Cifra de Vigenère

Técnicas de *Data Mining* para Criptoanálise

Mestrado em Sistemas de Dados e Processamento Analítico

Segurança e Privacidade em Sistemas de Armazenamento e Transporte de Dados

Joel Tiago Ribeiro

Introdução

Pretende-se com este trabalho combinar técnicas existentes de *data mining* com os tradicionais métodos de criptoanálise da cifra de Vigenère. O resultado desta combinação deve ser uma aplicação onde seja possível complementar a metodologia existente, no intuito de se obter melhores resultados e desempenhos.

Criptologia

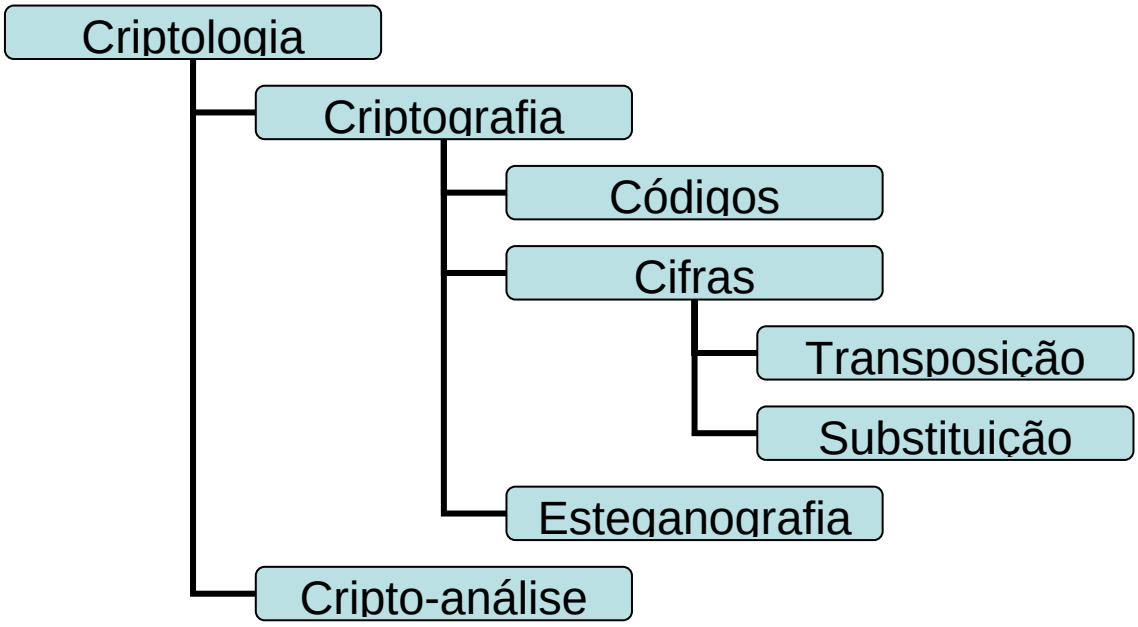

Figura 1 – Criptografia

Criptologia é a disciplina científica que reúne e estuda os conhecimentos (matemáticos, computacionais, psicológicos, filológicos, etc.) e técnicas necessários à criptoanálise (solução de criptogramas) e à criptografia (escrita codificada).

[http://pt.wikipedia.org/wiki/Criptologia]

Criptografia

Criptografia (Do Grego kryptós, "escondido", e gráphein, "escrever") é geralmente entendida como sendo o estudo dos princípios e das técnicas pelas quais a informação pode ser transformada da sua forma original para outra ilegível, a menos que seja conhecida uma "chave secreta", o que a torna difícil de ser lida por alguém não autorizado. Assim sendo, só o receptor da mensagem pode ler a informação com facilidade.

[http://pt.wikipedia.org/wiki/Criptografia]

Criptoanálise

Criptoanálise é o ramo da criptologia que estuda formas de descodificar uma mensagem sem conhecer a chave secreta. A técnica de criptoanálise é responsável por quebrar o código da mensagem cifrada, e não em decifra-lo.

