Universidade de Brasília (UnB)
Departamento de Ciência da Computação (CIC)
116297 - Mineração de Textos - 2019/1

Projeto - Parte I

1 Objetivo

Elaborar um método de mineração de textos para extrair conhecimento implícito e potencialmente relevante por meio de tarefas de classificação e agrupamento de dados. Os textos empregados serão de diferentes naturezas, podendo ser do domínio financeiro, político, jurídico, educacional, notícias, resenhas, redes sociais etc. O projeto, de maneira geral, compreende a definição dos objetivos, das tarefas de mineração ou visualização de textos, implementação do método proposto e na realização de experimentos de validação. Um artigo científico descrevendo o método proposto, os trabalhos relacionados da literatura e os experimentos de validação, será escrito para propósitos de documentação do projeto.

Esse projeto será dividido em duas partes e essa especificação aborda apenas a **Parte I**.

Escolha do tema

Escolha um problema do domínio do conhecimento que envolva a mineração de textos e defina uma tarefa a ser realizada. Como sugestões, alguns tópicos de interesse na área de mineração de textos podem ser vistos aqui:

- NLProgress: `http://nlpprogress.com/`

Analise um tópico de interesse que você gostaria de pesquisar e defina se a tarefa envolverá classificação, agrupamento ou visualização de textos.

Um exemplo de um problema a ser estudado seria: "Detecção de spam em e-mails", em que os e-mails são, em sua forma original, mensagens de textos de tamanho variável que podem ser ou não spam.

Conjunto(s) de textos

Você pode procurar por conjuntos de textos nos seguintes repositórios:

1. UCI Machile Learning Repository: `https://archive.ics.uci.edu/ml/datasets.html`
2. LABIC: `http://sites.labic.icmc.usp.br/text_collections/`
3. Delve datasets: `http://www.cs.toronto.edu/~delve/data/datasets.html`
4. UCI Knowledge Discovery: `http://kdd.ics.uci.edu/`
5. Data.World: `https://data.world/datasets/open-data`

No caso da temática "detecção de spam em e-mails", escolha um conjunto de textos em algum dos repositórios sugeridos ou construa um conjunto de textos de autoria própria. Como a referida tarefa pode ser vista como um problema de classificação de textos, deve-se ter o cuidado de definir categorias (rótulos) para cada instância de texto ou optar por um conjunto de textos já rotulado.

2 Parte I

Na primeira parte do projeto, deve-se realizar a leitura dos textos a partir de arquivos em disco, o pré-processamento e a caracterização. Além disso, deve-se realizar a revisão de literatura referente à temática escolhida para a pesquisa.

Pré-processamento e caracterização dos textos

Nesse sentido, vocês deverão tomar decisões no que concerne a escolha das técnicas de:

- Pré-processamento dos textos: lemmatização, stemmming, remoção de stop-words etc
- Caracterização dos textos: bag-of-words (BoW), term-frequency inverse document frequency (TF-IDF) ou Word2vec.

Revisão de literatura

Como um artigo científico deverá ser escrito para documentar a metodologia, vocês deverão fazer uma revisão de literatura considerando **a tarefa e o domínio do conhecimento** escolhidos para o projeto. Na temática de "Detecção de spam em e-mails", deve-se pesquisar artigos científicos que estudem/analisem/implementem outras similares. **Não serão aceitas referências de trabalhados provindas de blogs de internet**, *Wikipedia* ou fontes relacionadas. Pesquisem em artigos científicos publicados em anais de congressos, periódicos, dissertações de mestrado e teses de doutorado. Utilize as seguintes ferramentas para buscar artigos Sugestões:

Google Scholar: `http://scholar.google.com.br/`
CAPES Periódicos: `http://www.periodicos.capes.gov.br/`

Artigo - Parte I

Ao final do projeto, espera-se que o artigo científico tenha a seguinte estrutura e formato:

- Entre 4 e 5 páginas
- Língua portuguesa ou inglesa
- Deve conter as Seções: Introdução, revisão de literatura, método proposto, resultados experimentais, conclusão e referências.
- No início do documento, escrever um resumo do trabalho utilizando no **mínimo de 150 palavras** e **máximo 200 palavras**.
- Template disponível (em formato .doc ou LaTeX):
  `http://www.ieee.org/conferences_events/conferences/publishing/templates.html`

Para a parte I do projeto, o artigo deverá conter:

1. Seção de Introdução: contextualização do domínio do problema; descrição do problema a ser pesquisado por mineração de textos; como o problema já foi tratado na literatura; descrição do método proposto; e contribuição.

2. Seção de Revisão de Literatura: quatro trabalhos relacionados na literatura, em que deve-se mencionar (máximo 2 parágrafos por trabalho) o problema resolvido, o método proposto, resultados obtidos e possíveis limitações do método proposto.

Lembre-se que a qualidade da escrita do artigo é **sempre** fundamental. Evite parágrafos compostos por menos de quatro linhas, uso de pronomes pessoais na primeira pessoa ("nós propomos essa abordagem", "decidimos que a escolha da medida...") e mantenha o fluxo de leitura entre os parágrafos.

Instruções para envio

Devem ser enviados o(s) código-fonte(s) e o arquivo .PDF da parte I do artigo, em que os arquivos mencionados devem ser compactados em um único arquivo .zip ou .rar com o seguinte nome:

<numero de matrícula>_< primeiro nome e último nome >.zip

em que <numero do matrícula> deve ser substituído pelo seu número de matrícula e < primeiro nome e último nome > deve ser substituído pelo nome e último sem espaços em branco. Por exemplo, se o número de matrícula é 10/1587778 e o nome do aluno é Palmério Machado Orvalho, o nome do arquivo compactado será 101587778_PalmerioOrvalho.zip.

Para enviar o arquivo compactado, entre no Moodle e procure pelo link (tarefa) "Envio do Projeto - Parte I". Entre neste link, faça o upload do arquivo e confirme o envio do arquivo.

IMPORTANTE: Data limite para envio da Parte I do projeto: **26 de maio de 2019**.

Importante

- O projeto deverá ser realizado **individualmente**.
- Códigos-fontes ou trabalhos copiados da Internet ou qualquer outra fonte receberão nota zero.
- A nota da Parte I do Projeto receberá penalização de $2,5$ pontos por dia de atraso.
- Essa especificação pode sofrer modificações para melhor esclarecer determinados pontos do projeto.