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

Plano de atividades e cronograma

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

1. **Extração de texto e pré-processamento**: Remove elementos menos significativos para o sistema proposto como *stopwords*, *stems*, numerais e palavras menores que 2 caracteres. Nessa etapa identifica-se e os cabeçalhos e rodapés por meio de uma heurística que encontra repetições das primeiras e últimas palavras do texto as quais também são removidas. Nessa etapa também são identificados os finais de sentença conforme o pseudo-código mostrado no Algoritmo 1.

2. **Segmentação**: Divide o texto em segmentos com significado relativamente independente.

   2.1 - Implementação de Segmentadores: Implementou-se algoritmos de segmentação com diferentes abordagens afim de avaliá-los no contexto do trabalho:

   Coesão léxica : Implementou-se dois algoritmos mais tradicionais baseados em coesão léxica, o **TextTiling** e o **C99** por serem amplamente referenciados e usados como baseline para comparação com métodos mais recentes.

   Modelos estatísticos/Extração de tópicos : Implementou-se também os algoritmos baseados em modelos estatísticos como **TextSeg** e **BayesSeg** os quais utilizam modelos probabilísticos muito similares à modelo de extração de tópicos e permitem a utilização de frases-pista. O LCSeg e o TopicTiling são outros segmentadores baseados em tópicos ainda não implementados. A utilização de frases-pista requer uma lista com palavras e frases que correm próximas ao final ou início de segmentos. Essas palavras foram selecionadas manualmente para serem usadas dentro contexto das atas de reunião.

   Particionamento de grafos : Implementou-se também o **MinCutSeg** que trata a segmentação textual como um problema de corte mínimo em grafos onde as sentenças correspondem aos nós e a similaridades entre as sentenças correspondem à arestas. A segmentação se dá pelo particionamento do grafo que representa o texto.

   Outros (não aplicáveis) : Há outras técnicas na literatura, porém não aplicáveis a esse contexto como as baseadas no layout do documento e elementos da fala que exigem texto semi-estruturado em formato rich-text e áudio de conversas respectivamente.

   Foi implementado também um segmentador *"fake"* que gera um segmento por sentença, para servir de *baseline*. Como alternativa pode-se implementar um segmentador randômico.

   2.2 - Avaliação dos segmentadores

   Pretende-se avaliar os segmentadores implementados usando como referência os segmentos fornecidos pelos participantes do experimento e discutir os seguintes pontos:

2.2.1 - Pk e WinDiff : Medidas de concordância entre anotações de segmentações. Quais métodos tem desempenho melhor de acordo essas medidas.

2.2.2 - Medidas tradicionais Acurácia, precisão, revocação e F1 : O quão exata deve ser uma segmentação e o quanto pode ser tolerante a segmentações que ocorrem próximas ao esperado. *(rever, pois apresentam problemas e quase não são utilizadas em segmentação).*

2.2.3 - Impacto do preprocessamento : Como o pre-processamento influencia cada algoritmo/abordagem. (Tabela em anexo).

2.2.4 - Comparação entre diferentes abordagens : Como cada abordagem responde à segmentação de atas de reunião em termos de desempenho, quantidade de segmentos gerados, falsos positivos e falsos negativos visto as particularidades desses documentos.

2.2.5 - Texto reduzido a verbos e substantivos Qual o impacto de segmentar o texto após extrair somente verbos e substantivos do texto.

2.2.6 - Textos concatenados : Discutir a performance dos segmentadores usando como base a concatenação de textos escritos em português de domínios diferentes afim de verificar a influência da falta de parágrafos e marcações de seção, bem com a linguagem compacta entre outras características sobre os algoritmos.

2.3 - Parâmetros: Durante a avaliação utilizou-se para o TextTiling e C99 os parâmetros que obtiveram melhor resultado conforme testes estatísticos onde aplicou-se o teste de Friedman com pós-teste de Nemenyi para gerar um ranking das melhores configurações para uma medida. Para TextSeg, BayesSeg e MinCutSeg, utilizou-se as configurações fornecidas pelos autores.