[http://pt.wikipedia.org/wiki/Criptoan%C3%A1lise]

Códigos

Os códigos podem ser protocolos de comunicação, ou seja, um "conjunto de convenções que rege o tratamento e, especialmente, a formatação de dados num sistema de comunicação". Existem códigos abertos (como o código Morse) e códigos secretos.

Códigos também podem ser um conjunto de substitutos para letras, palavras ou frases inteiras. Geralmente são colocados em livros, os chamados livros de códigos ou nomenclaturas, como duas listas em ordem alfabética. Numa delas o texto claro está em ordem alfabética (para facilitar a cifragem), seguido dos substitutos. Na outra, os códigos estão em ordem alfabética (para facilitar a decifragem), seguidos do texto limpo correspondente.

[http://www.numaboa.com/content/category/11/130/136/]

Estenografia

Estenografia é o estudo e uso das técnicas para ocultar a existência de uma mensagem dentro de outra.

[http://pt.wikipedia.org/wiki/Esteganografia]

O termo estenografia vem do grego e significa "escrita coberta". É um ramo particular da criptologia que consiste, não em fazer com que uma mensagem seja ininteligível, mas em camuflá-la, mascarando a sua presença. Ao contrário da criptografia, que procura esconder a informação da mensagem, a estenografia procura esconder a existência da mensagem.

[http://www.numaboa.com/content/category/11/65/102/]

Cifras de Transposição

As cifras de transposição misturam as letras do texto original de acordo com uma qualquer regra reversível. Por outras palavras, o texto cifrado é obtido através da permutação do texto original.

[http://www.numaboa.com/content/category/11/121/124/]

Cifras de Substituição

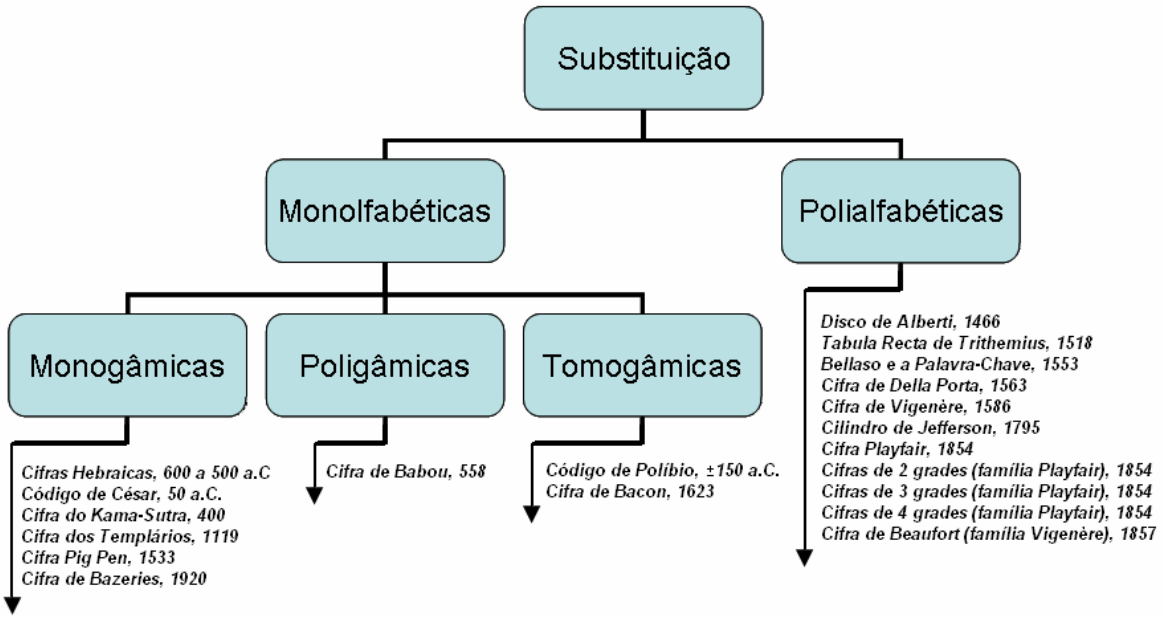

Figura 2 – Classificação das cifras de substituição

As cifras de substituição produzem criptogramas nos quais as letras do texto original, tratadas individualmente ou em grupos de comprimento constante, são substituídas por outras letras, figuras, símbolos ou uma combinação destes de acordo com um sistema predefinido e uma chave.

