V Workshop de Dissertações de Mestrado do PPgSI (2018)

FICHA DE PESQUISA

DADOS GERAIS				
Título do projeto de pesquisa	Atribuição autoral de textos digitais utilizando embeddings			
Orientando	José Eleandro Custódio			
Orientador(es)	Ivandré Paraboni			
Semestre no curso, na data do workshop	[] 2º semestre	[x] 3º semestre	[ ] 4º semestre	[ ] 5º semestre
Qualificação	[ ] Qualificação já realizada em: dd/mm/aaaa [x] Realização da qualificação planejada para: 29/10/2018			
Defesa	Prazo máximo para depósito: 29/10/2018 Realização da defesa planejada para: 01/05/2019			
Linha e Área de pesquisa	Gestão e Desenvolvimento de Sistemas: [ ] BD [ ] Engenharia de Software [ ] Gestão de TI [ ] IHC	Inteligência de Sistemas: [x] IA [ ] Processamento Gráfico [ ] Reconhecimento de Padrões		
Área de aplicação	[ ] Ambientes corporativos [ ] Bioinformática [ ] Biometria [ ] Dispositivos móveis [ ] Educação	[ ] Educação a distância [ ] Governo eletrônico [ ] Internet [ ] Jogos [ ] Jogos sérios	[x] Língua Natural [x] Linguística [ ] Processos de Negócio [ ] Quimioinformática	[ ] Redes Sociais [ ] Robótica [ ] Saúde [ ] [outro – escrever]
Publicações associadas ao projeto de mestrado	Foram publicados dois trabalhos: - "Similaridade de Textos aplicada à Verificação Autoral" apresentado em poster no HDRio2018 - EACH-USP Ensemble Cross-domain Authorship Attribution: Notebook for PAN at CLEF 2018 – Um artigo descrevendo o método utilizado no sistema de atribuição autoral utilizando na competição PAN-CLEF e que atingiu o melhor resultado geral da edição de 2018.			

DESCRIÇÃO DO PROJETO DE PESQUISA	
Contextualização / motivação	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	Métodos tradicionais de processamento de língua natural foram amplamente estudados para AA. No entanto, técnicas recentes de embeddings foram pouco estudadas e não há uma análise definitiva de suas eficácias para a AA.
Objetivo geral da pesquisa	Comparar os diferentes métodos de embeddings com métodos baselines estabelecidos para a atribuição autoral.
Trabalhos relacionados	**Sobre trabalhos que analisam um tipo de embeddings para AA:** Document embeddings learned on various types of n-grams for cross-topic authorship attribution. Gómez-Adorno, Helena et. al (2018); Continuous N-gram Representations for Authorship Attribution. Sari et al (2017); **Sobre trabalhos que analisam técnicas reconhecidas para AA:** Authorship attribution in portuguese using character N-grams Markov, Ila et al (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).
Proposta para solução	Estender o trabalho realizado para o PAN-CLEF-2018 através de: - Aplicação de embeddings aplicados nos níveis caracter, palavra e POS tag. - Aplicar um comitê de máquina para combinar os diferentes métodos.
Dados	O projeto utilizará conjuntos de dados disponibilizados pelos organizadores do PAN-CLEF para edição de 2018 e demais edições.
Validação	Será feito através da comparação de medidas estatísticas entre resultados obtidos e o baseline PAN-CLEF 2018.

	Juntamente com uma discussão sobre a efetividades dos embeddings.
Limitações	O trabalho irá utilizar apenas técnicas computacionais para atribuição autoral e não encorpará análise julgamental.
Resultados esperados	Contribuições científicas: é esperado um avanço no entendimento dos embeddings para AA.
	Contribuições tecnológicas: Não há

MÉTODO DE PESQUISA [para cada item, marque todas as opções adequadas para seu projeto] **(basear-se em: http://www.ufrgs.br/cursopgdr/downloadsSerie/derad005.pdf)**			
Gênero	[ ] Pesquisa teórica	[ ] Pesquisa prática	[ ] Pesquisa empírica [ ] Pesquisa metodológica
Natureza	[ ] Pesquisa básica/pura	[x] Pesquisa aplicada	
Objetivo	[ ] Pesquisa exploratória	[ ] Pesquisa descritiva	[x] Pesquisa explicativa [ ] [outro – escrever]
Abordagem	[x] Pesquisa quantitativa	[ ] Pesquisa qualitativa	[ ] Pesquisa mista (quali-quanti)
Procedimento técnico principal	[x] Pesquisa experimental [ ] Pesquisa bibliográfica [ ] Pesquisa documental [ ] Pesquisa *ex-post-facto* [ ] Pesquisa de levantamento	[ ] Pesquisa com *survey* [ ] Estudo de caso [ ] Pesquisa participante [ ] Pesquisa-ação [ ] Pesquisa etnográfica	[ ] Pesquisa netnográfica [ ] Teoria fundamentada em dados (*Grounded theory*) [ ] Ciência do projeto (*Design science*) [ ] [outro – escrever]
Técnica(s) / Instrumento(s) de coleta de dados	[ ] Medição [ ] Questionário [ ] Entrevista [ ] Grupos focais	[ ] Formulário [ ] *Benchmark*	[ ] Observação (direta / participante) [ ] Diário de campo / notas de campo [ ] Análise documental (ou de artefatos) [x] Extração
Técnica(s) de análise de dados	Análise quantitativa: [x] Estatística descritiva [ ] Estatística inferencial [ ] [outro – escrever]		Análise qualitativa: [ ] Análise de conteúdo [ ] Análise do discurso [ ] [outro – escrever]

CRONOGRAMA [altere conforme necessário]

	2017												2018												2019											
	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12
Estudo da literatura																																				
Revisão sistemática																																				
Definição da proposta																																				
Execução da proposta																																				
Coleta/obtenção de dados																																				
Análise dos dados																																				
Validação da proposta																																				
Preparo da qualificação																																				
Exame de qualificação																																				
Escrita da dissertação																																				
Escrita de artigo																																				
Depósito da dissertação																																				
(Outros)																																				

As bibliografias abaixo, ou correlatas, são bastante úteis para lhe ajudar a preencher esta ficha:

- **Gerhardt, T. E. & Silveira, D. T. Métodos de Pesquisa. EAD – Série Educação a Distância. Editora UFRGS, 2009. Disponível em: http://www.ufrgs.br/cursopgdr/downloadsSerie/derad005.pdf**
- **Wazlawick, R. S. Metodologia de Pesquisa para Ciência da Computação. 2ª. Ed. Elsevier, 2014.**
- **Creswell, J. W. Projeto de pesquisa: métodos qualitativo, quantitativo e misto. 3ª. Ed. Bookman/Artmed, 2010.**