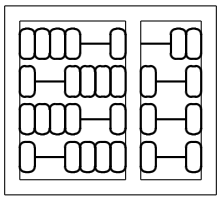

Instituto de Computação
Unicamp

MO 906 - Introdução à Inteligência Artificial

1º Semestre de 2013 - Prof. Siome Goldenstein

Processamento de Texto - Entrega: Sexta, 28/06/2013 às 23:50.

Este trabalho tem como objetivo trabalhar com mensagens de texto (em inglês). O conjunto de dados que utilizaremos possui 18828 mensagens coleteadas de 20 newsgroups (faça um google sobre "usenet" para entender melhor o que é/foi isso).

Para trabalhar com clusterização, assumimos que os dados são não-anotados, no entanto, o nome de cada arquivo nos diz sua procedência, para que seja possível fazer a análise e comparação dos resultados. Esta informação de pertinência é importante também para o aprendizado supervisionado.

O projeto deve ser feito em python com uso do pacote `scikit-learn`.

- Grupos de um ou dois alunos, como definido na P2. A quantidade de trabalho para dois alunos é maior, e tarefas obrigatórias para dois serão própriamente indicadas no enunciado. Tarefas extras dão pontos extras.
- Não é permitido compartilhamento de resultados e funções entre grupos distintos antes da entrega do trabalho.
- Os dados estão disponíveis para download no moodle, ou em `uma instância anterior da disciplina.`.

0. Relatório (4.0 Pontos)

A qualidade do relatório é fundamental (só a apresentação vale 4.0 pontos da nota). Justifique **tudo** o que fizer, e acrescente o código documentado de todas as implementações realizadas. A cobrança em relatórios de grupos de dois alunos é maior que para trabalhos individuais.

1. Preparação dos dados (1.0 Ponto)

1. Releia o artigo `A Plan for Spam,` do Paul Graham.
2. Construa um dicionário de 100 palavras que descreva bem as 18828 mensagens.
3. Extra: Experimente usar o módulo de python para pré-processar as palavras, removendo prefixos, sufixos e tempos verbais.
4. (Para grupos de dois) Crie outros dois dicionários com 200 e 500 palavras - cada dicionário gera novos experimentos em cada etapa adiante.
5. Utilize o dicionário para criar um vetor descritor para cada mensagem (na versão mais simples, este vetor é uma contagem de quantas vezes cada palavra do dicionário aparece na dada mensagem). Após esta etapa, apenas este vetor é utilizado para a análise de cada mensagem.

2. Clusterização (2.5 Pontos)

Utilize o método K-Means de clusterização já implementado no pacote scikit-learn para separar as 18828 mensagens em 20 grupos.

1. Descreva o resultado com uma matriz de pertinência (cluster X grupo).
2. Analise a sensibilidade do resultado do algoritmo para diferentes conjuntos iniciais de sementes.
3. (Para grupos de dois) Experimente a clusterização dos dados com outros algoritmos. Se implementar mostre o código, se utilizar algo de algum pacote indique exatamente o que e de onde utilizou. Compare os diferentes resultados

3. Classificação (2.5 Pontos)

1. Escolha uma técnica apropriada para fazer classificação já implementada no scikit-learn (sugestão KNN ou SVM).
2. Crie conjuntos apropriados de treinamento e teste (80% treinamento, 20% teste).
3. Treine e teste seu classificador - apresente seu resultado como uma matriz de confusão e também com uma média de acerto (média da diagonal principal).
4. (Para grupos de 2) repita o procedimento utilizando validação cruzada (cross-validation) com 5 folds, mostrando médias e desvios padrões.
5. Extra: teste com outro classificador.

4. Bibliografia

Junto com os dados, segue também uma série de artigos relevantes para leitura suplementar. Para os que tiverem interesse, eles servem também como ponto de partida para uma busca bibliográfica mais profunda.