comparação entre os resultados dos dois autores.

Autor	Número de palavras	Palavras
Silva	5	Or, fun, for, cursor, to
Leopoldo	12	bird, you, chamber, the, lore, with, thing, is, of, the, left, the

Quarto caso: Grau 4: A saída de um símbolo depende da saída agora dos quatro símbolos anteriores

A série gerada por Silva contendo 87 caracteres, segue abaixo:

ENTRANCE_INTERNAL_ANY_OF_TO_A_
NUMBERED_ONSCREEN_IT_SUCH_TO_
DISPLAYED_AT_IS_GONE_CURRENT

A série gerada pelo nosso programa, a partir do livro *The Raven*, apresentada contém 80 caracteres :

OF_POE_ONE_AT_WORKS_FROM_
THE_AND_LIGHTS_REPETEND_IS_PROJECT_
RACE_OF_REFUND_I_BET

Resultados: Neste grau é possível gerar séries em que todas as palavras podem ser identificadas como pertencentes ao inglês. A tabela 10 apresenta os resultados.

Tabela 10. Comparação entre os resultados dos dois autores.

Autor	Número de palavras	Palavras
Silva	16	Todas
Leopoldo	16	Todas

SÉRIES GERADAS POR PALAVRAS

Primeiro caso: Grau 1: Os símbolos gerados são palavras geradas de forma independente a partir de suas probabilidades conforme descrito na seção 3.3.3.

A série geradas por Shannon com 32 palavras é mostrada abaixo:

REPRESENTING_AND_SPEEDILY_IS_AN_
GOOD_APT_OR_COME_CAN_DIFFERENT_
NATURAL_HERE_HE_THE_A_IN_CAME_THE_TOO
F_TO_EXPERT_GRAY_COME_TO_FURNISHES
THE_LINE_MESSAGE_HAD_BE_THESE

A série gerada pelo nosso programa, a partir do livro *A Christimas Carol*, contendo também 32 palavras segue abaixo:

SLIPPERS__AND__MY__CAN__SCALE__LOOKED_
_WAS_THE__HE__IN__SHE__YEAR__AND
PLEASED__WERE__THAT__IT__HE__QUESTION
GOLD__IT__THIS__ARE__RECOVER__A
IN__MENTION__AT__CENTRE__HE__SOUND
SUFFERS

Resultados: Em ambos as palavras são apresentadas sem que as frases tenham necessariamente sentido, sendo difícil identificar coerência sequer entre pares de palavras.

Segundo caso: Grau 2: Os símbolos gerados são palavras geradas de forma independente a partir das probabilidades da dupla de palavras

A série gerada e apresentada por Shannon com 34 palavras:

THE_HEAD_AND_IN_FRONTAL_ATTACK_ON_
AN_ENGLISH_WRITER_THAT_THE_CHARACTER_
OF_THIS_POINT_IS_THEREFORE
ANOTHER_METHOD_FOR_THE_LETTERS_THAT_
THE_TIME_OF_WHO_EVER_TOLD_
THE_PROBLEM_FOR_AN_UNEXPECTED

A série gerada pelo nosso programa, a partir do livro *A Christimas Carol*, também apresenta 34 palavras:

FEW_DROPS_PUT_ON_UP_ON_DEAD_
TO_LAID_EACH_GRIEF_AND_YOUR_GOOD
SCROOGE_PLENTY_THE_SCHOOL_ARTIFICIAL_
LITTLE_MARLEY_MR_THE_TIME_RUNNING_TO
POOR_AND_FLARING_LINKS_NOTHING_
FROM_PASSION_PRIDE

Resultados: Já se torna possível identificar um sentido em algumas seqüência de palavras.

Terceiro caso: Grau 3: Os símbolos gerados são palavras geradas de forma independente a partir da probabilidades de ocorrência do trio de palavras

A série gerada pelo nosso programa com 39 palavras, a partir do livro *A Christimas Carol*, é apresentada abaixo:

STRENGTH_ITS_LEGS_WERE_TOO_WARM_THE_
CELEBRATED_HERD_NOT_JUSTIFIED_IN
FEZZIWIG_TOOK_THEIR_BEEN_IN_VAIN_
GENTLEMAN_ITS_NOT_LIGHTED_CHEERFULLY_A
ND_TO_CONSIDER_WHAT_LIBERALITY_SCROOG
E_FROWNED_WITH_HARD_WEATHER_THE_MATT
ER_ASKED_WHOOP_HOW_ARE

CÁLCULO DA ENTROPIA

A Entropia calculada por Shannon até a terceira ordem é apresentada na tabela 11 :

Tabela 11. Valores da entropia calculados por Shannon.

Ordem	Entropia (bits/letra)
1º	4,03
2º	3,32
3º	3,10

A Entropia calculada pelo programa até a quinta ordem é apresentada na tabela 12.

Tabela 12. Valores da entropia calculados pelo programa.

Ordem	Entropia (bits/letra)
1º	4,0387
2º	3,6632
3º	3,2670
4º	2,9161
5º	2,6282

Resultados: A diferença média entre as entropias obtidas por *Shannon* e pelo programa comparando apenas até a terceira ordem é igual a 0,173.

