EACH | campus capital USP LESTE
Escola de Artes, Ciências e Humanidades
Universidade de São Paulo

Programa de Pós-graduação em Sistemas de Informação
PPgSI

VI Workshop de Dissertações de Mestrado do PPgSI 2019

Atribuição autoral multicanais de textos digitais

Autoria de:	José Eleandro Custódio			
Orientação de:	Prof. Dr. Ivandré Paraboni			
Coorientação de:				
Linha de pesquisa:	[ ] Gestão e Desenvolvimento de Sistemas		[x] Inteligência de Sistemas	
Área de pesquisa:	[ ] Banco de dados [ ] Gestão de tecnologia da informação	[ ] Engenharia de software [ ] Interação humano computador	[x] Inteligência artificial [ ] Reconhecimento de padrões	[ ] Processamento gráfico [ ] Otimização
Área de aplicação:	[ ] Ambientes corporativos / Processos de negócio [ ] Economia [ ] Jogos / Jogos sérios [ ] Saúde	[ ] Bioinformática [ ] Educação / Educação a distância [x] Linguística / Língua natural [ ] Outra Qual? _____	[ ] Biometria [ ] Governo eletrônico [ ] Quimioinformática	[ ] Dispositivos móveis [ ] Internet / Redes sociais [ ] Robótica [ ] Geral*
Semestre no curso (na data do workshop):	[ ] 2º semestre	[ ] 3º semestre	[ ] 4º semestre	[x] 5º semestre
Qualificação:	[x] Qualificação já realizada em: 30/11/2018		[ ] Realização da qualificação planejada para: 30/11/2019	
Defesa:	Prazo máximo para depósito: 27/01/2020		Realização da defesa planejada para: 30/11/2019	
Publicações associadas ao projeto de mestrado:	• CUSTODIO, J. E. ; PARABONI, I. . Multichannel Open-set Cross-Domain Authorship Attribution. In: Conference and Labs of the Evaluation Forum (CLEF-2019), 2019, Lugano, Switzerland. Working Notes of the CLEF-2019 Evaluation Labs, 2019. v. 2380. • CUSTODIO, J. E. ; PARABONI, I. . Similaridade de textos aplicada à verificação autoral. In: I Congresso Internacional em Humanidades Digitais (HDRio-2018), 2018, Rio de Janeiro. Proceedings of HDRio-2018, 2018. p. 403-407. • CUSTODIO, J. E. ; PARABONI, I. . EACH-USP Ensemble Cross-Domain Authorship Attribution. In: Conference and Labs of the Evaluation Forum (CLEF-2018), 2018, Avignon. Working Notes Papers of the CLEF 2018 Evaluation Lab, 2018. v. 2125.			

Resumo do projeto de pesquisa

Contexto:
A atribuição autoral(AA) de textos digitais busca identificar o autor de um texto baseando-se nos rastros deixados pelo mesmo ao escrever. Os rastros, ou estilo de escrita, podem ser examinados através da análise da utilização das palavras, sequências de caracteres, pontuações e preferências gramaticais. No entanto, essas características são impactadas pelo domínio do texto, língua e quantidade de autores.

Problema de pesquisa:
Os métodos computacionais aplicados à atribuição autoral tendem a utilizar apenas um tipo de extração de características. No entanto, os rastros de autoria podem ser encontrados em diversos níveis linguísticos, como por exemplo, nível sintático, semântico ou lexical.

Objetivo de pesquisa:
Construir um método computacional que combine diversos tipos de conhecimento e de extração de características.

Caracterização da solução em desenvolvimento:
O método proposto irá construir classificadores que usem como extração de características n-gramas de palavras, n-gramas de caracteres, n-gramas de anotações POS e *embeddings*. Os modelos de POS e *embedings* usarão bases pré-treinadas bibliotecas prontas como *embeddings* Google e a biblioteca Python Spacy. Os modelos construídos serão então combinados em um método único que irá fornecer a classificação dos documentos gerando a marcação do autor do texto.

