PUCRS - Faculdade de Informática
Inteligência Artificial - Prof. Sílvia

Categorização de Texto

Trabalho 3 sobre Aprendizagem de Máquina

1. **Objetivo do trabalho:** Experimentar técnicas de classificação em um corpus de notícias da Lingua Portuguesa. Faz parte do escopo do trabalho processar os textos, classificá-los e analisar os resultados obtidos.

2. **Corpus:** é um nome dado a uma coleção de documentos (textos). O corpus[1], em anexo, possui 336 textos e foi extraído do Diário Gaúcho on-line durante o ano de 2010. Contém as seguintes seções:

   (a) "Esporte" : 95 textos, sendo muitos sobre Futebol

   (b) "Polícia": 89 textos sobre casos de polícia;

   (c) "Espaço do Trabalhador": 78 textos com foco em oportunidades de emprego e

   (d) "Seu problema é nosso" : 74 textos com relatos de problemas enfrentados pelos leitores como falta de infraestrutura e saúde pública.

3. **Etapas do trabalho -** Abaixo as macro-etapas do trabalho:

   (a) Pré-processamento do corpus: normalização morfológica, anotação linguística, extração dos termos, seleção dos termos mais relevantes, estruturação.

   (b) Categorização e análise dos resultados.

   (c) Escrita de um relatório sobre o trabalho realizado.

4. **Descrição da Etapa de Pré-processamento:** O pré-processamento é a etapa mais custosa de qualquer tarefa em Aprendizagem de Máquina (AM). A preparação dos textos é ainda um pouco mais custosa, pois textos são dados desestruturados. Para estruturá-los, ou seja, colocá-los em um formato que viabilize o processamento dos mesmos por um algoritmo de AM, vai incluir:

   (a) **Normalização Morfológica e Anotação Linguística:** Pode ser feita pelo parser VISL[2] (Figura 1) , Cogroo [3], ou ainda pelo anotador Tree Tagger[4]

[1] Esse corpus faz parte do Projeto PorPopular e foi obtido em http://www.ufrgs.br/textecc/porlexbras/porpopular/download_do_corpus.php

[2] Disponivel para consulta on-line em https://visl.sdu.dk/visl/pt/parsing/automatic/parse.php

[3] Disponivel para download em http://cogroo.sourceforge.net/(Disponivel em http://cogroo.sourceforge.net/

[4] Disponivel para download em http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ ou para consulta on-line em https://gramatica.usc.es/~gamallo/php/tagger/TaggerPT.php

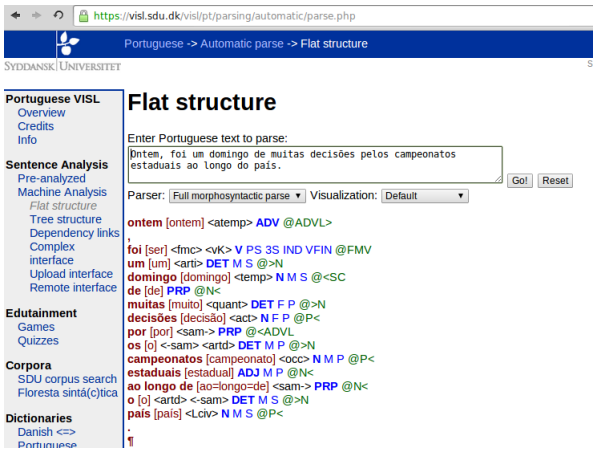

Figura 1- Exemplo de anotação provida pelo parser VISL (versão web do parser PALAVRAS)

(a) O objetivo da **normalização morfológica** é colocar os termos (strings) na mesma "forma". Por exemplo, verbos quando aparecem nos textos estão flexionados: "estudou", "estudaram" e "estuda". A meta, nesse caso, é transformar essas ocorrências em uma forma normal, que no caso dos verbos é o infinito: "estudar". Existem vários tipos de normalização morfológica. A que vamos usar, chama-se "Lematização". A lematização leva os termos para o lema, que, no caso de substantivos, corresponde à palavra no masculino, singular; e, no caso de verbos, no infinitivo. A imagem acima contém a anotação linguistica provida pelo parser VISL. Entre colchetes, imediatamente, após o termo, aparece o lema.

   i. O objetivo da **anotação linguistica** é prover informações sobre o texto para que possamos, em um segundo momento, escolher os termos mais relevantes de um texto. Anotar um texto é colocar tags (rótulos) em seus termos. Essas tags podem ser morfossintáticas e, até mesmo, semânticas. Nesse trabalho, vamos usar apenas a anotação de Part-Of-Speech (POS). As tags[5] de POS indicam as classes gramaticais das palavras: verbo (V), substantivo (N, PROP) adjetivo (ADJ) e advérbio (ADV).