4.2 Experimentos para a língua portuguesa

O objetivo deste experimento é demonstrar o funcionamento do programa para a língua portuguesa e compará-lo com trabalhos similares, contudo não foi encontrado nas pesquisas, material suficiente para realizar um estudo comparativo entre modelos, nos moldes do que foi realizado para o inglês. As séries disponíveis foram comparadas seguindo o padrão estabelecido para as séries em língua inglesa analogamente a seção 4.1. Assim como no processo realizado para o inglês, foram escolhidos ao acaso livros de autores brasileiros e alguns de autores portugueses, mas diferente do ocorrido anteriormente na seção [4.1] os experimentos ocorreram com obras nacionais, obtidas em formato .pdf e convertidas para o formato .txt, na biblioteca digital Domínio Público [12] e posteriormente comparadas com obras portuguesas, obtidas no Project Gutenberg [11], com o objetivo de verificar diferenças quanto à distribuição de probabilidades das letras e o cálculo da entropia. Abaixo segue a lista de obras literárias utilizadas nos experimentos:

- A Escrava Isaura – Bernardo Guimarães;
- A Viuvinha – José de Alencar;
- Dom Casmurro – Machado de Assis;
- Os Sertões – Euclides da Cunha;
- Os Lusíadas – Luis Vaz de Camões;
- O Mandarim – Eça de Queiroz;

4.2.1 Resultados para a língua portuguesa

O processo para a língua portuguesa é análogo ao descrito na seção 4.1.1.

CÁLCULO DAS PROBABILIDADES

A tabela 13 contém a comparação entre os valores das probabilidades referentes a 1-grama, ou seja, as letras, calculados pelo nosso programa, a partir do livro “Dom Casmurro” e os valores apresentados por Silva [4]. O valor do espaço apresenta a maior diferença devido ao programa contar um espaço cada vez que verifica o final de uma linha.

Tabela 13. Comparação entre os valores das probabilidades das letras calculados pelo programa e os apresentados por Silva [4].

Letra	Programa	Silva	Diferença	Letra	Programa	Silva	Diferença
a	12,14%	11,40%	0,74%	o	8,52%	8,00%	0,52%
b	0,70%	0,80%	-0,10%	p	2,22%	2,40%	-0,18%
c	3,01%	4,50%	-1,49%	q	1,20%	0,70%	0,50%
d	3,79%	4,60%	-0,81%	r	4,96%	5,30%	-0,34%
e	10,74%	11,40%	-0,66%	s	6,43%	7,10%	-0,67%
f	0,80%	1,00%	-0,20%	t	3,43%	4,60%	-1,17%
g	0,89%	1,10%	-0,21%	u	4,02%	3,00%	1,02%
h	1,09%	0,60%	0,49%	v	1,41%	1,20%	0,21%
i	5,26%	5,60%	-0,34%	w	0,00%	0,00%	0,00%
j	0,27%	0,20%	0,07%	x	0,23%	0,20%	0,03%
k	0,00%	0,10%	-0,10%	y	0,00%	0,00%	0,00%
l	2,20%	2,00%	0,20%	z	0,37%	0,30%	0,07%
m	4,27%	4,20%	0,07%	espaço	19,47%	15,60%	3,87%
n	3,91%	4,10%	-0,19%	Média			0,55%

SÉRIES DE APROXIMAÇÃO GERADAS POR N-GRAMAS

Primeiro caso: 1-grama: Os símbolos são gerados de forma aleatória e independentes mas as probabilidades de cada símbolo é considerada.

A série apresentada por silva com 37 caracteres segue abaixo:

CCENNIEPTSE_UQOCCAMS_AEJRNV_DDA_CTRAA

A série apresentada pelo programa, gerada a partir do livro "A Escrava Isaura" também com 37 caracteres segue abaixo:

ENER_SAEO_GS_O_ARATEEA_ES_ROTA_UERANA

Resultados: Assim como nos resultados observados para o inglês, neste estágio as palavras com sentido aparecem em pouca quantidade. A tabela 14 apresenta os resultados obtidos pelos dois autores.

Tabela 14. Comparação dos resultados obtidos pelos dois autores.

Autor	Número de palavras	Palavras
Silva	0	-
Leopoldo	3	o, es, rota

Segundo caso: 2-grama: Os símbolos são geradas de forma aleatória e independentes mas as probabilidades de cada digrama é considerada.

A série apresentada pelo programa, a partir do livro "Dom Casmurro" contendo 115 caracteres segue abaixo:

NTONS_QU_SSEPO_SO_S_TIQUMOSEMEOUA_
_R_SA_NA_A_S_A_SSA_QUSASEMAUEUSUMA_
OSUMNS_SNHLEQU_VAO_ONS__
SA_ORM_UMS_R_POOUTI_TMEA

Resultados: Considerando que a acentuação não está presente, três palavras foram identificadas. Os resultados estão dispostos na tabela 15.

Tabela 15. Resultados obtidos pelo programa.

Autor	Número de palavras	Palavras
Leopoldo	3	só, na,vão

Terceiro caso: 3-grama : Os símbolos são geradas de forma aleatória e independentes mas as probabilidades de cada trigrama é considerada.

