Pontifícia Universidade Católica
do Rio de Janeiro

Pedro Oguri

Aprendizado de Máquina para o Problema de Sentiment Classification

Dissertação de Mestrado

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós–graduação em Mestrado em Informática do Departamento de Informática da PUC–Rio

Orientador : Prof. Ruy Luiz Milidiú
Co–Orientador: Prof. Raúl Rentería

Rio de Janeiro
outubro de 2006

Pontifícia Universidade Católica
do Rio de Janeiro

Pedro Oguri

Aprendizado de Máquina para o Problema de Sentiment Classification

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós–graduação em Mestrado em Informática do Departamento de Informática do Centro Técnico Científico da PUC–Rio. Aprovada pela Comissão Examinadora abaixo assinada.

Prof. Ruy Luiz Milidiú
Orientador
Departamento de Informática — PUC–Rio

Prof. Raúl Rentería
Co–Orientador
Departamento de Informática — PUC–Rio

Prof. Marcus Vinícius Soledade Poggi de Aragão
Departamento de Informática — PUC–Rio

Prof. Marco Antonio Casanova
Departamento de Informática — PUC–Rio

Prof. José Eugênio Leal
Coordenador Setorial do Centro
Técnico Científico — PUC–Rio

Rio de Janeiro, 25 de outubro de 2006

Pedro Oguri

Graduou–se Bacharel em Informática pela PUC-Rio em 2003 tendo cursado parte de sua graduação na Universidad Autonoma de Madrid. Trabalhou como Engenheiro de Software na Accenture e na Fast Search and Transfer. Tem interesse na pesquisa de algoritmos, especialmente nas áreas de aprendizado de máquina e otimização combinatória.

Ficha Catalográfica

Oguri, Pedro

/ Pedro Oguri; orientador: Ruy Luiz Milidiú; co–orientador: Raúl Rentería. — Rio de Janeiro : PUC–Rio, Departamento de Informática, 2006.

v., 54 f: il. ; 29,7 cm

1. Dissertação (mestrado) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática.

Inclui referências bibliográficas.

1. Informática – Tese. I. Milidiú, Ruy Luiz. II. Renteria, Raúl. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. IV. Título.

CDD: 510

Agradecimentos

À minha esposa Vanessa e ao meu filho Breno, pelo apoio incondicional e compreensão.

Aos meus pais e à minha irmã que tanto me apoiaram e incentivaram.

Ao meu orientador Ruy Milidiú por ter me despertado tamanho fascínio pelo estudo de algoritmos através de seus ensinamentos.

Ao meu co-orientador e amigo Raul Rentería pelo apoio, incentivo e valiosas discussões.

Ao professor Marcus Poggi pelos ensinamentos, incentivo e entusiasmo contagiante.

Aos brasileiros que pagam impostos, por me sustentarem durante o primeiro ano do meu mestrado.

Resumo

Oguri, Pedro; Milidiú, Ruy Luiz; Renteria, Raúl. **Aprendizado de Máquina para o Problema de Sentiment Classification**. Rio de Janeiro, 2006. 54p. Dissertação de Mestrado — Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Sentiment Analysis é um problema de categorização de texto no qual deseja-se identificar opiniões favoráveis e desfavoráveis com relação a um tópico. Um exemplo destes tópicos de interesse são organizações e seus produtos. Neste problema, documentos são classificados pelo sentimento, conotação, atitudes e opiniões ao invés de se restringir aos fatos descritos neste.

O principal desafio em *Sentiment Classification* é identificar como sentimentos são expressados em textos e se tais sentimentos indicam uma opinião positiva (favorável) ou negativa (desfavorável) com relação a um tópico. Devido ao crescente volume de dados disponível na Web, onde todos tendem a ser geradores de conteúdo e expressarem opiniões sobre os mais variados assuntos, técnicas de Aprendizado de Máquina vem se tornando cada vez mais atraentes.

Nesta dissertação investigamos métodos de Aprendizado de Máquina para *Sentiment Analysis*. Apresentamos alguns modelos de representação de documentos como saco de palavras e N-grama. Testamos os classificadores SVM (Máquina de Vetores Suporte) e Naive Bayes com diferentes modelos de representação textual e comparamos seus desempenhos.

Palavras–chave

Aprendizado de Máquina. Sentiment Analysis. Classificação de Texto. Classificadores Bayesianos. Support Vector Machines.

Abstract

Oguri, Pedro; Milidiú, Ruy Luiz; Renteria, Raúl. **Machine Learning for Sentiment Classification**. Rio de Janeiro, 2006. 54p. MsC Thesis — Department of Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Sentiment Analysis is a text categorization problem in which we want to identify favorable and unfavorable opinions towards a given topic. Examples of such topics are organizations and its products. In this problem, documents are classifed according to their sentiment, connotation, attitudes and opinions instead of being limited to the facts described in it.

The main challenge in *Sentiment Classification* is identifying how sentiments are expressed in texts and whether they indicate a positive (favorable) or negative (unfavorable) opinion towards a topic. Due to the growing volume of information available online in an environment where we all tend to be content generators and express opinions on a variety of subjects, Machine Learning techniques have become more and more attractive.

In this dissertation, we investigate Machine Learning methods applied to Sentiment Analysis. We present document representation models such as bag-of-words and N-grams. We compare the performance of the Naive Bayes and the Support Vector Machine classifiers for each proposed model.

Keywords

Machine Learning. Sentiment Analysis. Text Classification. Bayesian Classifiers. Support Vector Machines.

Sumário

Lista de figuras

Lista de tabelas

Science is what we understand well enough to explain to a computer. Art is everything else we do.

Donald Knuth, *Prefácio do livro A=B.*