Universidade de Brasília
Departamento de Ciência da Computação

CIC0269 - Processamento de Linguagem Natural - 2022/1

Especificação do Projeto

Data de divulgação: 30/07/2022

1 Objetivo

Propôr um método ou metodologia que empregue técnicas de processamento de linguagem natural e/ou aprendizado de máquina para descoberta de conhecimento em corpos de texto (*corpus*). O projeto é o meio de avaliar todas as habilidades do discente no que tange à compreensão do conteúdo da disciplina, tomada de decisão (escolha do *corpus* e definição da(s) tarefa(s), objetivos ou hipóteses de pesquisa), análise dos resultados, escrita e apresentação.

2 Orientações de preparação do projeto

2.1 Escolha do tema

Escolha um problema ou tarefa em um domínio do conhecimento que demande o uso de técnicas de processamento de linguagem natural. É importante definir as hipóteses de pesquisa ou os objetivos geral e específicos da tarefa escolhida.

Caso tenha dúvidas em qual tipo de tarefa escolher, dê uma olhada nos seguintes links abaixo:

- NLP Progress: `https://github.com/sebastianruder/NLP-progress`
- NLP Tasks: `https://github.com/Kyubyong/nlp_tasks`

ou fale com o professor no horário de atendimento presencial (verifique no Fórum de Avisos da página da disciplina no Aprender3) ou por e-mail.

2.2 Conjunto(s) de dados

Você pode procurar por conjuntos de dados nos seguintes repositórios:

1. Dados Abertos BR: `https://dados.gov.br/`
2. UCI Machile Learning Repository: `https://archive.ics.uci.edu/ml/datasets.html`
3. OpenML: `https://www.openml.org/`
4. LABIC: `http://sites.labic.icmc.usp.br/text_collections/`
5. Delve datasets: `http://www.cs.toronto.edu/~delve/data/datasets.html`

6. UCI Knowledge Discovery: `http://kdd.ics.uci.edu/`
7. Data.World: `https://data.world/datasets/open-data`
8. Lattes: `http://scriptlattes.sourceforge.net/`
9. Dados Abertos SP: `http://www.governoaberto.sp.gov.br/`
10. Repositório com diversos datasets: `http://bagrow.com/dsv/datasets.html`

Por exemplo, a tarefa "Detecção de spam em e-mails" pode ser vista como uma tarefa de classificação. Assim, devemos investigar a aplicação dos modelos de classificação que podem ser testados e empregados para a realização da tarefa. Além disso, é de extrema importância escolher e realizar experimentos entre os modelos de classificação escolhidos com a finalidade de validar o método proposto.

2.3 Revisão de literatura

Como um artigo científico deverá ser escrito para documentar a metodologia, você deverá fazer uma revisão de literatura considerando **a tarefa e o domínio do conhecimento** escolhidos para o projeto. Na tarefa "Detecção de spam em e-mails", deve-se pesquisar artigos científicos que estudem/analisem/implementem outros métodos 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/`

Utilize seu login e senha da UnB no link **Acesso CAFE** para ter acesso livre aos periódicos de diversas de bases de conhecimento, como Scopus, Elsevier, IEEE, Web of Science etc. Recomenda-se fortemente uma **revisão sistemática da literatura** para ter conhecimentos dos trabalhos realizados na tarefa escolhida.

2.4 Pré-processamento dos Textos

Quando aplicável, o método proposto poderá conter etapas de pré-processamento, como:

- Deixar o texto com letras minúsculas, remoção de caracteres/palavras ininteligíveis, *lemmatization*, *stemming*, remoção de *stop-words* etc;
- Remoção de instâncias/atributos redundantes, com valores ausentes etc;

2.5 Extração de características

Nessa fase, você pode utilizar diversas abordagens para obter uma representação estruturada dos textos. Destacam-se os métodos Numericalização (One-hot Encoding), Term Frequency-Inverse Document Frequency (TF-IDF), Word Embeddings (word2vec, Paragraph Vector, Global Vectors), Sentence Embeddings, Transformers (Vetores BERT).

