PCC518 – Tópicos em Computação I - Recuperação de Informação

Professor: Denilson Alves Pereira

Trabalho Prático 2

- Trabalho individual.
- O trabalho deve ser entregue em versão eletrônica pelo Campus Virtual (https://campusvirtual.ufla.br). Envie somente arquivos texto sem formatação e pdf (não enviar .doc, .docx, .odt etc.). Arquivos compactados somente .zip e .tar.gz (não enviar .rar, .z etc.). Não use acentos e nem "ç" nos nomes de arquivo.
- **Data limite de entrega: 02/11/2021 Apresentação: 04/11/2021 (15 minutos)**
- **Valor: 18 pontos**

O objetivo deste trabalho é implementar classificadores de texto usando algoritmos de aprendizagem de máquina.

Siga as seguintes instruções para efetuar o trabalho:

- Faça o download da coleção Reuters-21578, 90 categories, do site http://disi.unitn.it/moschitti/corpora.htm. A coleção está organizada em documentos para treino e teste, armazenados em diretórios de acordo com a classe de cada um;
- Faça o preprocessamento dos dados, de forma a tratar codificações de caracteres, converter todo o texto para letras minúsculas, eliminar pontuações, símbolos desnecessários, tags (no caso de HTML, XML etc.) e stopwords, tokenizar cada palavra do texto etc. Prepare os documentos para serem usados como entrada para os algoritmos;
- Escolha três classificadores e compare sua eficácia para classificação dessa coleção. Use as palavras (*tokens*) como atributos (*features*) e o esquema de pesos TF-IDF como valor para os atributos;
- Ajuste os parâmetros dos classificadores fazendo uma busca em *grid* usando a técnica de validação cruzada em 10 partes (*10-fold cross validation*). Use os dados de treino para isso. Depois de escolhidos os melhores parâmetros, treine os classificadores com toda a coleção de treino e avalie-os usando a coleção de teste;
- Avalie a qualidade dos resultados usando as seguintes métricas: precision, recall e F1 por classe, e Macro-F1 e Micro-F1 (acurária) para o conjunto das classes. Apresente os resultados em forma de tabela, comparando os classificadores. Faça um teste estatístico usando, por exemplo, a medida ANOVA, para verificar se um classificador é realmente melhor do que o outro.

O que deve ser entregue:

- Um relatório técnico contendo introdução, referencial teórico, descrição do trabalho com suas estratégias de solução, experimentos executados e resultados obtidos, conclusão e referências bibliográficas. O relatório deve ter de 6 a 8 páginas, seguindo o template de artigos da SBC;
- Coloque no relatório detalhes como: os classificadores usados, parâmetros configurados, número de documentos de treino e de teste, número de tokens, esquema de pesos etc., de forma que uma outra pessoa que leia o seu relatório consiga reproduzir o mesmo experimento e obter o mesmo resultado;
- O código fonte dos programas, devidamente comentados;
- Slides para apresentação em sala de aula.

Pontos extras:

- 3 pontos: implementar o trabalho usando a biblioteca Apache Spark MLlib (ou simplesmente, Apache Spark ML) https://spark.apache.org/.
- 3 pontos: implementar uma nova abordagem que supere a abordagem básica descrita acima. Sugestões: combine diferentes estratégias de pré-processamento do texto, use *features* obtidas de ferramentas de PLN (*PoS-taggers*, *parser* de dependência, NER, …) combinadas, ou não, com *tokens*/TF-IDF, use *n-grams* como *features*, use modelos de *word embeddings*.