A série apresentada pelo programa, a partir da obra "Dom Casmurro" com 108 caracteres segue abaixo:

JA__GRE_PER_O_TAISCREALM_ATCIMRE_RCAU
E_E_SS_GADO_O_TU_AM__PUMULCREA_IDIZA_I
AM_SACO_QIVEO_PUMEARA_A_AVRDO_ERA

Resultados: Houve um aumento esperado na quantidade palavras identificadas como pertencentes a língua portuguesa. A tabela 16 apresenta os resultados obtidos.

Tabela 16. Resultados obtidos pelo programa.

Autor	Número de palavras	Palavras
Leopoldo	10	já, o, e, e, gado, o , tu, saco, a, era

Quarto caso: 4-grama : Os símbolos são geradas de forma aleatória e independentes mas as probabilidades de cada tetragrama é considerada

A série apresentada pelo programa, a partir de "Dom Casmurro" com 108 caracteres segue abaixo:

MEUIO_PE_RES_MAAS_B_O_RE_ALDARITUME_CO
M_ATE_E_CLADIN_COMELHOERNA_VINIA_THOS__
DIVTIO_ORA_TEM_MOS__CANE_HA_DE

Resultados: Ocorre um aumento na presença das palavras, na série é possível identificar o mesmo número de palavras que no similar em inglês. A tabela 17 apresenta os resultados obtidos.

Tabela 17. Resultados obtidos pelo programa.

Autor	Número de palavras	Palavras
Leopoldo	11	pé, rés, o, ré, com, até, e, ora, tem, hã, de

Quinto caso: 5-grama: Os símbolos são geradas de forma aleatória e independentes mas as probabilidades de cada pentagrama é considerada.

A série apresentada pelo programa também com 108 caracteres, gerados a partir da obra "A Escrava Isaura", segue abaixo:

```
A_E_URRIA_UMA_CISTO_LHADO_CARO
SENTI_TOMA_E_SUO_TINDOS_O
O_VEELA_JUJEITOU_JUESTES
AMORNHO_DCOM_ECONTRO_AMA_AO_PE
```

Resultados: A quantidade de caracteres não sofreu um grande aumento contudo aumenta o número de séries contendo palavras identificadas e o tamanho das palavras também aumenta. A tabela 18 apresenta os resultados obtidos.

Tabela 18. Resultados obtidos pelo programa.

Autor	Número de palavras	Palavras
Leopoldo	14	a, e, uma, cisto, caro, senti, toma, e, suo, o, o, ama, ao, pe

SÉRIES DE APROXIMAÇÃO GERADAS PELO MÉTODO *MARKOVIANO*

Primeiro caso: Grau 1: A saída de um símbolo depende da saída de um símbolo anterior conforme descrito na seção 3.3.2.

A série apresentada por Silva com 35 caracteres segue abaixo:

```
DE_TADER_A_STEMERARA_DANA_CUR_MESTE
```

A série apresentada pelo nosso programa, a partir do livro "A Viuvinha", com 34 caracteres segue abaixo:

```
CURE_FAIARA_NARETENA_NAO_AO_QUE_CE
```

Resultados: um grande número de palavras identificadas se considerado o tamanho das series. A tabela 19 apresenta a comparação entre os dois autores.

Tabela 19. Comparação entre os dois autores.

Autor	Número de palavras	Palavras
Silva	2	de, a
Leopoldo	4	cure, não, ao, que

Segundo caso: Grau 2: A saída de um símbolo depende da saída dos dois símbolos anteriores.

A série apresentada por Silva com 101 caracteres segue abaixo:

AS_CRICA_METRIPTO_A_DEAL_SISOMOS_
TITIOMENTECIO_REAMEDIM_QUATUICANDO_
US_EXIS_OU_ATORMACAO_AS_PORDDAGEM

A série apresentada pelo nosso programa, a partir da obra "A Escrava Isaura", com 100 caracteres segue abaixo:

ESTOU_DE_ESPERENTRENISCIDAO_
DA_SOM_NAO_QUAR_
UMARIMPLICABEM_EU_VIS_UMAISSAS_
LEIDAR_AS_ININHO_CURAZENC

Resultados: O número de palavras identificadas aumenta nos dois casos. A tabela 20 ilustra os resultados obtidos.

Tabela 20. Comparação dos resultados obtidos.

Autor	Número de palavras	Palavras
Silva	4	as, a, ou, as
Leopoldo	7	estou, de, da, som, não, eu, as

Terceiro caso: Grau 3: A saída de um símbolo depende da saída dos três símbolos anteriores.

A série apresentada por Silva com 103 caracteres segue abaixo:

UM_SEFURNACA_SISTO_SISTEMA_
DES_ELES_QUE_MENTE_SISTEMA_
DE_PODERANCIA_PORTARA_UM_
COMPATIVEL_OU_ESTA_MESMO

A série apresentada pelo nosso programa com 103 caracteres, a partir da obra "A Escrava Isaura", segue abaixo:

DEIXOU_COMO_UMA_FORCO_
QUELE_LHE_UMA_NENAS_PRIDO_
AS_E_SAO_NAO_NEGO_AMOS_RAZAO_
POR_AZUL_COM_UMA_ME_DIAVAM

Resultados: Neste estágio quase o texto inteiro é composto por palavras identificadas. A tabela 21 apresenta os resultados obtidos.

Tabela 21. Comparação dos resultados obtidos.