Caso você tenha escolhido trabalhar com textos em arquivos PDF, a extração de características com uso de técnicas baseadas em *Optical Character Recognition (OCR)* pode ser

necessária para extrair os textos dos PDF. Pode-se também utilizar as próprias imagens contidas nos arquivos PDFs (ou a imagem do próprio PDF) para o aprendizado do seu modelo, em que a extração de características pode ser feita implicitamente por meio de redes neurais convolucionais e transferência de aprendizado e vision transformers.

2.6 Processamento de Linguagem Natural e Aprendizado de Máquina

Compreende a implementação de tarefas de processamento de linguagem natural e aprendizado máquina (inclui aprendizado profundo) para a indução dos modelos matemáticos que aprendam os padrões existentes na representação estruturada dos textos. Para isso, a partir do problema e dos dados escolhidos a serem tratados, podem ser consideradas as técnicas vistas em sala de aula, como regressão logística, redes neurais artificiais, máquinas de vetores de suportes, redes neurais convolucionais, auto-encoders, redes neurais recorrentes, transformers etc.

Por exemplo, no contexto de "Detecção de spam em e-mails" pode ser visto como um problema de classificação, por isso, diversos modelos de classificação podem ser testados e empregados para a realização da tarefa. Além disso, é de extrema importância escolher e realizar experimentações comparando-se o desempenho de diferentes modelos de classificação com a finalidade de validar o método proposto.

2.7 Validação e Avaliação do Método Proposto

Se o projeto emprega textos rotulados, defina uma estratégia para avaliar o desempenho dos modelos de classificação escolhidos. Conforme visto em sala de aula, leve em consideração alguma estratégia como Holdout, validação cruzada (*cross validation*), validação cruzada estratificada (*stratified cross validation*) etc. Em seguida, calcule a matriz de confusão resultante do processo de classificação do conjunto de testes. Calcule algumas medidas justas (cuidado com o desbalanceamento de classes no *corpus* escolhido) como Acurácia, Precisão, Revocação, Curva ROC e medida F_1 para avaliar quantitativamente o desempenho da classificação.

Se os textos empregados não possuem rótulos, elabore uma estratégia para analisar e discutir o conteúdo extraído dos textos. Caso você empregue uma técnica de agrupamento (K-Means, extração de tópicos, visualização), deve-se avaliar a qualidade dos agrupamentos formados.

3 Instruções para redação de material

Redigir um artigo científico no formato da IEEE conforme as informações a seguir:

- Entre 4 e 5 páginas, considerando as figuras, referências bibliográficas etc;
- Deve-se utilizar **Língua Inglesa**;
- Deve conter as Seções: Introdução (*Introduction*), revisão de literatura (*Related works*), metodologia ou método proposto (*Proposed method*), resultados experimentais (*Experimental Results*), conclusão (*Conclusion*) e referências (*References*);
- No início do documento, escrever um resumo do trabalho utilizando no **mínimo de 150 palavras** e **máximo 200 palavras**.
- O Template da IEEE está disponível aqui (em formato .doc ou LaTeX): `http://www.ieee.org/conferences_events/conferences/publishing/templates.html`

Detalhadamente, o artigo deverá conter as seguintes seções:

1. Introdução (*Introduction*): contextualização do domínio do problema; descrição do problema a ser pesquisado utilizando aprendizado de máquina; como o problema já foi tratado na literatura; descrição do método proposto; objetivos (geral e específicos) ou hipótese de pesquisa, contribuição e estrutura do artigo (descrever de maneira breve: cada seção e seu propósito);

2. Revisão de Literatura (*Related works*): entre dois e três trabalhos relacionados na literatura, em que deve-se mencionar (máximo 2 parágrafos por trabalho) o problema resolvido, o método proposto e os resultados obtidos;