[http://www.numaboa.com/content/view/624/209/]

Monoalfabéticas

O sistema que substitui cada um dos caracteres de um texto limpo usando outros caracteres (letras, números, símbolos, etc) conforme uma tabela de substituição preestabelecida é o sistema mais antigo que se conhece. As tabelas de substituição contêm os caracteres que serão substituídos e os caracteres substitutos e são conhecidas como alfabeto. Quando apenas um alfabeto é aplicado, a substituição é chamada de monoalfabética.

[http://www.numaboa.com/content/view/624/209/]

Monogâmicas

Como cada um dos caracteres do texto limpo é substituído por outro, o comprimento da mensagem cifrada é igual ao comprimento da mensagem original. Da mesma forma a frequência de ocorrência das letras (números ou símbolos) do criptograma também é a mesma que a frequência de ocorrência das letras da língua usada no texto claro. Este sistema é chamado de monoalfabético porque é aplicado apenas um alfabeto (ou tabela de substituição) e classificado como substituição monogâmica (ou monográfica) porque cada caracter é tratado individualmente. A substituição monogâmica também é conhecida como uniliteral (uni = uma e literal = letra).

[http://www.numaboa.com/content/view/330/210/]

Poligâmicas

A substituição monoalfabética poligâmica tem as mesmas características da substituição simples (ou monoalfabética monogâmica), com a diferença de que se substitui grupos de caracteres do texto original por um ou mais caracteres. Portanto, o comprimento da mensagem cifrada nem sempre é o mesmo da mensagem original.

[http://www.numaboa.com/content/view/330/210/]

Tomogâmicas

Os sistemas tomográficos, também conhecidos como tomogâmicos, são aqueles nos quais cada letra é substituída por um grupo de duas ou mais letras ou números. Assim sendo, o comprimento do criptograma será necessariamente maior do que o do texto original.

[http://www.numaboa.com/content/view/330/210/]

Polialfabéticas

Quando mais de um alfabeto é utilizado para cifrar um texto limpo, o método é denominado de substituição polialfabética.

[http://www.numaboa.com/content/view/624/209/]

Os alfabetos não precisam necessariamente ter origens diferentes, por exemplo, um alfabeto latino e outro cirílico. O simples facto de alterar a ordem na sequência das letras já caracteriza um "novo" alfabeto. Por exemplo, b-c-d-...-y-z-a é um alfabeto de substituição; c-d-e-f-... é um alfabeto de substituição diferente. Se ambos forem utilizados para cifrar uma mesma

mensagem, substituindo as letras originais, então o método utilizado é uma substituição polialfabética.

[http://www.numaboa.com/content/view/625/211/]

Cifra de Vigenère

A cifra de Vigenère é um método de encriptação que usa diferentes séries da cifra de César baseadas nas letras de uma palavra-chave. Esta é uma forma simples de uma cifra de substituição polialfabética.

[http://en.wikipedia.org/wiki/Vigen%C3%A8re_cipher]

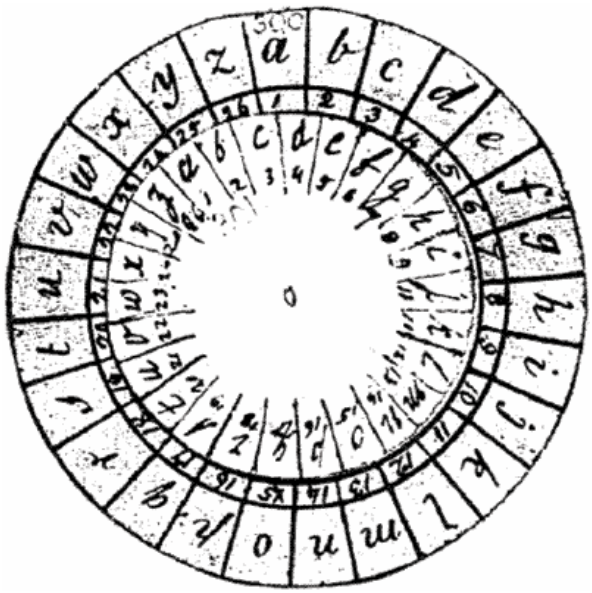

Figura 3 – Disco de Vigenère

Os processos de cifragem e decifragem da cifra de Vigenère podem ser definidos pelas seguintes equações:

- **Cifragem**

$$Cifra_i = (Texto_i + Chave_i) \bmod 26$$

- **Decifragem**

$$Texto_i = (Cifra_i - Chave_i) \bmod 26$$

De forma análoga, pode-se usar a matriz representada na figura 4. Assim sendo, o processo de cifragem será definido pela equação $Cifra_i = M_c[Texto_i; Chave_i]$, onde $M_c[x; y]$ corresponde, na matriz, ao valor da linha *x* e da coluna *y*.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
B	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A
C	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B
D	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C
E	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D
F	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E
G	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F
H	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G
I	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H
J	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I
K	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J
L	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K
M	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L
N	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M
O	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N
P	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
Q	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
R	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
S	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
T	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
U	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
V	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
W	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
X	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
Y	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
Z	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y

Figura 4 – Matriz de Vigenère

Exemplo

Neste ponto pretende-se apresentar um pequeno exemplo do processo de cifragem e decifragem da cifra de Vigenère, usando as equações apresentadas no ponto anterior.

- **Cifragem**

Texto		*PALAVRA*
Chave	+	*CHAVE*
Cifra		*RHLVZTH*

Figura 5 – Exemplo do Processo de Cifragem

– **Decifragem**

Cifra		RHLVZTH
Chave	-	CHAVE
Texto		PALAVRA

Figura 6 – Exemplo do Processo de Decifragem

Análise à Cifra de Vigenère

A análise à cifra de Vigenère começa pela determinação do tamanho da palavra-chave, através da identificação de padrões repetidos na cifra. Numa fase seguinte, sabendo o número de letras que constituem a chave e as frequências das letras da cifra e da respectiva linguagem, é possível prever potenciais valores da palavra-chave.

[http://islab.oregonstate.edu/koc/ece575/02Project/Mun+Lee/VigenereCipher.html]

Data Mining

O conceito *data mining* pode ser definido como *"a extracção não trivial de informação implícita, desconhecida, e potencialmente útil, a partir de um conjunto de dados"*.

[W. Frawley and G. Piatetsky-Shapiro and C. Matheus. "Knowledge Discovery in Databases: An Overview". *AI Magazine*: pp. 213-228. ISSN 0738-4602]

Possíveis abordagens para criptoanálise

Neste ponto pretende-se introduzir algumas técnicas de *data mining* que podem ser úteis em processos de criptoanálise.

Sequence Mining

Basicamente, esta técnica é usada para *"pesquisar padrões estatisticamente relevantes (Motifs) em sequências de dados"* – tipicamente séries temporais, com valores discretos –.

[http://en.wikipedia.org/wiki/Sequence_mining]

Na criptoanálise, a técnica de *sequence mining* poderá ser particularmente útil na pesquisa de padrões de letras – que se repetem ao longo da cifra –, permitindo determinar o tamanho da chave com mais precisão.

Modelos de Previsão

Os modelos de previsão constituem o tipo de *data mining* mais conhecido e usado. Estes modelos são usados para prever o valor de um atributo, a partir de um conjunto de outros atributos conhecidos. Assim sendo, podem ser usados modelos de previsão – como classificadores Naive Bayes, árvores de decisão, entre outros –, para se prever as palavras a decifrar. A ideia é, considerando que algumas letras da palavra-chave estão erradas, conseguir encontrar as palavras correctas através de métodos de previsão.

Graph Mining

Recorrendo a grafos, é possível representar o relacionamento entre as diversas palavras, e pontuações, de uma qualquer linguagem. Assim sendo, é possível fazer a reconstrução de um texto, cujas pontuações, acentuações e espaços foram omitidos – facto que acontece na cifra de Vigenère –, e com eventuais erros.

Conceitos

Neste ponto pretende-se introduzir alguns conceitos usados na área de *data mining* e que podem ser úteis em processos de criptoanálise.