Fundamentos:
Os n-gramas de caracteres e palavras mais frequentes são reconhecidos na literatura por capturar elementos inconscientes da escrita.

Trabalhos relacionados:

- Sobre trabalhos que analisam um tipo de embeddings para AA: Document embeddings learned on various types o authorship attribution. Gómez-Adorno, Helena et. al (2018); Continuous N-gram Representations for Authorship Att
- Sobre trabalhos que analisam técnicas reconhecidas para AA: Authorship attribution in portuguese using character N (2017); Authorship attribution using text distortion Stamatatos, Efetathios (2017).
- Modelos matemáticos que descrevem o conceito de embeddings:
- Distributed Representations ofWords and Phrases and their Compositionality. Mikolov, Thomas et al (2013); A Neural Probabilistic Language Model. Bengio, Yoshua et al (2003).

Validação:
São usados conjuntos de dados públicos que possuam desempenho conhecido, e os experimentos serão desenvolvidos usando validação cruzada, exploração de parâmetros com grid search.

Limitações, riscos e ameaças:
Esse trabalho irá lidar primordialmente com o problema de atribuição autoral de conjunto fechado, onde os autores dos textos desconhecidos são necessariamente um dos autores do conjunto de treinamento. Demais variações poderão ser exploradas apenas para expansão do conhecimento sobre o tema

Os métodos propostos serão limitados a um conjunto de até 30 autores candidatos (30 classes), estando fora do escopo, o cenário com milhares de autores.

Contribuição científica:
O trabalho pretende aumentar o conhecimento sobre o funcionamento dos n-gramas de caracteres e outros para atribuição autoral, pretende expandir o conhecimento sobre as formas de combinação de modelos computacionais.

Contribuição tecnológica (se pertinente):

Método de pesquisa			
Gênero (escolha UM)	[ ] Pesquisa teórica	[ ] Pesquisa prática	[x] Pesquisa empírica / [ ] Pesquisa metodológica
Natureza (escolha UMA)	[ ] Pesquisa básica	[x] Pesquisa aplicada	
Abordagem (escolha UMA)	[x] Pesquisa quantitativa	[ ] Pesquisa qualitativa	[ ] Pesquisa mista (quali-quanti)
Revisão de literatura* (você pode escolher mais de uma)	[ ] Revisão narrativa [ ] Revisão descritiva [ ] Revisão de escopo	[ ] Meta-análise [x] Revisão sistemática qualitativa [ ] Revisão *guarda-chuva*	[ ] Revisão teórica [ ] Revisão realística [ ] Revisão crítica
Procedimento técnico principal (escolha UM)	[x] Pesquisa experimental [ ] Pesquisa bibliográfica [ ] Pesquisa documental [ ] Pesquisa *ex-post-facto*	[ ] Pesquisa com *survey* [ ] Estudo de caso [ ] Pesquisa participante [ ] Pesquisa-ação	[ ] Pesquisa etnográfica [ ] Teoria fundamentada em dados [ ] Ciência do projeto [ ] Outra Qual? _____
Análise de dados (você pode escolher mais de uma)	[x] Estatística descritiva [x] Estatística inferencial	[ ] Teste estatístico [ ] Análise de conteúdo	[ ] Análise do discurso [ ] Outros: _____

* Definição de tipos de revisões de literatura estabelecida por Paré, G., Trudel M-C., Jaana M., Kitsiou, S. Synthesizing Information systems knowledge: A typology of literature reviews. In: Information & Management 52, p. 183-199, 2015. DOI: 10.1016/j.im.2014.08.008

Próximas atividades:
Descrição dos experimentos e confecção da dissertação.

Opcional: Forneça um esquema gráfico que mostre aspectos de sua pesquisa. Por exemplo: um fluxograma para construção da sua solução ou um infográfico para sua proposta de pesquisa. Se necessário, use a quarta página.