Autor	Número de palavras	Palavras
Silva	13	um, sistema, eles, que, mente, sistema, de, portara, um, compatível, ou, este, mesmo
Leopoldo	16	deixou, como, uma, lhe, uma, as, e, são, não, nego, razão, por, azul, com, uma, me

Quarto caso: Grau 4: A saída de um símbolo depende da saída de quatro símbolos anteriores.

A série apresentada por Silva com 95 caracteres segue abaixo:

SISTEMAS_DE_DISTICACAO_OU_
MESMO_POR_EXEMPLO_SIMETRICOS_
DE_SUBSTITUI_O_CONCLUSIVO_
QUE_O_PROBLEMA

A série apresentada pelo programa com 84 caracteres, a partir da obra "A Escrava Isaura", segue abaixo:

E_UM_MOVIMENTO_DE_ME_COM_OS_
SE_NOVAS_ERA_AMIGOS_SALA_
DE_SANTUAR_A_CASA_E_ENTRO_HOMEM

Resultados: Em ambos os casos apenas um palavra não pode ser identificada com pertencente ao português é o caso de "disticacao" para a série gerada por Silva e "santuar" para a série gerada pelo nosso programa.

SÉRIES GERADAS POR PALAVRAS

Primeiro caso : Grau 1: Os símbolos gerados são palavras geradas de forma independente a partir de suas probabilidades.

A série apresentada pelo programa com 37 palavras, gerada a partir da obra "A Viuvinha", segue abaixo:

AFEICAO__RENDAS__A__AGRADECER_LHE_
_CUJO__A__COM_CONSISTIA__UMA__MENINA_
_A__UM__OBRAS_ALVAS__UMA_
_FERRO__QUE__TEMPO__OS
CRONICO__VALOR__ADIVINHAR_
_SENHOR__ALGUM__SUA
TINHA__QUE__NA__DA__E__PRIMEIRO
MENOS__DO__SENTAR_SE__SENHOR

Resultados: Nesse nível as palavras em geral se apresentam sem formar sentido, sequer considerando apenas os pares de palavras.

Segundo caso: Grau 2: Os símbolos gerados são palavras geradas de forma independente a partir das probabilidades das duplas de palavras.

A série apresentada pelo programa com 33 palavras, gerada a partir do livro "A Viuvinha", segue abaixo:

DE_MAIS_IA_VER_SENHOR_GRACAS_O_
BRACO_ADMIRADO_FICOU_TENAZ_QUE_
E_A_VIUVA_NO_UM_ANO_E_SEM_
AQUELE_AMANHA_HUMANA_PUDESSEM_
AMOR_TÃO_QUE_PELOS_CONHECIMENTO_
DO__PESAVA_AS_FORCAS

Resultados: Já se observa sentido em pares de palavras e em alguns casos em trios de palavras como em "sem aquele amanhã".

Terceiro caso : Grau 3: Os símbolos gerados são palavras geradas de forma independente a partir das probabilidades das triplas de palavras.

A série apresentada pelo nosso programa com 36 palavras, gerada a partir da obra "A Viuvinha" segue abaixo:

ALGUM_TEMPO_ELE_HORRIVELMENTE_
PALIDO_CAMINHOU_GROSSAS_
LETRAS_POR_NAQUELE_MOMENTO_ESSE
BEM_DIVERSO_DO_PROVANCAS_
E_MISERIAS_A_SUA_PLENITUDE
_TALVEZ_ACHE_A_MESMA_COR_
COMO_NAO_CAIU_DOS_PENSAMENTO_
LOUCO_E_SE_PASSOU_DEPOIS

Resultados: Aumenta a relação entre as palavras, sendo possível identificar sentido em trios de palavras e em certos casos em número maior como em "algum tempo ele horrivelmente pálido caminhou".

CÁLCULO DA ENTROPIA

Resultados: A diferença média entre os valores da entropia calculados pelo programa, a partir da obra literária "Dom Casmurro" e os valores calculados por Silva foi de 0,06. A tabela 22 apresenta os valores da entropia para o português, calculados pelo nosso programa e os valores calculados por Silva.

Tabela 22. Valores da entropia do português.

Leopoldo		Silva	
Ordem	Entropia (bits/letras)	Ordem	Entropia (bits/letras)
1º	3,94	1º	3,97
2º	3,56	2º	3,53
3º	3,27	3º	-
4º	3,00	4º	-
5º	2,76	5º	-

4.3 Análise comparativa dos resultados obtidos pelos livros

O objetivo desta seção é avaliar os resultados obtidos pelos diversos livros utilizados, tanto nos estudos conduzidos para a língua inglesa como para a língua portuguesa. A outra avaliação objetiva verificar qual das técnicas de geração de séries de aproximação por caracteres, ou seja, o método de geração por n-gramas e o método *Markoviano*, obteve melhor desempenho.

Para verificar qual dos livros e qual das séries lograram maior sucesso, realizou-se uma comparação inicialmente entre os resultados obtidos pelos livros em cada um dos métodos e posteriormente a comparação entre os dois métodos, independente de qual livro obteve o melhor resultado em cada método.

O computador utilizado na geração de todas as séries envolvidas nessa avaliação possui a seguinte configuração:

- Processador: Intel Core 2 duo de 2,4 GHz;
- Memória principal: 4 Gb;
- Sistema Operacional: Windows XP.