Top K

O Top K pode ser definido como sendo uma estrutura de dados onde é possível armazenar um máximo de K elementos. Nesta, o critério de selecção consiste num valor de ranking dos elementos, permanecendo na estrutura um máximo de *K* elementos, ordenados por ranking. Assim sendo, existem para o efeito dois vectores ordenados, um para os valores de ranking dos elementos, e outro para os próprios elementos. Sempre que um elemento é inserido ou "empurrado" para a posição *K* + 1, será eliminado.

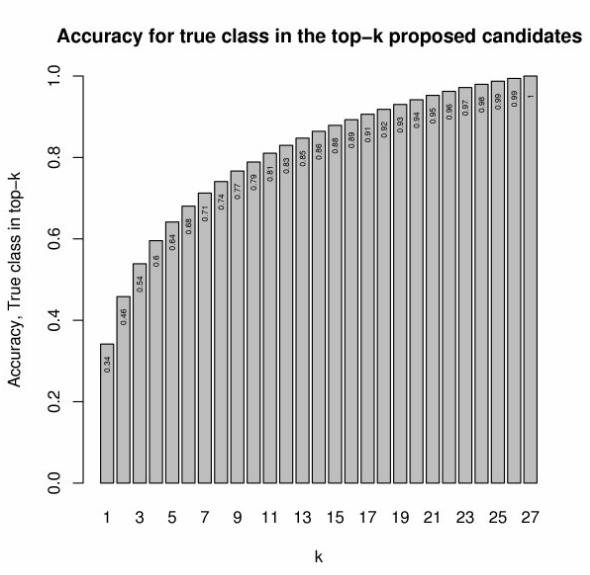

Figura 7 – Exemplo de uma aplicação do Top K

Aplicação: Vigenère Cipher

Um dos objectivos deste trabalho prendeu-se com a definição e desenvolvimento de uma aplicação de criptoanálise à cifra de Vigenère.

Objectivo

Dada uma palavra-chave e uma cifra, pretende-se identificar quais as letras da chave que estão incorrectas, sugerindo a sua correcção.

Estruturas de Dados

Neste ponto pretende-se descrever as estruturas de dados implementadas na aplicação Vigenère Cipher.

Histograma

Será implementado um histograma para representar as frequências das letras de uma qualquer linguagem. Através desta estrutura será possível identificar quais as letras com mais probabilidade de ocorrência, não só na linguagem mas também na chave.

Árvore *n*-ária

Será implementada uma árvore *n*-ária para representar as palavras, letra a letra. Através desta estrutura será possível identificar se uma qualquer palavra pertence à linguagem, mesmo que algumas das suas letras estejam trocadas.

Grafo

Será implementado um grafo para representar o relacionamento das palavras e pontuações de uma determinada linguagem. Através desta estrutura será possível reconstruir frases, ou identificar as palavras ou pontuações que possam surgir depois de uma dada palavra.

Algoritmo

Neste ponto pretende-se descrever o algoritmo da aplicação Vigenère Cipher, que pode ser dividido nas fases de treino e de previsão.

Fase de Treino (Aprendizagem)

A fase de treino consiste em construir, a partir de um conjunto de textos dado pelo utilizador, o histograma, a árvore *n*-ária, e o grafo.

Fase de Previsão

A fase de previsão começa com a execução dos processos tradicionais de criptoanálise da cifra de Vigenère, ou seja, o cálculo do tamanho da chave e a determinação dos possíveis valores da chave.

Usando um dos valores da chave determinados, começa-se por decifrar a cifra. Através do resultado da decifragem e do grafo, serão determinados os seus Top 10 textos mais similares. Para finalizar, usando estes textos, reconstrói-se cada letra da chave calculando o seu *score* (grau de certeza).

Demonstração

Neste ponto pretende-se demonstrar as funcionalidades da aplicação Vigenère Cipher.

Codificação/Descodificação

Esta funcionalidade permite ao utilizador cifrar e decifrar um qualquer texto, usando a cifra de Vigenère.

