UNIVERSIDADE FEDERAL DA BAHIA
PRÓ-REITORIA DE PESQUISA, CRIAÇÃO E INOVAÇÃO
COORDENAÇÃO DE INICIAÇÃO A PESQUISA, CRIAÇÃO E INOVAÇÃO

Plano de Trabalho do Estudante

<Observação: Favor não alterar o layout desta página de rosto. Apenas preencha os dados nos campos solicitados. A partir da segunda página estão os demais itens do modelo a serem preenchidos.>

EDITAL – PROGRAMA (Digitar o nome e número do edital – Programa (ver Edital))
EDITAL PROPCI/UFBA 01/2015 – PIBIC

Título do Plano de Trabalho:
(completo, sem abreviações)

Pré-processamento de documentos de textuais

Orientador:
(Nome completo, sem abreviações)

TATIANE NOGUEIRA RIOS

Salvador
2015

1. Objetivos específicos do estudante

Lista de objetivos específicos do estudante que será indicado para desenvolver este Plano de Trabalho.

Este plano de trabalho tem como objetivos específicos o estudo de técnicas de pré-processamento de textos, bem como o desenvolvimento de uma ferramenta com as técnicas estudas implementadas. Espera-se que a ferramenta viabilize o pré-processamento de documentos utilizando uma interface amigável na qual seja possível filtragem de documentos de textos, bem como a estruturação dos documentos de maneira a torná-los processáveis por parte dos algoritmos de extração de padrões.

Compreende-se por filtragem de documentos a eliminação de repetições de documentos; o balanceamento da coleção de documentos por reamostragem; a redução da quantidade de documentos, quando o objetivo assim permitir; verificação da existência de uma estrutura prévia nos documentos, a fim de utilizar esta informação na estruturação final da coleção; análise do tamanho dos documentos na coleção, verificando a necessidade de uma normalização dos pesos atribuídos aos termos em função do tamanho dos textos; escolha dos termos relevantes ao longo do texto, bem como dos termos a serem desconsiderados e outras tarefas comumente realizadas nesta etapa.

Por outro lado, a estruturação mais comum para dados textuais é a representação destes em um espaço vetorial no formato de tabelas atributo-valor, de maneira que cada linha corresponda a um documento da coleção e cada coluna corresponda a uma característica (dimensão) do documento, na qual cada característica pode ser representada por um termo ou a um conjunto de termos presentes na coleção de documentos. Desta maneira, a cada célula da tabela formada é associada uma medida, como a medida binária, que indica a presença ou não de um termo em um documento, a frequência de um termo em um documento e a frequência ponderada de um termo em um documento em função de sua distribuição ao longo da coleção.

Geralmente, os termos presentes nas tabelas atributo-valor são previamente analisados e tratados. Em um esforço inicial, busca-se desconsiderar dos textos termos que não representem conhecimento útil, via eliminação de *stopwords*, as quais são palavras não relevantes na análise dos textos, sendo geralmente constituídas por preposições, pronomes, artigos, interjeições e outras. Posteriormente, busca-se identificar similaridades de significados entre palavras, como em casos de variações morfológicas ou de palavras sinônimas. No caso de palavras sinônimas, pode-se fazer uso de vocabulário controlado do domínio. No caso de variações morfológicas, pode-se usar a simplificação de termos, como a radicalização, a lematização e a substantivação.

Sendo assim, o aluno deverá iniciar seu trabalho com um levantamento bibliográfico dos trabalhos relacionados a fim de definir quais desses métodos podem/devem ser utilizados, bem como quais são as tarefas mais relevantes para o pré-processamento de textos considerando o Tratamento de imprecisão e incerteza na categorização de documentos de texto.

2. Resultados específicos do estudante

Lista dos resultados específicos a serem alcançados pelo estudante que será indicado para desenvolver este Plano de Trabalho, entre eles a capacitação a ser atingida ao final dos 12 meses.

Com a ferramenta a ser desenvolvida, espera-se que a etapa de pré-processamento contribua para o tratamento de imprecisão e incerteza na categorização de documentos de texto como um todo, uma vez que a boa escolha de documentos e termos representativos dos documentos evita a necessidade de ajustes nas demais etapas do processo de mineração de textos.
Também como resultado, espera-se que o trabalho realizado fomente a produção científica em eventos de iniciação científica regionais e nacionais, como por exemplo o Seminário de Pesquisa Estudantil da UFBA, o Workshop de Trabalhos de Iniciação Científica e Graduação (WTICGBASE) da Escola Regional de Computação Bahia-Alagoas-Sergipe (ERBASE) e o Concurso de Trabalhos de Iniciação Científica (CTIC) do Congresso da Sociedade Brasileira de Computação (CSBC)

3. Cronograma específico de execução

Relação itemizada, em ordem sequencial e temporal, das atividades que deverão ser realizadas pelo estudante que será indicado para desenvolver este Plano de Trabalho, ao longo do período de desenvolvimento deste Plano de Trabalho (12 meses).

1. Revisão bibliográfica. (1º ao 3º mês)
   1.1. Estudo/análise do trabalho que vem sendo desenvolvido pelos principais grupos de pesquisa na linha do projeto aqui proposto.
   1.2. Levantamento bibliográfico com escolha das principais referências a serem consideradas no pré-processamento de documentos de texto.
2. Implementação dos métodos escolhidos na etapa anterior. (4º ao 7º mês)
   2.1. Filtragem de documentos de texto.
   2.2. Estruturação de documentos de texto.
3. Coleta de documentos de texto e composição das coleções de texto a serem utilizadas. (8º mês)
4. Avaliação da ferramenta proposta. (9º ao 11º mês)
   4.1. Pré-processamento das coleções escolhidas utilizando os métodos implementados.
5. Documentação dos resultados obtidos, com posterior publicação de artigos e resumos.

Salvador, 02 de ABRIL de 2015.

Tatiane Nogueira Rios
Orientador(a)

Secretaria do Programa
Rua Basílio da Gama, 06. Canela.
Salvador – BA. 40.110-040.
Tel.: 71 3283-7968 Fax: 71 3283-7964
E-mail: pibic@ufba.br