PROPOR 2018

Tutorial

Linguística Forense Computacional: processamento de textos forenses em português

Rui Sousa Silva
Universidade do Porto – Faculdade de Letras / Centro de Linguística da Universidade do Porto
rssilva@letras.up.pt
@rsousasilva

Descrição:

A linguística forense computacional, que consiste na utilização de conhecimentos, métodos e ferramentas computacionais no processamento de textos de linguagem natural de natureza forense, tem sido utilizada com êxito na resolução de casos forenses, sobretudo com textos redigidos em língua inglesa. Entre as aplicações da linguística que têm recorrido com êxito a ferramentas computacionais encontram-se a análise de autoria, a deteção de plágio e a determinação de perfis sociolinguísticos. Embora alguns destes métodos e ferramentas também possam ser utilizados no processamento de textos de outras línguas, baseiam-se, normalmente, nas características da língua inglesa, relegando para segundo plano características linguísticas específicas cruciais na análise forense.

Este tutorial tem como objetivo suprimir esta lacuna, incentivando a investigação científica nesta área através do recurso a (potenciais) casos forenses com utilização de textos escritos em português, nas diversas aplicações da análise linguística forense (da análise de autoria à determinação de significados), e nas diferentes variedades linguísticas do português (da variedade europeia à variedade brasileira). O tutorial chama a atenção para o facto de esta análise ser cada vez mais crucial em casos de cibercrime, mostrando como uma abordagem computacional é fulcral para apoiar o linguista forense na tomada de decisões periciais eficientes, eficazes e informadas, sobretudo em casos com volumes de dados mais elevados. O tutorial destina-se a linguistas que pretendam ficar a conhecer melhor a investigação na área da linguística forense, bem como a especialistas em ciências de computadores que pretendam obter conhecimentos de base para conceção de ferramentas computacionais adequadas.

Parte I:

1. **Abordagens computacionais à análise de autoria: da atribuição à identificação.**
    a. Atribuição vs. Análise de autoria.
    b. Aplicações da análise se autoria.
    c. Critérios aplicáveis à análise de autoria.
    d. Desafios computacionais à análise de autoria.

Data set 1_Autoria: Este data set inclui 4 ficheiros .csv: três ficheiros com tweets de três autores (identificados como Autor_1, Autor_2 e Autor_3) e um ficheiro com tweets de autoria desconhecida. O objetivo desta tarefa é atribuir a autoria dos textos de autoria desconhecida a um dos três autores (Autor_1, Autor_2 ou Autor_3). Nota: Sustente a sua decisão em dados de análise concretos.

Intervalo para café

2. **Deteção de plágio.**
    a. Plágio como problema forense.
    b. Da deteção à análise de plágio.
    c. Plágio externo e plágio intrínseco.
    d. Tipos e estratégias de plágio.
    e. Critérios de deteção.
    f. Desafios computacionais à deteção de plágio.

Data set 2_Plágio: Este data set inclui 4 ficheiros .pdf referentes a um caso de plágio que ocorreu no jornal português *Público*. O texto da notícia (ficheiro Texto_Suspeito_ClaraBarata_EmBuscaDoAutoBronzeadorIdeal.pdf) foi acusado de plagiar, entre outros, fontes como a NewScientist e a própria Wikipedia. Quando confrontada com o caso, a jornalista responsável pelo texto negou as acusações. Comparando os documentos fornecidos, (1) verifique se a jornalista plagiou e (2) explique como sustentar computacionalmente as suas decisões.

Almoço

Parte II:

1. Perfis sociolinguísticos.
    a. Características sociolinguísticas.
    b. Critérios de determinação de perfis sociolinguísticos.
    c. Ferramentas de apoio à obtenção de perfis.

Intervalo para café

3. Determinação de significados.
    a. Análise de significados em Linguística Forense.
    b. Disputas de marcas comerciais.
    c. Deteção de linguagem ofensiva.

Caso para análise: Caso FDP

Em 2012, Manuel Leitão, responsável pela publicação de um guia de restaurantes na cidade do Porto (Portugal), o Porto Menú, foi acusado de insulto pelo antigo presidente da Câmara Municipal do Porto, Rui Rio, depois de o guia ser publicado mostrando, na capa, uma fotografia do principal mercado da cidade, o Mercado do Bolhão, com uma inscrição onde se lia: "Rio és um FDP". Inicialmente, julgou-se que a inscrição existia na fachada do Mercado; porém, verificou-se depois que ela não existia e que tinha sido resultado da manipulação da fotografia (fotomontagem) apresentada na capa da revista.

Alvo de dois processos - um cível e um criminal -, Manuel Leitão alegou que "FDP" não significava "filho da puta", como reclamava Rui Rio, mas sim "fanático dos popós", em alusão às corridas de automóveis antigos então organizadas pelo antigo Presidente da Câmara (nota: em português europeu, "popó" significa "carrinho", em linguagem infantil).

A sigla foi também utilizada noutro contexto, em que o acusado vestia uma t-shirt onde estava escrito: "quem vem e atravessa o rio; és um FDP* - *fora do Porto."

A questão que se coloca, do ponto de vista da linguística forense, é a seguinte: "FDP" significa inequivocamente "filho da puta"? Que análise poderemos fazer, utilizando ferramentas computacionais, para determinar o significado da sigla?

Literatura recomendada para ser lida antes de assistirem:

Johnson, A. & Wright, D. (2014). 'Identifying idiolect in forensic authorship attribution : an n-gram textbite approach'. *Language and Law / Linguagem e Direito*, 1(1), 70–94. http://ojs.letras.up.pt/index.php/LLLD/article/view/2443/2233

Sousa-Silva, R. (2014). 'Detecting translingual plagiarism and the backlash against translation plagiarists'. *Language and Law / Linguagem e Direito*, 1(1), 70–94. http://ojs.letras.up.pt/index.php/LLLD/article/view/2444

Sousa-silva, R., & Abreu, B. B. (2015). Plágio : um problema forense. *Language and Law / Linguagem e Direito*, 2(2), 90–113. http://ojs.letras.up.pt/index.php/LLLD/article/view/2405

Sousa Silva, R., Laboreiro, G., Sarmento, L., Grant, T., Oliveira, E., & Maia, B. (2011). "twazn me!!! ;(" automatic authorship analysis of micro-blogging messages. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 6716 LNCS). http://doi.org/10.1007/978-3-642-22327-3_16