COH-PIAH

1 Prólogo

Neste último exercício da Parte 1, iremos praticar não só o que vimos até agora no curso mas também outra habilidade importante de um programador: utilizar e interagir com código escrito por terceiros. Aqui, você não irá implementar o seu programa do zero. Você irá partir de um programa já iniciado e irá completá-lo. Na verdade, esse é o caso mais comum na indústria de software, onde muitos desenvolvedores trabalham colaborativamente em um mesmo programa.

2 Introdução

Manuel Estandarte é monitor na disciplina Introdução à Produção Textual I na Universidade de Pasárgada (UPA). Durante o período letivo, Manuel descobriu que uma epidemia de COH-PIAH estava se espalhando pela UPA. Essa doença rara e altamente contagiosa faz com que indivíduos contaminados produzam, involuntariamente, textos muito semelhantes aos de outras pessoas. Após a entrega da primeira redação, Manuel desconfiou que alguns alunos estavam sofrendo de COH-PIAH. Manuel, preocupado com a saúde da turma, resolveu buscar um método para identificar os casos de COH-PIAH. Para isso, ele necessita da sua ajuda para desenvolver um programa que o auxilie a identificar os alunos contaminados.

3 Detecção de autoria

Diferentes pessoas possuem diferentes estilos de escrita; por exemplo, algumas pessoas preferem sentenças mais curtas, outras preferem sentenças mais longas. Utilizando diversas estatísticas do texto, é possível identificar aspectos que funcionam como uma "assinatura" do seu autor e, portanto, é possível detectar se dois textos dados foram escritos por uma mesma pessoa. Ou seja, essa "assinatura" pode ser utilizada para detecção de plágio, evidência forense ou, neste caso, para diagnosticar a grave doença COH-PIAH.

4 Traços linguísticos

Neste exercício utilizaremos as seguintes estatísticas para detectar a doença:

1. Tamanho médio de palavra: Média simples do número de caracteres por palavra.
2. Relação Type-Token: Número de palavras diferentes utilizadas em um texto divididas pelo total de palavras.
3. Razão Hapax Legomana: Número de palavras utilizadas uma única vez dividido pelo número total de palavras.

4. Tamanho médio de sentença: Média simples do número de caracteres por sentença.

5. Complexidade de sentença: Média simples do número de frases por sentença.

6. Tamanho médio de frase: Média simples do número de caracteres por frase.

5 Funcionamento do programa

A partir da assinatura conhecida de um portador de COH-PIAH, seu programa deverá receber diversos textos e calcular os valores dos diferentes traços linguísticos desses textos para compará-los com a assinatura dada. Os traços linguísticos que seu programa deve utilizar são calculados da seguinte forma:

1. Tamanho médio de palavra é a soma dos tamanhos das palavras dividida pelo número total de palavras.

2. Relação Type-Token é o número de palavras diferentes dividido pelo número total de palavras. Por exemplo, na frase "O gato caçava o rato", temos 5 palavras no total (o, gato, caçava, o, rato) mas somente 4 diferentes (o, gato, caçava, rato). Nessa frase, a relação Type-Token vale $4/5 = 0,8$;

3. Razão Hapax Legomana é o número de palavras que aparecem uma única vez dividido pelo total de palavras. Por exemplo, na frase "O gato caçava o rato", temos 5 palavras no total (o, gato, caçava, o, rato) mas somente 3 que aparecem só uma vez (gato, caçava, rato). Nessa frase, a relação Hapax Legomana vale $3/5 = 0,6$

4. Tamanho médio de sentença é a soma dos números de caracteres em todas as sentenças dividida pelo número de sentenças (os caracteres que separam uma sentença da outra não devem ser contabilizados como parte da sentença).

5. Complexidade de sentença é o número total de frases divido pelo número de sentenças.

6. Tamanho médio de frase é a soma do número de caracteres em cada frase dividida pelo número de frases no texto (os caracteres que separam uma frase da outra não devem ser contabilizados como parte da frase).

Após calcular esses valores para cada texto, você deve compará-los com a assinatura fornecida para os infectados por COH-PIAH. O grau de similaridade entre dois textos, a e b, é dado pela fórmula:

$$S_{ab} = \frac{\sum_{i=1}^{6} \|f_{i,a} - f_{i,b}\|}{6} \tag{1}$$

onde,

- S_{ab} é o grau de similaridade entre os textos a e b;
- $f_{i,a}$ é o valor de cada traço linguístico i no texto a; e
- $f_{i,b}$ o valor de cada traço linguísticoi no texto b.

No nosso caso, o texto b não é conhecido, mas temos a assinatura correspondente: a assinatura de um aluno infectado com COH-PIAH. Ou seja, sabemos o valor de $f_{i,b}$ que é dado como valor de entrada do programa.

Perceba que quanto mais similares a e b forem, menor S_{ab} será. Para cada texto, você deve calcular o grau de similaridade com a assinatura do portador de COH-PIAH e, no final, exibir qual texto mais provavelmente foi escrito por algum aluno infectado (ou seja, o texto com assinatura mais similar à assinatura dada).

Observação: texto tirado do curso introdução a ciência da computação parte um, localizado na plataforma coursera.