Anais do SEPE – Seminário de Ensino, Pesquisa e Extensão
Vol. IX (2019) – ISSN 2317-7489

MODELAGEM DE TÓPICOS EM REGISTROS ELETRÔNICOS DE SAÚDE

Ivair Puerari[1]
Denio Duarte[2]

Resumo: O volume de dados na área da saúde é volumoso, pois informações de pacientes são coletados diariamente. Dentre os dados gerados, pode-se citar os registros eletrônicos de saúde. Estes registros contém dados textuais sobre as ocorrências de pacientes em setores de saúde. Esses dados podem ser utilizados no desenvolvimento de métodos para extração automática de informações úteis porém escondidos ao olho humano. Estes métodos podem fornecer avaliações rápidas e precisas sobre, por exemplo, a condição de um paciente e, assim, serem úteis nas tomadas de decisões feitas por profissionais de saúde. Para este fim, Modelagem de tópicos apresenta um conjunto de algoritmos estatísticos que analisam palavras, visando extrair de uma coleção de documentos, os principais tópicos que representam os assuntos abordados pela coleção. Modelagem de tópicos considera que tópicos são formados por distribuições probabilísticas de palavras e que documentos (textos) podem ser gerados a partir de diferentes distribuições sobre tópicos. O presente estudo tem como objetivo, realizar análise exploratório sobre duas coleções de documentos retirados de registros eletrônicos de saúde. Os dados estão contidos em formato de texto, sendo que, uma coleção de documentos contém dados de internações onde pacientes que tiveram alta hospitalar e outra coleção de documentos formado por pacientes que vieram a óbito na Unidade de Terapia Intensiva (UTI). Os registros eletrônicos de saúde foram retirados da base de dados *Medical Information Mart for Intensive Care III* (MIMIC-III) que inclui dados relacionados à saúde com 53.423 internações hospitalares clínicas que permaneceram na unidade de terapia intensiva do centro médico Beth israel Deaconess em Boston, Massachusetts entre 2001 e 2012. Os registros de internação têm dados como, diagnósticos de entrada, diagnósticos médicos, procedimentos realizados e notas sobre os eventos que se sucederam durante a internação, bem como o relatório final de internação. Inicialmente, cada documento passou por um pré-processamento, ou seja, uma limpeza do texto, onde foram removidos palavras e elementos não uteis, como pronomes, pontuação e palavras que não oferecerão vantagem na análise. Após a etapa de pré-processamento das

1 Acadêmico de Ciência da Computação, Universidade Federal da Fronteira Sul, campus Chapecó, ivair.puerari@estudante.uffs.edu.br.
2 Doutor em Ciência da Computação, Universidade Federal da Fronteira Sul, campus Chapecó, duarte@uffs.edu.br.

coleções de documentos, foi escolhido e aplicado um modelo probabilístico de tópico chamado *Latent Dirichlet Allocation* (LDA), no qual tem o intuito de classificar o texto de um documento para um tópico específico. Desta forma, com os tópicos gerados foi possível realizar uma análise exploratória sobre as coleções de documentos, a fim de verificar o assunto representado dentro de cada tópico, bem como identificar quais as intersecções e disjunções nos tópicos pertencentes às duas coleções (alta e óbito) e avaliar modelagem de tópicos na tarefa de extração automática de informação em registros eletrônicos de saúde.

Palavras-chave: Modelagem de Tópicos. Registros Eletrônicos de Saúde. Documentos. Tópicos.

Categoria: UFFS - Pesquisa

Área do Conhecimento: Ciências Exatas e da Ter

Formato: Comunicação Oral