

Inteligência Artificial

Trabalho 2

Gustavo Alves Rodrigues (121151605)

Marcos Vinícius Treviso (121150107)

Thayson Rafael Karlinski (111151990)

1 – Análise do programa:

O corretor ortográfico implementado depende das probabilidades à priori, probabilidade à posteriori, e a probabilidade de erro dado uma letra. Para isto, foi criado uma classe específica denominada *gerador_probabilidades*.

1.1 – Gerador de probabilidades:

Nessa classe é possível gerar três tipos de probabilidades:

- Dado que o alfabeto permite letras e bigramas minúsculas e maiúsculas.
- Dado que o alfabeto permite letras e bigramas acentuadas.
- Dado que o alfabeto permite números.

Isso porque o trabalho foi feito para analisar erros ortográficos na língua portuguesa, além disso, a possibilidade dos números é acrescentada devido ao fato de muitos usuários cometerem erros de digitar um número ao invés de uma letra.

As probabilidades à priori (letras) e as probabilidades à posteriori (bigramas) foram calculadas em relação a uma base de texto, utilizando métodos simples de amostragem. As bases de texto utilizadas foram retiradas de livros em modo texto na Internet, e encontram-se no diretório *books/* na raiz da pasta do trabalho. Segue abaixo uma tabela indicando seus respectivos nomes e tamanhos.

Nome	Tamanho
A Metamorfose	120 KB
O Apanhador no Campo de Centeio	404 KB
Sherlock Holmes – Obra Completa	3.6 MB

Entretanto, as probabilidades dos erros foram aleatoriamente inseridas a partir da posição da letra relativa ao teclado português no formato ABNT2, onde os integrantes do grupo optaram pelos valores mostrados na tabela abaixo.

Acertar a letra:	Letras ao lado:	Outras letras:
60%	35%	5%

Tabela 1. Probabilidade de erro dada uma letra.

Segundo os dados da Tabela 1 é possível observar que a probabilidade de acertar a letra é maior do que errar, e errar uma letra que está logo ao lado é maior do que uma que está longe. Dito isso, é importante salientar também que esses dados não representam sua real veracidade da probabilidade de erro num texto digitado em português.

1.2 – Corretor ortográfico:

Para o corretor ortográfico funcionar, é necessário as probabilidades geradas pelo gerador de probabilidades, e de um arquivo contendo pares de palavras, tal que a primeira palavra corresponde a errada e a segunda corresponde a correta, onde toda palavra deve estar dentro do alfabeto das probabilidades geradas, pois senão elas não terão probabilidades associadas (tanto à priori como à posteriori).

Cada instância do corretor ortográfico tem como objetivo analisar um par de palavras do arquivo de palavras, por isso, foi criado um método de ajuda para colocar todas as palavras do arquivo para uma lista, desse modo a iteração dos pares ocorre sobre a lista e não sobre o arquivo em aberto.

Após setada as dependências do corretor ortográfico, é inicializado o algoritmo de Viterbi, com a opção *booleana* de normalizar as probabilidades no modelo.

O algoritmo de Viterbi se resume em fazer uma filtragem com maximização sobre os estados num instante t . Desse modo, foi feito com que o algoritmo consiga a maior probabilidade de transição de uma letra para outra, gerando então a palavra mais provável de acordo com a maximização dessas probabilidades de transições.

Ao fim do algoritmo de Viterbi, é feito a procura do erro através do caminho mais provável de um estado x_1 até o estado x_{t-1} , percorrendo os antecessores do último estado maximizado. E com isso é encontrado a palavra gerada pelo corretor, indicando qual seria a sequência de estados mais provável (palavra gerada) dada a evidência (palavra errada).

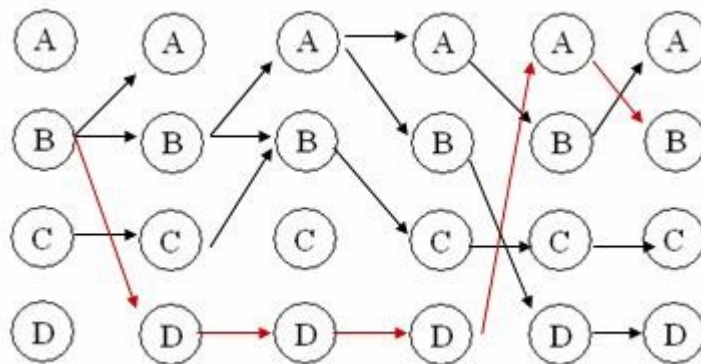


Figura 1. Representação do algoritmo de Viterbi.
(retirado de: <http://compbio.pbworks.com>)

Para a exibição dos resultados na tela, são utilizadas funções para medir a similaridade de duas palavras. As duas funções foram extraídas de repositórios públicos na Internet.

- *Levenshtein distance:*
http://rosettacode.org/wiki/Levenshtein_distance#Python
- *Dice's coefficient:*
[http://en.wikibooks.org/wiki/Algorithm_Implementation/Strings/Dice
%27s_coefficient#Python](http://en.wikibooks.org/wiki/Algorithm_Implementation/Strings/Dice_%27s_coefficient#Python)

Suas funcionalidades no programa se resumem em comparar a palavra gerada com a errada e com a correta, identificando assim estatísticas de erro do programa. Como observação, é interessante dizer que o erro em relação a cada caractere da palavra gerada em relação a palavra correta pode ser visto em razão dessa similaridade entre as duas palavras.