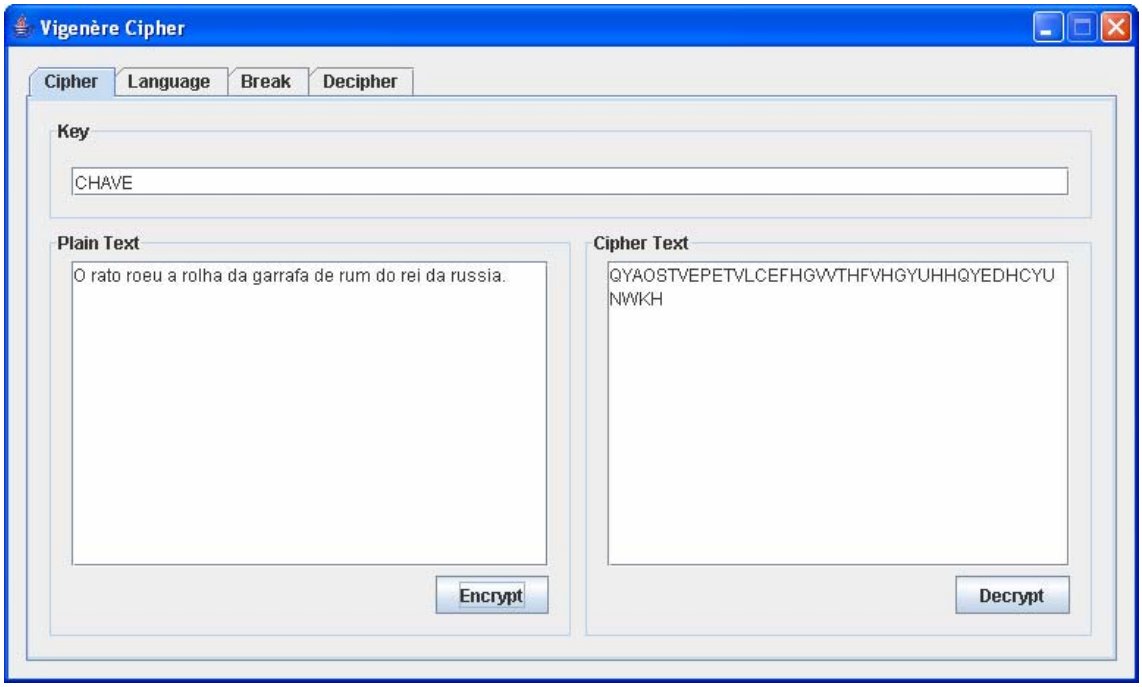

Figura 8 – Vigenère Cipher: Codificação/Descodificação

Fase de treino

Esta funcionalidade permite ao utilizador executar a fase de treino. Como já foi referido, esta fase consiste na construção, através de um conjunto de textos dados pelo utilizador, do histograma, da árvore *n*-ária, e do grafo.

