Deteção Automática de Plágio

Proposta de Dissertação de Mestrado

Orientador: João Paulo Cordeiro

1 Objetivos

A *Deteção Automática de Plágio* (DAP) é uma área relativamente recente [1, 2, 3], derivada das áreas de Recuperação da Informação e Semelhança Documental. Há um conjunto de particularidades interessantes, de um foro mais criminal, subjacentes à DAP. De um modo geral, o objetivo é descobrir se um documento ou parte dele foi plagiado de uma outra qualquer fonte textual — *documento(s) fonte*. O plágio académico constitui atualmente uma crescente e séria preocupação de muitas instituições de ensino, a nível internacional [1, 2], motivando assim uma crescente investigação nesta área. Esta ganhou ainda mais força a partir de 2009 com a realização da primeira conferência/competição internacional de DAP ("PAN'09"), tendo-se repetido o evento por vários anos seguidos [6]. Benno Stein e Martin Potthast, da Universidade Bauhaus — Weimar, são duas figuras proeminentes na área, já com algum contributo significativo, em vários aspetos, incluindo o estabelecimento de material e métodos automáticos de avaliação de sistemas de DAP. Todavia, a comunidade reconhece que a área é complexa e ainda existe um caminho a percorrer [2], especialmente no aumento da eficácia e eficiência computacional, para grandes documentos e coleções fonte.

O trabalho aqui proposto consiste, numa primeira fase, na experimentação de algoritmos de ponta de DAP [4, 5] e, numa segunda fase, explorar e contribuir com algo de novo neste domínio. Em especial, ambiciona-se propor métodos que ajudem a diminuir a complexidade no processo de alinhamento de segmentos de texto — os segmentos plagiados com as fontes originais. As abordagens atuais continuam a denotar dificuldades na eficiência da resposta, à medida que os tamanhos dos pares de textos tendem a aumentar. Numa primeira tentativa, torna-se necessário explorar novas formas de representar as características chave do texto de modo a tornar o processo de alinhamento mais rápido, especialmente em grandes textos,

2 Tarefas a Realizar

T1 Estudo da área de *Deteção Automática de Texto*.

T2 Experiências piloto com os algoritmos existentes.

T3 Exploração e implementação de novos métodos.

T4 Escrita de um artigo científico.

T5 Escrita da dissertação.

3 Cronograma

T1 - 2 meses	T2 - 1 mês	T3 - 3 meses
T4 - 1 mês	T5 – 2 meses	

7 Referências

[1] Stein B., Koppel M., Stamatatos E. (2007). Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection. ACM SIGIR Forum, Volume 41, No. 2, Pages: 68–71.

[2] Potthast M., Stein B., Barrón-Cedeño A., Rosso P. (2010). An Evaluation Framework for Plagiarism Detection. Proceedings of the 23rd International Conference on Computational Linguistics, COLING 2010.

[3] Maurer H., Kappe F., Zaka B. (2006). Plagiarism - A Survey. Journal of Universal Computer Science, Volume 12, No. 8, Pages: 1449–1481.

[4] Haggag, O., & El-Beltagy, S. (2013). Plagiarism candidate retrieval using selective query formulation and discriminative query scoring. In Working Notes for the CLEF 2013 Conference.

[5] Unger, N., Thandra, S., & Goldberg, I. (2016, October). Elxa: Scalable Privacy-Preserving Plagiarism Detection. In Proceedings of the 2016 ACM on Workshop on Privacy in the Electronic Society (pp. 153-164). ACM.

[6] PAN. PAN @ CLEF 2016. URL: http://pan.webis.de/ (visited on 2017-07-07).

8 Contactos

João Paulo da Costa Cordeiro, Boco 6 / Gabinete 4.3

UBI, Departamento de Informática
Rua Marquês d'Ávila e Bolama
6201-001 Covilhã