3. **Extração de Tópicos**: O sistema usa como extratores de tópicos o LDA, PLSA e K-Means (códigos cedidos pelo Rafael). Esses podem ser avaliados subjetivamente por meio de questionários. Há ainda os descritores fornecidos pelos participantes que podem utilizados como referência (discutir).

4. **Interface com o usuário**: A interface do sistema permite que o usuário crie uma coleção de documentos que deseja pesquisar e insira novo documentos. Por meio de um campo de busca é possível pesquisar por palavras-chave e lhe será apresento a visualização dos resultados obtidos pelo sistema.

5. **Módulo de preparação**: Recebe uma coleção de documentos os quais são pre-processados, segmentados, um extrator de tópicos agrupa os sub-documentos por tópico e identifica os descritores para cada tópico. Esses dados são armazenados internamente em uma estrutura de arquivos *texto* para os sub-documentos legíveis, *arff* para a representação textual e *csv* para os tópicos obtidos.

6. **Módulo Consulta** : É apresentado ao usuário um ranking com os resultados mais relevantes (com base nas palavras-chave).
   Deve ser ainda 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.
   A avaliação desse módulo envolve a segmentação (se os resultados apresentados contém um assunto relativamente independente relacionado com a aquilo que o usuário espera)

Dissertação

Para a conclusão da dissertação restam as seguintes tarefas:

Introdução : Explicar melhor os objetivos e justificativa.

Conceituação Teórica : Aprofundar principalmente as técnicas utilizadas em Segmentação, Extração de tópicos. Incluir uma revisão sobre Recuperação de informação.

Trabalhos relacionados Apresentar trabalhos relacionados a segmentação de textos em línguas diferentes do inglês, às diferentes aplicações dos métodos de segmentação, trabalhos relacionados a segmentação textos transcritos de reuniões com múltiplos participantes e trabalhos de recuperação de informação em sub-documentos atas de reunião (até agora 1 trabalho).

Sistema proposto Melhorar o detalhamento dos módulos. Atualizar a figura que mostra a visão geral do sistema. **Resultados**: Apresentar e discutir os resultados obtidos na segmentação e recuperação dos sub-documentos (segmentos) pelo usuário.

Conclusão : Apresentar as contribuições do trabalho e trabalhos futuros (Classificação para apontar um segmento como tratando de uma decisão ou não).

Figuras : Incluir figuras para ilustrar alguns pontos como os vales de similaridade do TextTiling e cálculo do Pk.
Melhorar a qualidade das imagens utilizado imagens vetoriais (algumas já refiz).

Cronograma

Etapa	Status	Entrega
Extração de texto e pré-processamento	Pronto	Pronto
Implementação de Segmentadores	5 segmentadores prontos	Pronto
Avaliação dos segmentadores		
Pk e WinDiff	Discutir	15-12
Medidas tradicionais Acurácia, precisão, revocação e F1	Discutir	15-12
Impacto do preprocessamento	Discutir	15-12
Comparação entre diferentes abordagens	Discutir	29-12
Texto reduzido a verbos e substantivos	Implementar e discutir	29-12
Textos concatenados	Implementar e discutir	29-12
Módulo de preparação	Pronto para essas técnicas	Pronto
Trabalhos relacionados	Escrever	15-1
Introdução	Melhorar	20-1
Conceituação Teórica	Melhorar	20-2
Módulo Consulta	Implementar técnicas de IR e extração de tópicos	20-3
Interface com o usuário	Inserir visualização/navegação por tópicos	20-3
Navegação por agrupamentos	Implementar	20-3
Avaliação do Sistema	Discutir/Planejar	10-4
Sistema proposto	Melhorar	15-4
Conclusão	Começar	20-4
Correções		20-5