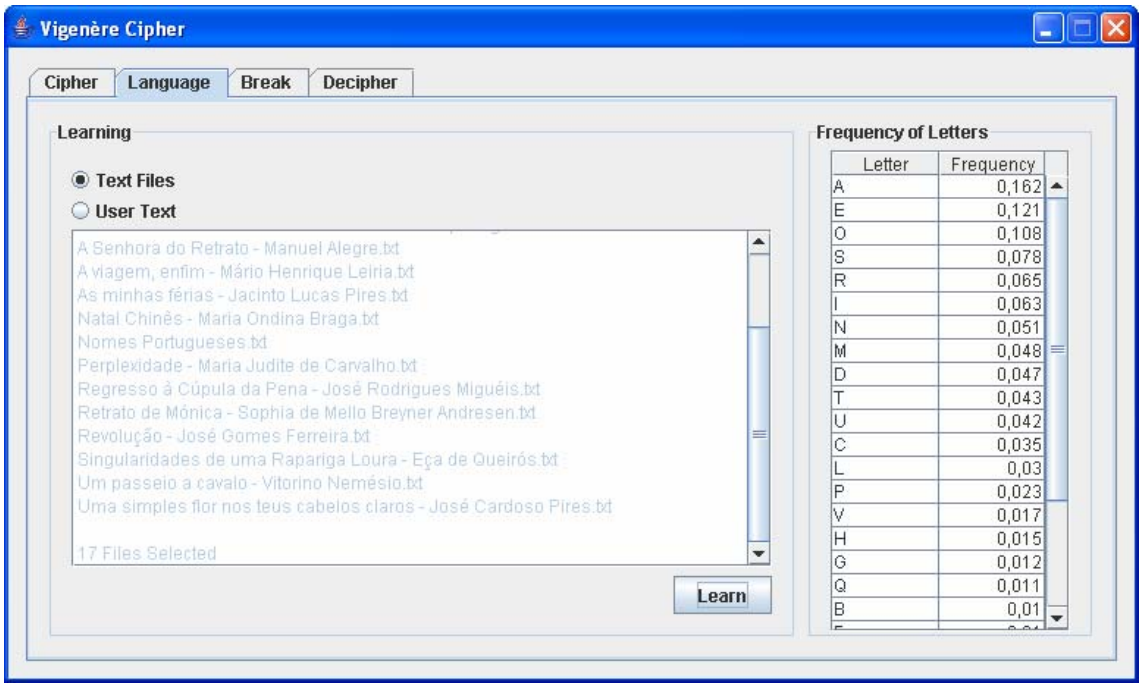

Figura 9 – Vigenère Cipher: Fase de Treino

Análise tradicional

Esta funcionalidade permite ao utilizador executar os tradicionais métodos de análise da cifra de Vigenère. Assim sendo, ao utilizador é possível determinar os Top K tamanhos de chave e os respectivos valores.

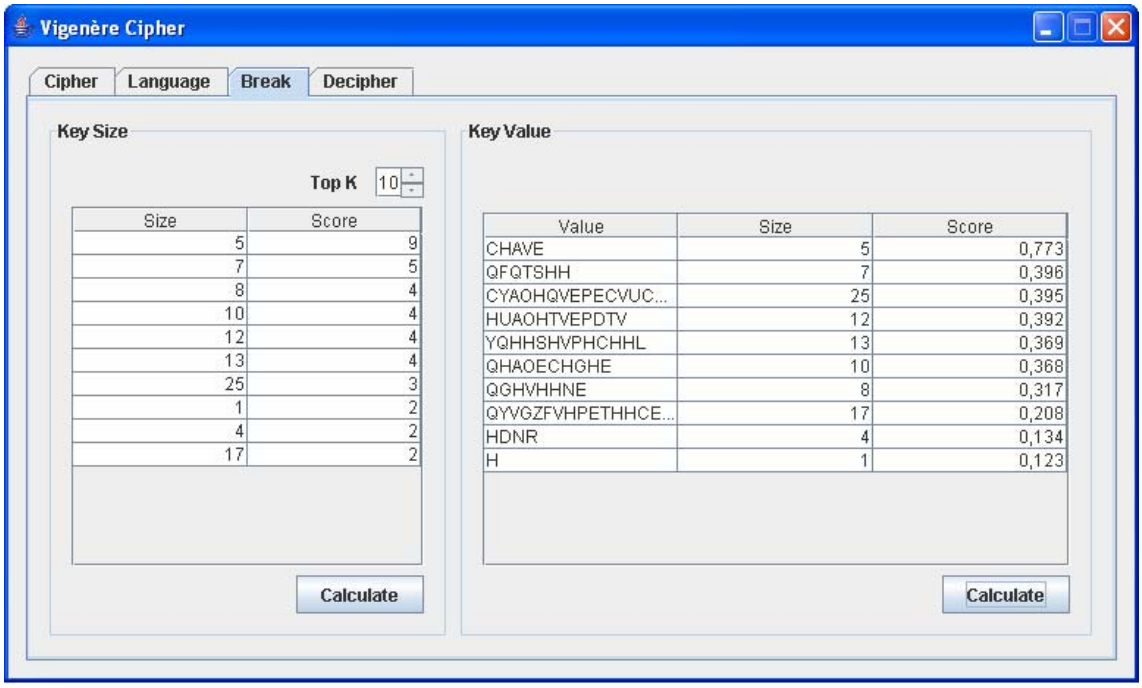

Figura 10 – Vigenère Cipher: Análise Tradicional

Análise recorrendo a técnicas de *data mining*

Esta funcionalidade permite ao utilizador executar a análise à cifra de Vigenère, recorrendo a técnicas de *data mining*. Assim sendo, ao utilizador é possível determinar quais as letras da chave introduzida que estão potencialmente erradas. É ainda apresentado ao utilizador o grau de certeza de cada uma das letras (*Score*) e de toda a chave (*Best Match*).

