Ovídio José Francisco

Aplicação de técnicas de Mineração de Textos para Organização e Extração de Históricos de Decisões de Documentos de Reuniões

Sorocaba, SP

15 de setembro de 2017

z

Sumário

1 PROPOSTA 3

1.1 Preparação dos documentos 3

1.1.1 Segmentação 4

Referências 5

1 Proposta

Essa seção apresenta as etapas de desenvolvimento do sistema, bem como o seu funcionamento geral, desde a preparação dos documentos até a entrega dos históricos de ocorrência ao usuário. Inicialmente será descrito a seleção e pré-processamento. Em seguida, será relatado como as técnicas de representação computacional de textos e as abordagens de mineração de texto e resgate de informação são utilizadas nesse trabalho.

O objetivo do sistema é permitir ao usuário consultar uma coleção de documentos de reuniões a fim de obter todo o histórico de ocorrências de um determinado tema pesquisado, podendo identificar nos documentos onde o tema foi mencionado como informe ou onde houve uma decisão sobre o tema.

Para isso, o sistema é divido em dois módulos principais: Módulo de preparação e manutenção e Módulo de consulta.

1.1 Preparação dos documentos

As atas são extraídas de documentos do tipo *pdf*, *doc*, *docx* ou *odt* que normalmente possuem formato binário. Aplicou-se um processo para transformar esses arquivos em texto plano.

A fim de preparar o texto e selecionar as palavras mais significativas, as atas passaram por processos de transformação os quais serão apresentados a seguir.

1. Remoção de cabeçalhos e rodapés: as atas contém trechos que podem ser considerados pouco informativos e descartados durante o pré-processamento, como cabeçalhos e rodapés que se misturam aos tópicos tratados na reunião, podendo ser inseridos no meio de um tópico prejudicando tanto o algoritmo de segmentação, quanto a leitura do texto pelo usuário.

2. Identificação de finais sentenças: devido ao estilo de pontuação desses documentos, como encerrar sentenças usando um *";"* e inserção de linhas extras, foram usadas as regras especiais para identificação de finais de sentença.

3. Redução de termos: eliminou-se a acentuação, sinais de pontuação, numerais e todos os *tokens* menores que três caracteres. Palavras de uso muito frequente como artigos, preposições e pronomes, chamadas de *stop words*, foram removidas utilizando-se uma lista de 438 palavras.

4. *Stemming*: extraiu-se o radical de cada palavra. Para isso, as letras foram convertidas em caixa baixa e aplicou-se o algoritmo *Orengo*[1] para remoção de sufixos.

1.1.1 Segmentação

Como já mencionado, uma ata registra a sucessão de assuntos discutidos em uma reunião, porém apresenta-se com poucas quebras de parágrafo e sem marcações de estrutura, como capítulos, seções ou quaisquer indicações sobre o assunto do texto. Logo, faz-se necessário descobrir quando há uma mudança de assunto no texto da ata.

Para isso, os algoritmos *TextTiling* e

1

Referências