European Network for Academic Integrity

Atividade do estudante [resultado O1-A-4, pt, licença CC BY 4.0, 23 agosto 2018]

Como funciona a deteção de plágio de textos

Data: 2018-08-23

Citação:

[autor] Foltýnek, Tomáš, Mendel University em Brno, República Checa

[título] Como funciona a deteção de plágio de textos

[data] 2018-08-23

[fonte] http://www.academicintegrity.eu/wp/all-materials]

[tradução] Laura Ribeiro, Faculdade de Medicina, Universidade do Porto, Portugal; Sandra F. Gomes, Faculdade de Medicina, Universidade do Porto, Portugal.

[data de acesso]

Como funciona a deteção de plágio de textos

Iremos falar apenas sobre a semelhança de textos. A semelhança de figuras, sons ou vídeos é reconhecida por métodos completamente diferentes e deixaremos esses problemas de lado.

Os *softwares* apropriados para a correspondência de texto (ou sistema de deteção de plágio) devem tratar não apenas da cópia exata mas também da sua reordenação, paráfrase, resumo ou mesmo da tradução para outro idioma. Ainda é um problema considerado difícil e quem desenvolve os sistemas profissionais, muitas vezes, tem de optar entre velocidade e precisão.

Temos um documento suspeito de ser plágio de outro(s) documento(s). Vamos chamá-lo de documento suspeito que pode ser potencialmente plagiado. A nossa tarefa é encontrar possíveis documentos de origem usados pelo autor do documento suspeito.

O processo completo tem três fases (Stamatatos et al., 2015):

Fase 1 - **Recuperação de Candidatos.** Naturalmente, não faz sentido comparar o documento suspeito com tudo o que foi escrito anteriormente. Portanto, é essencial reduzirmos as fontes alvo. Por isso, apenas escolhemos documentos que merecem ser analisados mais detalhadamente, quer seja a partir de uma base de dados extensa de acordo com os metadados, quer seja com um mecanismo de pesquisa (*web*) que use consultas bem direcionadas.

Fase 2 - **Análise detalhada.** A partir da fase 1, temos um conjunto limitado de documentos candidatos disponíveis. Destes, precisamos de selecionar apenas aqueles que realmente podem ser as fontes de plágio e identificar as passagens reutilizadas.

Fase 3 - **Pós-processamento.** A fase 2 pode identificar muitos resultados, alguns deles são sobrepostos e outros são de pouca relevância. É por isso que se segue a terceira fase. A sua finalidade é filtrar e visualizar os resultados. O resultado é um documento suspeito em que as partes reutilizadas são indicadas por cor diferente e existe uma hiperligação para os documentos originais.

Vamos concentrar-nos agora na segunda fase. Existem diferentes abordagens (Meuschke & Gipp, 2013):

Similaridade lexical

É também chamada de análise de similaridade de sequências de texto. Começa pela remoção de informações do documento sem significado e outros passos de pré-processamento, que são executados para aumentar a velocidade e a precisão da análise. O objetivo é revelar uma mudança na ordem das palavras; diferentes formas de palavras, reescrevendo frases inteiras, parágrafos ou capítulos. Provavelmente, o método mais comum é o *"chunking"* (junção), no qual o documento é dividido em *chunks* (sequências de palavras consecutivas), esses *chunks* são impressos e armazenados numa base de dados. A partilha de *chunks* comuns serve então como medida de similaridade.

Análise de semântica

A análise de semântica é uma análise do significado do texto. Ao substituir sinónimos, diferentes formas de análise estatística e cálculo sofisticado da distância semântica, obtemos um conjunto de documentos que estão mais relacionados com o texto suspeito.

Uma forma possível para fazer isto é usar o modelo de espaço vetorial. Neste modelo, olhamos para o documento como um vetor num espaço vetorial n-dimensional [onde n é o número de palavras que ocorrem em ambos os documentos (união definida)]. O valor de cada coordenada representa o número de ocorrências de uma determinada palavra no documento, geralmente ponderada pela frequência geral da tal palavra. Para aumentar a precisão da análise semântica podem ser utilizados vários esquemas de ponderação, reordenação e redução de dimensão. Assim, a similaridade de dois documentos ou passagens é calculada como o cosseno de dois vetores, que expressa o ângulo entre eles. As duas passagens mais semelhantes são os vetores correspondentes ao ângulo com menor valor. Note-se que este método pode levar-nos a potenciais fontes mas não fornece evidência clara de que o autor do documento suspeito realmente copiou o texto.

Análise estilométrica

A análise estilométrica baseia-se na composição de sequências, expressões usadas, frases, preposições, pessoas, números, etc., com a qual podemos reconhecer o estilo de escrita de um autor e identificar locais onde o estilo de escrita muda. Esses locais, provavelmente, indicam a mudança de um autor. Se tivermos documentos suficientes escritos por um autor, podemos confirmar ou refutar a autoria de um texto suspeito. No entanto, este método ainda não é muito confiável e não fornece evidência clara de plágio.

Recursos:

Stamatatos, E., Potthast, M., Rangel, F., Rosso, P., & Stein, B. (2015). Overview of the PAN/CLEF 2015 Evaluation Lab. In J. Mothe et al. (Eds.), *Experimental IR Meets Multilinguality, Multimodality and Interaction. 6th International Conference of the CLEF Initiative (CLEF 15)* (pp. 518–538). Berlin Heidelberg New York: Springer. https://doi.org/http://dx.doi.org/10.1007/978-3-319-24027-5_49

Meuschke, N., & Gipp, B. (2013). State-of-the-art in detecting academic plagiarism. *International Journal for Educational Integrity*, *9*(1), 50–71. Disponível em http://www.ojs.unisa.edu.au/index.php/IJEI/article/view/847

Notas para os professores

Este documento fornece informações realmente breves sobre os métodos de correspondência de texto. O objetivo é obter a visão geral principal. Os leitores mais interessados devem consultar a revisão de Meuschke & Gipp (2013) para encontrar mais detalhes.