3. Metodologia/Método proposto (*Proposed method*): apresentar um fluxograma que ilustre todas as etapas da metodologia ou do método proposto relacionado com abordagens de aprendizado de máquina. Descrever detalhadamente cada etapa do fluxograma em uma subseção específica. Não se esqueça de justificar suas decisões quanto à escolha das técnicas de processamento de linguagem natural, aprendizado de máquina, pré-processamento etc;

4. Resultados experimentais (*Experimental Results*): descrever detalhadamente o processo de experimentação para validar e avaliar o método proposto ou a metodologia. Indicar as medidas de avaliação de performance de classificadores ou de qualidade de agrupamentos. Apresente experimentação para escolha de parâmetros, caso alguma técnica empregada demande ajuste de parâmetros. Gráficos, tabelas ou imagens que ilustrem algum experimento devem possuir legendas e as informações nelas contidas (por exemplo, o que pode ser interpretado a partir da tabela ou da imagem) devem ser discutidas no texto;

5. Conclusão (*Conclusion*): descrever de maneira sucinta a essência do método, as principais contribuições, os objetivos cumpridos e não-cumpridos de acordo com os resultados experimentais. Finalizar com possibilidades de trabalhos futuros;

6. Referências: inclua todos as informações completas de artigos e trabalhos científicos referenciados no texto (utilize BibTex para inserir as referências).

Lembre-se que a qualidade da escrita do artigo é **fundamental** para causar uma boa impressão do seu projeto. Evite parágrafos compostos por menos de quatro linhas, uso de pronomes pessoais na primeira pessoa ("I did this approach...", "I read the paper...") e mantenha o fluxo de leitura entre os parágrafos. Não utilize linguagem informal no artigo (haverá penalizações). Por exemplo: ao invés de "This technique was chosen because is good for the research", prefira escrever "This technique was chosen due to its prior successful use in literature in several natural language processing tasks [reference 1], [reference 2]".

4 Entregas

4.1 1ª Entrega

Entregar um arquivo PDF do artigo científico na tarefa específica na página da disciplina na plataforma Aprender3/Moodle. Procure pelo tarefa (link) "Envio do Projeto (1a Entrega)". Entre nesse link, faça o upload do arquivo PDF e finalize a submissão. O artigo deve conter, no mínimo, as Seções *Introduction* (exceto com a descrição do método proposto) e *Related work*.

Deadline: 19 de agosto de 2022, às 23:59h.

O professor fornecerá *feedback* da 1ª Entrega antes do prazo final para envio do projeto completo (*2ª Entrega*).

4.2 2ª Entrega

Todo o material produzido no desenvolvimento do projeto deverá ser publicado em um repositório no Github. O arquivo `README.md` principal do projeto deve conter o seguinte cabeçalho:

Universidade de Brasília
Departamento de Ciência da Computação
CIC0269 - Processamento de Linguagem Natural - 2022/1

Em seguida, deve-se colocar seu nome e do professor:

Professor: Vinícius R. P. Borges

Coloque todas as informações essenciais e indicações de localização dos arquivos gerados pelo projeto no `README.md`. Por exemplo, indique como acessar o código-fonte, o artigo científico e o *corpus*.

Deve-se enviar a URL do repositório do Github em uma tarefa específica na página da disciplina na plataforma Aprender3/Moodle. Procure pelo tarefa (link) "Envio do Projeto (2a Entrega)". Entre neste link, digite a URL completa do repositório do GitHub contendo todos os arquivos relacionados com o projeto e finalize a submissão.

Deadline: 15 de setembro de 2022, às 23:59h.

5 Apresentação

A apresentação ocorrerá em horário de aula, no dia 17/09/2022, sendo necessário apresentar a tarefa de pesquisa, os objetivos ou hipóteses de pesquisa, o método/metodologia proposto, os resultados experimentais e a conclusão. A apresentação tem duração mínima de 5 minutos e máxima de 7 minutos.

Importante

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