4.3.1 Análise dos resultados obtidos pelos livros em língua portuguesa

Os livros utilizados para a realização da análise comparativa dos resultados são os mesmos listados na seção 4.2.

GERAÇÃO DE SÉRIES POR N-GRAMAS

O método adotado para definir o melhor resultado obtido entre os livros foi comparar as médias das quantidades de palavras identificadas. Essas médias consistem no resultado da divisão da soma do número de palavras identificadas em cada série por cinco. Caso ocorra de dois ou mais

livros obterem o mesmo resultado, será considerado também como critério de avaliação o tempo relativo ao cálculo das probabilidades e nesse caso o livro cujo tempo dos cálculos seja o menor será considerado como o tendo o melhor resultado final. O procedimento utilizado na análise é descrito a seguir.

1. Inserção do texto no programa;
2. Cálculo das probabilidades do n-gramas;
3. Anotação do tempo de cálculo das probabilidades;
4. Geração de séries, do 1-grama até 5-grama, contendo sempre cem caracteres;
5. Anotação do número de palavras identificadas nas séries geradas.

A tabela 23 apresenta o tempo de execução do programa para calcular as probabilidades dos n-gramas.

Tabela 23. Tempo de cálculo das probabilidades dos n-gramas.

Livro	Tempo de execução (Minutos)
A Escrava Isaura	15
A Viuvinha	5
Dom Casmurro	18
O Abolicionismo	17
Os Lusíadas	16
O Mandarim	11

Primeiro caso: 1-grama: A tabela 24 apresenta os resultados obtidos por cada livro em séries geradas a partir das probabilidades do 1-grama, ou seja, das letras.

Tabela 24. Número de palavras identificadas em cada livro para séries do primeiro caso.

Livro	Nº de palavras identificadas
A Escrava Isaura	3
A Viuvinha	5
Dom Casmurro	4
O Abolicionismo	4
Os Lusíadas	5
O Mandarim	5

Segundo caso: 2-grama: A tabela 25 apresenta os resultados obtidos por cada livro em séries geradas considerando as probabilidades dos digramas.

Tabela 25. Número de palavras identificadas em cada livro para séries do segundo caso.

Livro	Nº de palavras identificadas
A Escrava Isaura	5
A Viuvinha	8
Dom Casmurro	7
O Abolicionismo	5
Os Lusíadas	7
O Mandarim	7

Terceiro caso: 3-grama: A tabela 26 apresenta os resultados obtidos em séries geradas a partir das probabilidades dos trigramas por cada livro.

Tabela 26. Número de palavras identificadas em cada livro para séries do terceiro caso.

Livro	Nº de palavras identificadas
A Escrava Isaura	9
A Viuvinha	11
Dom Casmurro	10
O Abolicionismo	9
Os Lusíadas	11
O Mandarim	9

Quarto caso: 4-grama: A tabela 27 apresenta os resultados obtidos por cada livro em séries geradas considerando o uso das probabilidades dos tetragramas.

Tabela 27. Número de palavras identificadas em cada livros para séries do quarto caso.

Livro	Nº de palavras identificadas
A Escrava Isaura	11
A Viuvinha	12
Dom Casmurro	11
O Abolicionismo	10
Os Lusíadas	11
O Mandarim	10

Quinto caso: 5-grama: A tabela 28 apresenta os resultados que cada livro obteve em séries geradas a partir das probabilidades dos pentagramas.

Tabela 28. Número de palavras identificadas em cada livro para séries do quinto caso.

Livro	Nº de palavras identificadas
A Escrava Isaura	14
A Viuvinha	13
Dom Casmurro	14
O Abolicionismo	11
Os Lusíadas	12
O Mandarim	13

GERAÇÃO DE SÉRIES PELO MÉTODO *MARKOVIANO*

Para o caso das séries de aproximação geradas pelo método Markoviano, o processo é similar ao adotado na seção anterior, contudo não há a contagem do tempo para calcular as probabilidades, já que esse método não faz uso das probabilidades para a geração das séries. Para esse método, as séries são geradas até o quarto grau.

Primeiro caso: 1º Grau: A tabela 29 apresenta os resultados obtidos por cada livro na geração de séries pelo método *Markoviano* para o primeiro grau.

Tabela 29. Número de palavras identificadas em cada livro em séries do primeiro grau.

Livro	Nº de palavras identificadas
A Escrava Isaura	5
A Viuvinha	8
Dom Casmurro	6
O Abolicionismo	5
Os Lusíadas	6
O Mandarim	7

Segundo caso: 2º Grau: A tabela 30 apresenta os resultados obtidos por cada livro na geração de séries do segundo grau do método *Markoviano*.

Tabela 30. Número de palavras identificadas em cada livro em séries do segundo grau.

Livro	Nº de palavras identificadas
A Escrava Isaura	7
A Viuvinha	13
Dom Casmurro	14
O Abolicionismo	11
Os Lusíadas	12
O Mandarim	13

Terceiro caso: 3º Grau: A tabela 31 apresenta os resultados obtidos por cada livro na geração de séries do terceiro grau do método *Markoviano*.

Tabela 31. Número de palavras identificadas em cada livro em séries do terceiro grau.

