Universidade de Brasília
Departamento de Ciência da Computação

ENAP - Processamento de Linguagem Natural

Especificação do Projeto

Data de divulgação: 22/04/2024

1 Objetivo

Propôr um método ou metodologia que empregue técnicas de processamento de linguagem natural para a extração de informações em corpos de texto (*corpus*). O projeto é o meio de avaliar todas as habilidades do participante do curso 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 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 algum dos professores no horário de aula ou de mentoria.

2.2 Corpora

Você pode procurar por conjuntos de dados nos seguintes repositórios:

1. Hugging Faces Datasets: `https://huggingface.co/docs/datasets/index`
2. TensorFlow Datasets: `https://www.tensorflow.org/datasets/catalog/overview`
3. Dados Abertos BR: `https://dados.gov.br/`
4. UCI Machile Learning Repository: `https://archive.ics.uci.edu/ml/datasets.html`
5. OpenML: `https://www.openml.org/`
6. LABIC: `http://sites.labic.icmc.usp.br/text_collections/`

7. Delve datasets: `http://www.cs.toronto.edu/~delve/data/datasets.html`
8. UCI Knowledge Discovery: `http://kdd.ics.uci.edu/`
9. Data.World: `https://data.world/datasets/open-data`
10. Dados Abertos SP: `http://www.governoaberto.sp.gov.br/`
11. 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 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;

2.4 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).

2.5 Técnicas de Processamento de Linguagem Natural

Compreende a implementação de tarefas de processamento de linguagem natural 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 recorrentes, modelos seq2seq, 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.

2.6 Validação e Avaliação do Método Proposto

Se o seu método/metodologia emprega textos rotulados, defina uma estratégia para avaliar o desempenho dos modelos associados à tarefa em consideração a serem obtidos. Nesse sentido, é importante empregar uma estratégia como Holdout, validação cruzada (*cross validation*), validação cruzada estratificada (*stratified cross validation*) etc. Isso significa que as amostras do *corpus* devem, ao menos, ser divididas em treinamento e teste. Os exemplos de treinamento

devem ser considerados para o treinamento do(s) modelo(s) de processamento de linguagem natural visando obter o melhor modelo possível para a tarefa em questão. Os exemplos de teste devem ser utilizados unicamente para avaliação, em que estratégias como matriz de confusão, Precisão, Revocação, Curva ROC e F_1-Score podem ser empregadas para avaliar o desempenho e a eficácia do(s) modelo(s) obtido(s).

Se os textos empregados não possuem rótulos, elabore uma estratégia baseada em similaridade de textos, analisando 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 Apresentação

A apresentação ocorrerá em horário de aula, no dia 29/05/2024, 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** ou em **duplas**;
- Essa especificação pode sofrer modificações para melhor esclarecer determinados pontos do projeto;
- Os critérios de avaliação do projeto serão informados oportunamente.