2 – Resultados:

A probabilidade de erro está levando em consideração o algoritmo de Levenshtein para analisar a similaridade entre a correta e a gerada pelo Viterbi. O alfabeto usado foi sensível ao caso, acentuado e com possibilidade de usar dígitos.

palavra errada | palavra correta → palavra mais provável [similiaridade com a correta]

<i>Palavras/Base Texto</i>	A Metamorfose	O Apanhador ...	Sherlock Holmes ...
caza casa	caza [75.00%]	caza [75.00%]	caza [75.00%]
m3sa mesa	mesa [100%]	mesa [100%]	mesa [100%]
inferno interno	interno [100%]	inderno [85.71%]	interno [100%]
pezo peso	peso [100%]	peso [100%]	peso [100%]
gogador jogador	górador [57.14%]	gogador [85.71%]	gorador [71.43%]
gakeria galeria	gameria [85.71%]	gameria [85.71%]	gameria [85.71%]
tijela tigela	timela [83.33%]	timela [83.33%]	timela [83.33%]
cadeifa cadeira	cadeira [100%]	cadeira [100%]	cadeifa [85.71%]
senjora senhora	senhora [100%]	senhora [100%]	senhora [100%]
af ar	ar [100%]	ar [100%]	ar [100%]
musica música	musica [83.33%]	mueica [66.67%]	musica [83.33%]
mae mãe	mãe [100%]	mar [33.33%]	mar [33.33%]
negocio negócio	negocio [85.71%]	negocio [85.71%]	negocio [85.71%]
coração coração	coração [100%]	coração [100%]	coração [100%]
Média Aritmética:	90.73	85.7978	85.9678
Desvio Padrão:	12.6013	17.8318	17.5212

Tabela 2. Resultados dada palavras erradas ao programa.

De acordo com os resultados obtidos, é possível ver que o algoritmo conseguiu corrigir palavras cujo bigrama da palavra errada não seja muito comum e o bigrama da palavra correta seja bastante comum. Entretanto, devemos também levar em consideração a posição do caractere errado no bigrama, onde seguem as regras das probabilidades de erro dado uma evidência.

Sendo assim, é correto afirmar que o programa se comportou de maneira esperada, que é corrigir apenas os erros mais grotescos e se atrapalhar com os outros. Os principais motivos é que a comparação está sendo apenas em cima de bigramas, e que as probabilidades dos erros não são coerentes com o domínio da língua portuguesa.

Podemos visualizar mais claramente essa conclusão através da obtenção das probabilidades gerais de aparecimento de um bigrama, onde o bigrama acentuado é maior que seu não acentuado correspondente.

A Metamorfose	
gó: 0.3068%	go: 0.1698%
ãe: 0.1109%	ae: 0.0000%
ão: 0.8929%	op: 0.0247%

óp: 0.0288%	op: 0.0247%
õe: 0.0767%	oe: 0.0123%

O Apanhador no Campo de Centeio	
áx: 0.0154%	ax: 0.0030%
ãe: 0.0503%	ae: 0.0009%
ão: 0.9782%	ao: 0.0793%
ôx: 0.0047%	ox: 0.0034%

Sherlock Holmes – Obra Completa	
áx: 0.0026%	áx: 0.0023%
ãe: 0.0050%	ãe: 0.0015%
ão: 0.9446%	ão: 0.1131%
ça: 0.1801%	ça: 0.1598%
õe: 0.0638%	õe: 0.0238%

Na penúltima linha da tabela de resultados, a palavra desejada era **mãe**, onde a probabilidade do bigrama **ãe** era uma parte fundamental para a corretude do algoritmo. No caso da base textual A Metamorfose, como esse bigrama acentuado foi muito maior que o não acentuado (além das outras regras já citadas), o algoritmo conseguiu descobrir o melhor caminho correto. Já nas outras duas bases, isso não aconteceu.

Também foi realizado testes com a palavra correta no lugar da errada, segue abaixo uma tabela com os resultados obtidos.

palavra correta | palavra errada → palavra mais provável [similiaridade com a correta]

Palavras/Base Texto	A Metamorfose	O Apanhador ...	Sherlock Holmes ...
casa casa	casa [100%]	casa [100%]	casa [100%]
mesa mesa	mesa [100%]	mesa [100%]	mesa [100%]
interno interno	interno [100%]	interno [100%]	interno [100%]
peso peso	peso [100%]	peso [100%]	peso [100%]
jogador jogador	morador [71.43%]	mogador [85.71%]	morador [71.43%]
galeria galeria	gameria [100%]	galeria [100%]	galeria [100%]
tigela tigela	tirela [83.33%]	tigela [100%]	tigela [100%]
cadeira cadeira	cadeira [100%]	cadeira [100%]	cadeira [100%]
senhora senhora	senhora [100%]	senhora [100%]	senhora [100%]
ar ar	ar [100%]	ar [100%]	ar [100%]

música música	musica [83.33%]	mueica [66.67%]	musica [83.33%]
mãe mãe	mãe [100%]	mar [33.33%]	mar [33.33%]
negócio negócio	negocio [85.71%]	negocio [85.71%]	negocio [85.71%]
coração coração	coração [100%]	coração [100%]	coração [100%]
Média Aritmética:	94.5571	90.8157	90.9857
Desvio Padrão:	9.1083	18.4841	18.1376

Tabela 3. Resultados dada palavras corretas ao programa.

De acordo com o resultados obtidos na Tabela 3, conseguimos visualizar que o corretor ortográfico não teve problemas em corrigir palavras que estavam certas, ou seja, na grande maioria dos casos o caminho mais provável foi justamente o caminho mais correto. Com isso, fica comprovado que apesar de não usar probabilidades reais dos textos em português, o algoritmo se comportou bem de acordo com as bases obtidas. Também deve ser levado em consideração que todos os testes realizados foram avaliados sobre um alfabeto específico, caso testes sejam feitos em outros tipos de alfabeto, os resultados certamente serão diferentes, principalmente para palavras acentuadas.