Livro	Nº de palavras identificadas
A Escrava Isaura	15
A Viuvinha	16
Dom Casmurro	15
O Abolicionismo	14
Os Lusíadas	15
O Mandarim	17

Quarto caso: 4º Grau: A tabela 32 apresenta os resultados obtidos por cada livro através da geração de séries do quarto grau do método *Markoviano*.

Tabela 32. Número de palavras identificadas em cada livro em séries do quarto grau.

Livro	Nº de palavras identificadas
A Escrava Isaura	17
A Viuvinha	17
Dom Casmurro	18
O Abolicionismo	17
Os Lusíadas	19
O Mandarim	18

A tabela 33 apresenta a média de palavras identificadas em séries geradas pelos dois métodos, obtida por cada livro.

Tabela 33. Média de palavras identificadas em cada livro em séries dos dois métodos.

Livro	Média de palavras (n-gramas)	Média de palavras (Markoviano)
A Escrava Isaura	8,4	11,00
A Viuvinha	**9,8**	13,50
Dom Casmurro	9,2	13,25
O Abolicionismo	7,8	11,75
Os Lusíadas	9,2	13,00
O Mandarim	8,8	**13,75**

Resultados: Considerando apenas o método de geração por n-gramas, o livro com a melhor média de palavras identificadas por série foi "A Viuvinha", já pelo método *Markoviano* o melhor resultado obtido foi do livro "O Mandarim". Considerando os dois métodos, o livro com o melhor resultado foi "A Viuvinha". Na comparação entre os métodos, ficou claro que o método *Markoviano* se aproxima da língua portuguesa mais rapidamente que o método baseado nas probabilidades dos n-gramas, mesmo o pior resultado obtido por um livro no método markoviano ainda foi melhor que resultado obtido pelo método dos n-gramas.

4.3.2 Análise dos resultados obtidos pelos livros em língua inglesa

Assim como foi feito com os livros escritos em língua portuguesa, também foi realizado uma análise comparativa dos resultados obtidos em livros da língua inglesa mediante geração de séries de aproximação por caracteres.

GERAÇÃO DE SÉRIES POR N-GRAMAS

O mesmo método empregado na comparação dos resultados dos livros em língua portuguesa é adotado para a comparação dos resultados de obras literárias escritas em língua inglesa. A tabela 34 apresenta os resultados com os tempos dos cálculos das probabilidades dos n-gramas de cada livro. A tabela a seguir apresenta os valores referentes aos tempos necessários para calcular as probabilidades dos n-gramas.

Tabela 34. Tempo de cálculo das probabilidades dos n-gramas.

Livro	Tempo de execução (Minutos)
Hamlet	15
A Christmas Carol	11
Alice's Adventures in Wonderland	3
The Adventures of Huckleberry Finn	26
The Raven	4
The Wonderful Wizard of Oz	9

Primeiro caso: 1-grama: A tabela 35 apresenta os resultados obtidos por cada livro através da geração de séries de aproximação considerando as probabilidades das letras.

Tabela 35. Número de palavras identificadas em cada livro para séries do primeiro caso.

Livro	Nº de palavras identificadas
Hamlet	2
A Christmas Carol	3
Alice's Adventures in Wonderland	3
The Adventures of Huckleberry Finn	4
The Raven	3
The Wonderful Wizard of Oz	3

Segundo caso: 2-grama: A tabela 36 apresenta os resultados obtidos por cada livro em séries geradas considerando as probabilidades dos digramas.

Tabela 36. Número de palavras identificadas em cada livro para séries do segundo caso.

Livro	Nº de palavras identificadas
Hamlet	3
A Christmas Carol	4
Alice's Adventures in Wonderland	4
The Adventures of Huckleberry Finn	5
The Raven	4
The Wonderful Wizard of Oz	4

Terceiro caso: 3-grama: A tabela 37 apresenta os resultados obtidos em séries geradas considerando as probabilidades dos trigramas por cada livro.

Tabela 37. Número de palavras identificadas em cada livro para séries do terceiro caso.

Livro	Nº de palavras identificadas
Hamlet	7
A Christmas Carol	7
Alice's Adventures in Wonderland	7
The Adventures of Huckleberry Finn	7
The Raven	8
The Wonderful Wizard of Oz	8

Quarto caso: 4-grama: A tabela 38 apresenta os resultados obtidos por cada livro em séries geradas a partir das probabilidades dos tetragramas.

Tabela 38. Número de palavras identificadas em cada livro para séries do quarto caso.

Livro	Nº de palavras identificadas
Hamlet	9
A Christmas Carol	9
Alice's Adventures in Wonderland	10
The Adventures of Huckleberry Finn	11
The Raven	11
The Wonderful Wizard of Oz	9

Quinto caso: 5-grama: A tabela 39 apresenta os resultados que cada livro obteve em séries geradas a partir das probabilidades dos pentagramas.

Tabela 39. Número de palavras identificadas em cada livro para séries do quinto caso.

Livro	Nº de palavras identificadas
Hamlet	12
A Christmas Carol	13
Alice's Adventures in Wonderland	13
The Adventures of Huckleberry Finn	14
The Raven	13
The Wonderful Wizard of Oz	12

GERAÇÃO DE SÉRIES PELO MÉTODO *MARKOVIANO*

O método de comparação dos resultados para o método Markoviano obedece ao mesmo padrão estabelecido para a comparação de livros escritos em português.