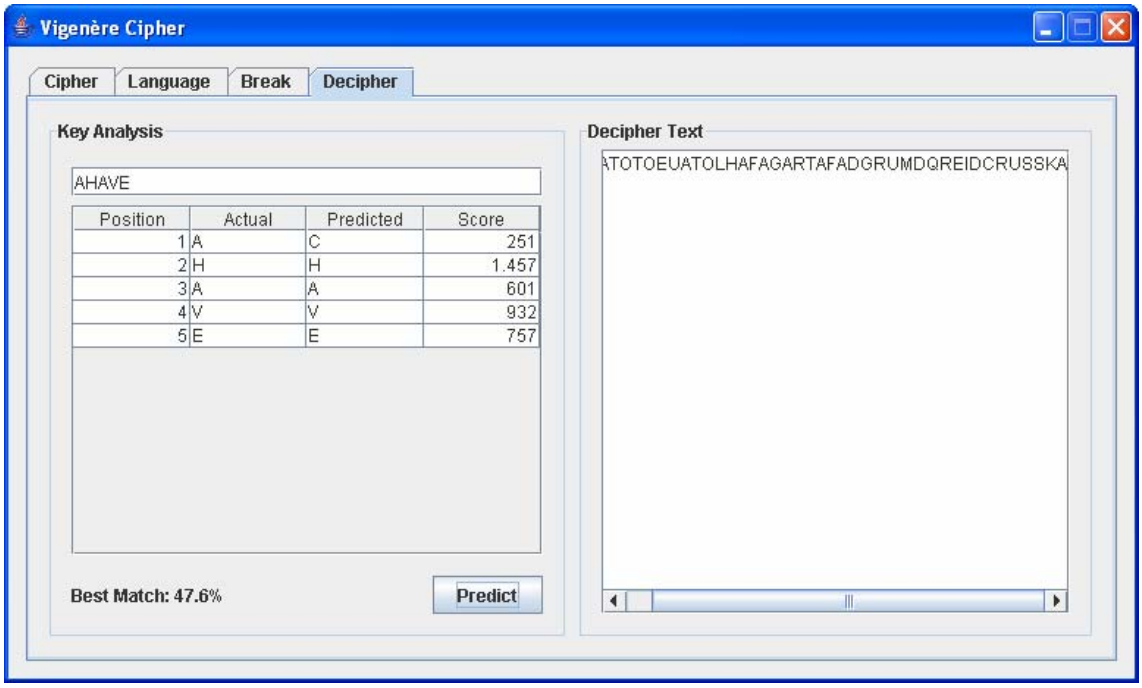

Figura 11 – Vigenère Cipher: Análise recorrendo a técnicas de *data mining*

Conclusões

Neste ponto pretende-se, em jeito de conclusão, identificar os pontos fracos e fortes da abordagem sugerida e implementada na aplicação Vigenère Cipher.

Pontos Fortes

Identificam-se os seguintes pontos fortes:

- A boa capacidade de resposta para cifras pequenas.
- O facto do algoritmo ser adaptativo, ou seja, consegue automaticamente adaptar os seus parâmetros no sentido de minimizar o seu espaço de pesquisa.

Pontos Fracos

Identificam-se os seguintes pontos fracos:

- O facto desta abordagem depender da fase de treino (aprendizagem).
- Os resultados obtidos não são exactos, podendo haver casos onde não seja possível encontrar a palavra-chave correcta.

Trabalho Futuro

Sugere-se como trabalho futuro:

- **Implementação de outras funcionalidades** como por exemplo a reposição de espaços, pontuações e acentuações.
- **Implementação de outros métodos** como por exemplo a previsão do valor da palavra-chave através de métodos baseados na força bruta, usando como referência o grau de certeza da chave (*Best Match*).
- **Teste ao desempenho e eficácia** para comparar a abordagem sugerida aos métodos tradicionais.
- **…**