| | |
|---|---|
| Programa: | Engenharia Elétrica |
| Área de Concentração: | Engenharia de Computação |
| Aluno: | Jefferson Caros de Mendonça |
| Orientador: | Edson Satoshi Gomi |
| Curso: | Mestrado |
| Data de Ingresso: | 14/09/2016 |
| Título: | Modelo para Mineração de Dados - Análise das Perguntas e Respostas do Site Stack Overflow |

## Resumo

As transformações no cenário tecnológico provocam mudanças constantes que exigem atualização contínua. Como fonte de estudos e atualização, profissionais da área de computação diariamente recorrem à diversas fontes de informação, dentre elas destaca-se o site StackOverflow, maior comunidade de perguntas e respostas onde os usuários podem aprender, trocar experiências e compartilhar conhecimento. O objeto deste projeto de pesquisa é propor um modelo de mineração de dados que consiga catalogar as perguntas desta comunidade, facilitando a busca por informações e detecção dos principais questionamentos discutidos .

**Palavras-chave:** information retrieval, text mining, text classification, automated text categorization, keyword extraction, document indexing, StackOverflow.

# Modelo para Mineração de Dados - Análise das Perguntas e Respostas do Site Stack Overflow

## 1 Introdução

Não raro a evolução na área computacional se desenvolve em ritmo acelerado, novos algoritmos, técnicas para processamento de imagem, computação distribuída, *linguagens de programação* aprimoram suas API's constantemente, construindo e desconstruindo métodos e muitas vezes incorporam novos paradigmas. Essas transformações no cenário tecnológico provocam mudanças constantes exigindo atualização contínua. Para dar vazão a este dinamismo, profissionais da área de computação diariamente recorrem a cursos presenciais, a distância, livros, revistas e claro *websites*. Dentre estas mídias, merece destaque o fórum *Stack Overflow*, maior comunidade *online* de perguntas e respostas onde os usuários podem aprender, trocar experiências e compartilhar conhecimento.

### 1.1 Contextualização do Problema

O número de questões em sites de perguntas e respostas crescem diariamente, faz-se necessário categorizar os assuntos discutidos em tópicos, para uma busca mais fácil e dinâmica. Yasotha and Charles (2016) observam que a categorização manual de textos, pode ser feita somente por especialistas e requer muito tempo. Como consequência é de grande importância a categorização e classificação de documentos de forma automática.

### 1.2 Objetivos

Propor um modelo de mineração de dados sobre a comunidade online de perguntas e respostas StackOverflow, onde seja possível categorizar de forma automática os tópicos mais relevantes discutidos.

### 1.3 Justificativas

O método proposto por Arash et al. (2016) categorizou os dados do StackOverflow. Em seu projeto os autores classificaram os assuntos utilizando as tags da própria pergunta, e então utilizaram o site Wikipédia para validar o tópico encontrado e identificar a qual categoria ele pertence. A proposta desta pesquisa, é prover a categorização e classificação de textos que faça apenas uso do texto disponível nas perguntas e respostas, não utilizando recursos adicionais como tags que fornecem dica sobre qual é o tópico de um assunto específico.

### 1.4 Organização do texto

Para melhor definir qual o posicionamento do presento projeto, no capítulo seguinte será detalhado em maior profundidade os projetos que abordaram a categorização de documentos, inclusive àqueles que também fizeram uso do site StackOverflow como base

de dados. Então a proposta será detalhada quantos aos procedimentos para a indexação das perguntas e respostas, desde a seleção do conteúdo original, armazenamento em banco de dados e extração e por fim serão exibidos os resultados esperados.

# 2 Revisão da Literatura

# 3 Detalhamento da Proposta

# 4 Plano de Trabalho

## 4.1 Resultados Desejados e Validação

A análise será feita sobre os mesmo dados da pesquisa realizada por Arash et al. (2016), os resultados obtidos anteriormente serão a base para a medição da acurácia da nova proposta e então será possível aplicar o modelo desenvolvido sem a utilização de tags pré-definidas, possibilitando a construção de catálogos para outros sites de perguntas e respostas.

## 4.2 Atividades e Cronograma

Atividades e Cronograma.

# Referências


Arash, M., English, E., and Mahdi (2016). Text mining stackoverflow An insight into challenges and subject-related difficulties faced An insight into challenges and subject-related difficulties faced by computer science learners subject-related difficulties faced by computer science learners. *Journal of Enterprise Information Management Journal of Enterprise Information Management Journal of Enterprise Information Management Web of Science Database Library Review*, 29(3):255–275.

Joorabchi, A., English, M., and Mahdi, A. E. (2015). Automatic mapping of user tags to Wikipedia concepts: The case of a Q&A website – StackOverflow. *Article Journal of Information Science*, 41(5):570–583.

Manning, C. D., Raghavan, P., and Schutze, H. (2009). An Introduction to Information Retrieval. *Online*, (c):569.

Mihalcea, R. (2007). Using Wikipedia for Automatic Word Sense Disambiguation. In *Proceedings of NAACL HLT*, volume 2007, pages 142–147.

Mihalcea, R. and Tarau, P. (2004). TextRank: Bringing order into texts. *Proceedings of EMNLP*, 85:404–411.

Mihalcea, R. F. and Mihalcea, S. I. (2001). Word Semantics for Information Retrieval: Moving One Step Closer to the Semantic Web. *13th IEEE International Conference on Tools with Artificial Intelligence. ICTAI 2001*, pages 280–287.

Miotto, R. and Weng, C. (2013). Unsupervised mining of frequent tags for clinical eligibility text indexing. *Journal of Biomedical Informatics*, 46(6):1145–1151.

Posch, L. (2014). Enriching Ontologies with Encyclopedic Background Knowledge for Document Indexing. *Proceedings of the 13th International Semantic Web Conference*, pages 537–544.

Roul, R. K., Asthana, S. R., and Sahay, S. K. (2015). Automated document indexing via intelligent hierarchical clustering: A novel approach. *2014 International Conference on High Performance Computing and Applications, ICHPCA 2014*.

Udell, J. (2005). UIMA and the Blogosphere. *1803072320050822*, page 30.

Yasotha, R. and Charles, E. Y. A. (2016). Automated text document categorization. *2015 IEEE 7th International Conference on Intelligent Computing and Information Systems, ICICIS 2015*, pages 522–528.