Primeiro caso: 1º Grau: A tabela 40 apresenta os resultados referentes às séries geradas pelo primeiro grau do método Markoviano por cada livro.

Tabela 40. Número de palavras identificadas em cada livro para séries do primeiro grau.

Livro	Nº de palavras identificadas
Hamlet	5
A Christmas Carol	6
Alice's Adventures in Wonderland	7
The Adventures of Huckleberry Finn	9
The Raven	5
The Wonderful Wizard of Oz	5

Segundo caso: 2º Grau: A tabela 41 apresenta os resultados obtidos por cada livro através da geração de séries pelo segundo grau do método *Markoviano*.

Tabela 41. Número de palvras identificadas em cada livro para séries do segundo grau.

Livro	Nº de palavras identificadas
Hamlet	10
A Christmas Carol	10
Alice's Adventures in Wonderland	12
The Adventures of Huckleberry Finn	13
The Raven	9
The Wonderful Wizard of Oz	11

Terceiro caso: 3º Grau: A tabela 42 apresenta os resultados obtidos por cada livros em séries geradas pelo terceiro grau do método *Markoviano*.

Tabela 42. Número de palavras identificadas em cada livro para séries do terceiro grau.

Livro	Nº de palavras identificadas
Hamlet	14
A Christmas Carol	15
Alice's Adventures in Wonderland	17
The Adventures of Huckleberry Finn	19
The Raven	13
The Wonderful Wizard of Oz	17

Quarto caso: 4º Grau: A tabela 43 apresenta os resultados obtidos com a geração de séries do quarto grau do método Markoviano para cada livro.

Tabela 43. Número de palavras identificadas em cada livro para séries do quarto grau.

Livro	Nº de palavras identificadas
Hamlet	17
A Christmas Carol	18
Alice's Adventures in Wonderland	19
The Adventures of Huckleberry Finn	22
The Raven	18
The Wonderful Wizard of Oz	19

A tabela 44 apresenta a média de palavras identificadas obtidas por cada livro em ambos os métodos analisados.

Tabela 44. Média de palavras identificadas em cada livro em séries dos dois métodos.

Livro	Média de palavras (n-gramas)	Média de palavras (Markoviano)
Hamlet	6,6	9,2
A Christmas Carol	7,2	9,8
Alice's Adventures in Wonderland	7,4	11,00
The Adventures of Huckleberry Finn	**8,2**	**12,60**
The Raven	7,8	9,00
The Wonderful Wizard of Oz	7,2	10,40

Resultados: Em ambos os métodos, o livro que obteve o melhor resultado foi "*The Adventures of Huckleberry Finn*". Na comparação entre os métodos, novamente o método *Markoviano* obteve melhor desempenho, indicando ser o método que se aproxima de modo mais rápido da língua estudada.

Capítulo 5

Conclusões e Trabalhos Futuros

A realização deste trabalho resultou no desenvolvimento de um programa que permite realizar um estudo maior sobre a língua portuguesa nos âmbitos da influencia da probabilidade na formação de textos e do valor de sua entropia. O programa desenvolvido permite calcular os valores das probabilidades e da entropia a partir de textos inseridos pelo usuário e é possível também gerar a partir desse mesmo texto de forma automática séries de aproximação para a língua portuguesa de três formas possíveis.

Este trabalho permite reproduzir para o português as experiências realizadas por *Shannon* para o inglês em 1948 e apresentadas na obra *A Mathematical Theory of Communication* que demonstrou a relevância do conhecimento da linguagem para realizar uma codificação apropriada e obter uma maior eficiência de um canal de comunicação.

Devido ao aumento na geração de informações e na preocupação com a sua segurança, cada vez mais surge a necessidade de mecanismos capazes de compactar e codificar essas informações. Os resultados obtidos no nosso programa podem revelar informações úteis para aplicações que lidam com o problema da compactação e da segurança de informações. Além de aplicações nas áreas de codificação e compactação, diversas outras aplicações, pertencentes a inúmeras áreas, como transmissão de dados, recuperação de dados, busca e correção de palavras, também se utilizam das informações referentes às probabilidades das letras, digramas e até trigramas.

As aplicações desenvolvidas para lidar com textos escritos em inglês não terão necessariamente o mesmo desempenho se aplicada ao português, mesmo possuindo o mesmo alfabeto, devido às diferenças entre as linguagens, dessa forma se faz necessário realizar um estudo sobre as características para cada linguagem caso se pretenda adequar essa aplicação a uma nova língua, e o programa desenvolvido permite agilizar um eventual estudo para a língua portuguesa levantando as informações referentes às probabilidades e entropia. O programa desenvolvido neste projeto permite realizar estudos comparativos entre textos escritos tanto em inglês como em português e avaliar as diferenças entre os dois casos.

5.1 Contribuições

O nosso projeto deixa como contribuição um estudo realizado para a língua portuguesa de forma mais detalhada que os encontrados na literatura, apresentando resultados seguindo o mesmo método utilizado anteriormente para o inglês.

O programa desenvolvido tem como capacidade realizar estudos a partir de textos tanto em inglês como em português de forma automatizada e permite realizar todas as etapas apresentadas na experiência para o inglês no mesmo ambiente. Permite também ao usuário determinar o tamanho das séries geradas conforme sua conveniência.