(b) **Extração do Termos:** Após o processo de anotação, já podemos retirar do texto termos que podem ser úteis na etapa de estruturação. Vamos usar n-gramas[6] (n variando de 1 a 3). Para cada texto do corpus, construa uma lista com os n-gramas mais relevantes. Preserve, em sua implementação, a informação sobre o texto do qual esses termos foram extraídos e a seção do jornal desse texto. Quando n=1,

[5] As tags exemplificadas são do VISL. Variam conforme o anotador usado.

[6] n-grama: sequência de n palavras. Por exemplo, na frase "Isso é um teste". Para n=2, existem os seguintes n-gramas: "isso é", "é um", "um teste".

usar apenas as palavras com as tags correspondentes a verbos, substantivos, advérbios e adjetivos. Nos demais casos, incluir além das citadas, as preposições (PRP).

(c) **Seleção dos Termos mais relevantes:** É nessa etapa que precisamos escolher os termos mais relevantes (redução de dimensionalidade) visando a representação dos textos (estruturação). Usando as listas criadas na etapa anterior, crie uma lista geral de termos (sem repetição). Para cada termo dessa lista, contabilize a frequência desse termo no corpus. A seguir, selecione os k primeiros termos mais frequentes. Faz parte do seu trabalho definir o valor de k mais adequado. O resultado dessa seleção é uma lista de termos, conhecida como Bag-of-Words (BoW).

(d) **Estruturação:** Nessa fase, vamos usar uma representação vetorial para estruturar os textos. A BoW funcionará com os atributos (campos) do texto. A representação vetorial mais simples é a binária, que indica se um termo da BoW está ou não no texto. Por exemplo, supondo que a BoW é formada pelo vetor [P1,P2,P3,P4] e existem os seguintes textos já pré-processados (cada texto com sua lista de termos): T1 = {P4, P5, P6} de Esporte; T2 = {P1,P3, P7} de Policia; T3 = {P8,P4, P5} de Esporte; T4 = {P1,P8, P9} de Esporte e T5 = {P1,P4, P9} de Polícia. O arquivo arff para o Weka, (ferramenta que usaremos nas etapas seguintes), com os textos estruturados, ficaria assim:

```
@relation Arquivo
@attribute P1 integer
@attribute P2 integer
@attribute P3 integer
@attribute P4 integer
@attribute classe {Esporte,Policia}
@data
0, 0, 0, 1, Esporte
1, 0, 1, 0, Policia
0, 0, 0, 1, Esporte
1, 0, 0, 0, Esporte
1, 0, 0, 1, Policia
```

5. **Descrição da Etapa de Categorização:** A categorização (ou classificação) de textos é o processo de automaticamente atribuir uma ou mais categorias predefinidas a documentos textuais. Nessa etapa testaremos algoritmos de classificação sobre o corpus pré-processado. O objetivo dessa etapa é testar com diferentes algoritmos (ao menos 3, incluindo k-nn e MultiLayer Perceptron) e comentar aquele que melhor classificou os textos. Usar, nessa etapa, 80% dos textos (de cada classe) para treino e os restantes para teste.

6. **Descrição do Relatório:** Deve ser entregue um relatório descrevendo: o objetivo do trabalho, pré-processamento (descrever o pré-processamento

realizado, configurações da BoW); para cada tarefa, mencionar os algoritmos testados e detalhar a análise dos resultados (tomar como base as medidas usuais), bem como incluir comentários sobre o desenvolvimento do mesmo e a sua conclusão.

7. **Desenvolvimento e Entrega:** O trabalho poderá ser desenvolvido ao longo das aulas práticas da disciplina. Entrega Preliminar: 21/11/2017 via moodle (fontes e versão inicial do relatório). Entrega final: 27/11/2017 também via moodle (fontes e versão final do relatório)..

8. **Forma de avaliação**:

   (a) Etapas de pre-processamento (da normalização à estruturação): 5,0 pts (dependente da avaliação presencial do dia 21/11/2017)

   (b) Etapa de Aprendizagem - tarefa de Categorização (Weka): 2,0 pts (dependente da avaliação presencial do dia 21/11/2017)

   (c) Análise dos resultados (relatório): 3,0 pts

   (d) Pontos Extras:

   - Categorização de novos documentos (fora do corpus dado e fornecidos pelo professor) : 1,0 pt
   - Categorização (treino e teste) usando BoW com unigramas e bigramas (com análise dos resultados): 1,0 pt
   - Categorização (treino e teste) usando BoW com unigramas, bigramas e trigramas (com análise dos resultados):1,0 pt
   - Categorização (treino e teste) usando valores diferentes para k na etapa de seleção de termos (com análise dos resultados) para identificar o valor adequado: 1,0 pt