**Universidade Federal de São Carlos**
Programa de Pós-Graduação em Ciência da Computação
Campus Sorocaba

# Solicitação de Prorrogação de Prazo de Defesa

## Relatório de Atividades Desenvolvidas

Aprovado como aluno regular em agosto de 2015, concluí as disciplinas exigidas para obtenção dos créditos e fui qualificado em agosto do ano seguinte.

O objetivo do trabalho é aplicar técnicas de recuperação de informação e mineração de texto em atas de reunião para fornecer históricos de menções a assuntos de interesse por meio de um sistema de busca. O desenvolvimento envolve basicamente o tratamento de uma base de documentos e a identificação dos assuntos nela contidos por meio de técnicas de segmentação textual e extração de tópicos. Por meio de uma interface, o sistema receberá uma consulta do usuário para proceder a busca e devolver um histórico de menções relacionadas a intenção do usuário.

Segue as principais atividades desenvolvidas:

**Contribuições:** Como parte do sistema proposto, desenvolveu-se uma ferramenta que visa a segmentação automática de documentos permitindo a configuração do preprocessamento, a detecção de sentenças e os parâmetros dos algoritmos de segmentação. Oferece uma interface gráfica para configuração e visualização dos segmentos extraídos bem como as *features* selecionadas na etapa de preprocessamento. Os algoritmos de segmentação foram avaliados comparando-os com atas segmentadas manualmente por participantes das reuniões. Por meio de análises estatísticas chegou-se ao modelo que melhor segmenta os textos no contexto das atas de reunião. A biblioteca desenvolvida para a ferramenta está disponibilizada para estudo e pode ser aproveitada na etapa de preprocessamento em outros projetos.

**Experimento com usuários:** A fim de obter referências para os algoritmos de segmentação e extração de tópicos, foram obtidas atas de reuniões do Conselhos de Pós Graduação, Conselho de Cursos e Conselho de Departamento de da UFSCar-Sorocaba. Um conjunto de seis documentos foi oferecido à profissionais que participam de reuniões desse departamento para dividir cada documento em segmentos e rotulá-los. Para isso, desenvolveu-se um *software* que permitiu aos voluntários visualizar um documento, selecionar o texto correspondente a um assunto e removê-lo para então indicar quais palavras melhor indicavam o tópico. O software possibilitou a segmentação e rotulação das atas e ao final gerou-se um aquivo contendo os dados coletados, o qual foi tratado e serviu como referência para a avaliação dos algoritmos de segmentação. Os arquivos gerados foram tratados para remoção de ruídos e ajustes para que os segmentos sempre terminem em uma sentença reconhecida pelo algoritmo, uma vez que as medidas de avaliação de segmentos recebem sentenças como unidade mínima de informação.

**Submissão de Artigo:** Como parte da revisão bibliográfica e desenvolvimento do sistema proposto, a segmentação automática das atas resultou em um artigo submetido ao ENIAC

– Encontro Nacional de Inteligência Artificial e Computacional. No artigo são descritos as principais técnicas empregadas nessa tarefa, bem como métodos mais utilizados para avaliação de segmentadores. Mostra também como o pré-processamento e configurações de parâmetros podem ser ajustados para esse tipo de documento. Para isso, utilizou-se a segmentação manual como referência de segmentação ideal a qual foi comparada com os segmentos extraídos automaticamente. Por meio de testes estatísticos descritos no artigo, chegou-se a um modelo que melhor identifica mudanças de assunto no contexto das atas.

**Sistema proposto:** Atualmente o sistema proposto recebe um conjunto de documentos divide cada um em segmentos com assunto relativamente independente. Em seguida, utiliza técnicas de extração de tópicos para fornecer as palavras que podem melhor representar o tópico de cada segmento. Para isso, possui a implementação das principais técnicas de extração de tópicos como LDA (*Latent Dirichlet Allocation*) e PLSA (*Probabilistic Latent Semantic Analysis*). Após a extração, os segmentos rotulados ficam disponíveis para consulta ou armazenamento. A segmentação dos documentos e a extração de tópicos permite a identificação trechos com um assunto de interesse.

**Universidade Federal de São Carlos**
Programa de Pós-Graduação em Ciência da Computação
Campus Sorocaba

# Solicitação de Prorrogação de Prazo de Defesa

## Plano de atividades e cronograma

Para a conclusão dos trabalhos e defesa, são necessárias as seguintes tarefas:

**1 - Módulo de Preparação e Manutenção:** Deve receber uma coleção de documentos, realizar a etapa de preprocessamento que entregará cada texto dividido em sub-documentos. Em seguida, cada sub-documento deve receber rótulos os quais irão compor uma estrutura de dados que será utilizada para consulta. Para isso, serão empregadas técnicas de segmentação textual como BayesSeg e técnicas de extração de tópicos como o LDA (*Latent Dirichlet Allocation*) e PLSA (*Probabilistic Latent Semantic Analysis*) para atribuir tópicos aos documentos.

**2 - Módulo de Busca:** O sistema deve permitir ao usuário buscar por menções sobre um assunto, bem como encontrar e navegar pelos documentos. Assim, o sistema necessita de um módulo de busca por assuntos. Nesse módulo, a *string* de entrada do usuário deve ser tratada e comparada à base de dados dos documentos já processados, para então exibir ao usuário os trechos correspondentes à busca. Deve ainda ser implementada a busca aproveitando o agrupamento dos sub-documentos em tópicos. Para isso, serão empregadas as técnicas de recuperação de informação e extração de tópicos da literatura para rankear e agrupar os sub-documentos.

**3 - Avaliação do Sistema:** O sistema final deve ser avaliado junto a usuários a fim de avaliar a eficiência do sistema em suas respostas bem como funcionalidades do ponto de vista de experiência do usuário. Para isso, novamente será necessário a ajuda de voluntários que se enquadram no perfil de usuários alvo. Após avaliação, uma eventual otimização das técnicas deve ser considerada para o aprimoramento das ferramentas.

**4 - Conclusão da Dissertação:** Redigir o texto da dissertação.

Tabela 1: Cronograma das atividades

| Atividades | Fevereiro | Março | Abril | Maio | Junho | Julho |
|---|---|---|---|---|---|---|
| 1 | X | X | | | | |
| 2 | | X | X | X | | |
| 3 | | | | X | X | |
| 4 | | | X | X | X | |