O programa estende o que foi apresentado no trabalho para língua inglesa, calculando as probabilidades da ocorrência do 4-grama e do 5-grama e com isso é possível gerar séries de aproximação através dessas probabilidades e apresentar o cálculo da entropia para os níveis quatro e cinco.

O programa também acrescenta o cálculo das probabilidades das palavras contidas no texto para o terceiro nível e assim permite a geração de séries de aproximação geradas pelas palavras até o nível três. O programa calcula a entropia associada às probabilidades das palavras nos níveis um, dois e três.

5.2 Dificuldades

Para a realização do projeto surgiram diversas dificuldades que se iniciaram já no processo de pesquisa, quando um dos locais planejados como fonte de informação, no caso a Universidade Federal de Pernambuco, encontrava-se com suas bibliotecas funcionando de forma atípica devido à greve dos funcionários daquela instituição, sem que fosse possível acessar seu acervo.

Uma dificuldade encontrada decorreu da falta de explicação detalhada do processo de geração dos digramas a partir de um texto levando a duas interpretações possíveis explicadas na seção 3.2.1.

A escolha da linguagem de programação (Java) utilizada também se apresentou em alguns momentos como um elemento dificultador pela dificuldade de se implementar as funções que envolviam a geração aleatória de números.

O processamento dos cálculos de probabilidade tanto dos n-gramas quanto das palavras envolve cálculos que podem demora mais de uma hora para sua realização dependendo da configuração da máquina utilizada e isso consumiu um tempo de desenvolvimento acima do esperado.

Outra dificuldade encontrada decorreu da necessidade de realizar em alguns textos utilizados nos teste um processo de filtragem além do disponível na ferramenta, pois estes textos apresentavam símbolos além do planejado e erros de impressão.

A dificuldade principal encontrada foi a falta de publicações e trabalhos sobre o tema realizados para o português. Os trabalhos em geral se restringem a apresentar as tabelas de probabilidade das letras e até dos digramas e o cálculo da entropia, mas não apresentam a geração de séries de aproximação. A falta de trabalhos similares impediu de realizar uma comparação das séries de aproximação da mesma forma que o realizado para a língua inglesa.

5.3 Trabalhos futuros

O trabalho, apesar de ter sido realizado em um linguagem de programa orientada a objetos (Java), não faz uso de todas suas características sendo interessante reorganizar os códigos de forma modularizada para permitir que a manutenção e a extensão do programa possa ser realizada mais facilmente e de forma mais eficiente.

Resolver ou atenuar a questão da contabilização exagerada do espaço, através da mudança na forma como o programa verifica o final de cada linha.

A aplicação de técnicas de refatoração dos códigos também se faz necessária para prover uma maior legibilidade destes com implicações também na manutenção e extensão do programa.

Permitir ao usuário definir quais são os elementos de seu alfabeto alterando os níveis de filtragem, que por sua vez necessitam considerar um número maior de símbolos que podem estar presentes em um texto.

Implementar algumas melhorias na interface do programa, como o reposicionamento dos botões e o aumento na área de exibição dos valores tornando-o mais fácil de ser utilizado.

Estender o programa para permitir a geração de séries de aproximação pelo processo *Markoviano* para palavras.

Bibliografia

[1] SAMPAIO, Inês Sílvia. Conceitos e modelos da comunicação. Disponível em: <http://www.uff.br/mestcii/ines1.htm> Acesso em: 22 de novembro de 2007.

[2] SHANNON, Claude E. *The Mathematical theory of communication.* The Bell System Technical Journal, V. 27, p. 379-423, julho. 1948.

[3] *The Shannon and Weaver model.* Disponível em:<http://discovery.bits-pilani.ac.in/Homepage/disciplines/languages/My%20Web%20Sites/Profiles/Nirban/model.doc > Acesso em: 22 de Novembro de 2007.

[4] SILVA, Joel. Informação, codificação e segurança de dados, Brasília, c 02, março. 1998.

[5] ABRANSOM, N. *Information Theory and Coding*. McGraw-Hill Book Company, 1963.

[6] PRATT, Fletcher. *Secret and Urgent*, Blue Ribbon Books, 1939

[7] BRAGA, Bruno R. Analise de freqüência de línguas. UFRJ. Março. 2003

[8] FERNANDES, Rafael M.S., AZEVEDO, Tiago S. Teoria da Informação e suas Aplicações em Compressão e Aleatoriedade. UFRJ, maio, 2006

[9] R., Janow. *Shannon entropy applied to productivity of organizations*. Disponível em <http://ieeexplore.ieee.org/Xplore/login.jsp?url=/iel5/8871/28027/01252225.pdf&arnumber=1252225> Acesso em : 22 de novembro de 2007

[10] SCHEIDER, Tom. *Introduction to Tom Schneider's Laboratory.* Disponível em: <http://www.ccrnp.ncifcrf.gov/~toms/introduction.html> . Acesso em: 22 de novembro de 2007

[11] Project Gutenberg, disponível em : <http://www.gutenberg.org/wiki/Main_Page> , Acesso em: 23 de novembro de 2007

[12] Domínio Público, disponível em: <http://www.dominiopublico.gov.br/> , Acesso em 23 de novembro de 2007