**ACH-2016 Inteligência Artificial - 2o Sem /2014**

**Profa. Patrícia R. Oliveira**

# Especificação do Trabalho Prático

**Entrega no TIDIA:** até a meia-noite do dia 28/11/14. Devem ser postados o código fonte, o código executável (junto com eventuais instruções de execução do mesmo) e o relatório de resultados (extensão .pdf ou .doc). Basta uma postagem por grupo.

**Observação 1:** Este trabalho deve ser realizado por grupos de 3 (três) ou 4 (quatro) alunos.

**Observação 2: Plágios, mesmo que parciais, serão punidos com nota zero para os copiadores e copiados envolvidos. O mesmo acontecerá caso seja detectado plágio com arquivos disponíveis na internet.**

**A. Objetivo geral:** Esse trabalho tem como objetivo geral a implementação do algoritmo de aprendizagem Naive Bayes para realizar uma análise de sentimento de *tweets*, de modo a classificá-los como sendo de "sentimento positivo" ou "sentimento negativo".

**B. Conjunto de dados:** O conjunto de dados disponível em (http://thinknook.com/wp-content/uploads/2012/09/Sentiment-Analysis-Dataset.zip) contém 1.578.627 *tweets* rotulados, sendo o rótulo 1 para sentimento positivo e o rótulo 0 para sentimento negativo.

## Parte 1: Preparação (pré-processamento) dos dados

a) Façam o download do arquivo contendo o conjunto de dados.

b) Removam eventuais dados de cabeçalho e assinatura que existirem nos textos.

c) Construam o vocabulário para os dados de texto.

d) Transformem cada texto de acordo com a freqüência calculada para cada palavra do vocabulário (ver, como exemplo, o exercício sobre Naive Bayes disponível no Tidia e discutido em sala de aula).

## Parte 2: Implementação do Classificador Naive Bayes

Implementem o classificador Naive Bayes para resolver o problema de classificação dos tweets, de acordo com as seguintes etapas:

Etapa I: Calculem todos os termos de probabilidade necessários, utilizando um conjunto de treinamento, formado a partir de um dos métodos de amostragem (a ser em definidos nas Partes 3 e 4 deste documento).

Etapa II: Computem os valores de classe estimados para cada texto do conjunto de teste e calculem a acurácia do classificador.

A forma com a qual os experimentos foram conduzidos, além dos resultados e impressões obtidos nessa fase devem ser descritos no relatório final a ser entregue neste trabalho. Mais detalhes sobre o conteúdo desse relatório serão expostos no final desta especificação.

**Parte 3: Treinamento e Avaliação de Desempenho do Classificador Bayesiano utilizando a Abordagem Holdout.**

O programa para a Parte 3 deve ser implementado para execução via linha de comando, tendo como argumentos os nomes dos arquivos de entrada (conjunto de dados) e de saída (resultados obtidos), e então:

a) Carregar o conjunto de dados pré-processados e o vocabulário.

b) Dividir, aleatoriamente, os dados em conjunto de treinamento e conjunto de teste, em que o conjunto de treinamento deve ter aproximadamente 2/3 do tamanho do conjunto de dados original.

c) Executar a Etapa I para os dados de treinamento.

d) Executar a Etapa II para os dados de teste.

O programa pode ser escrito em Java, C/C++, ou qualquer outra linguagem que gere código executável.

**Parte 4: Treinamento e Avaliação de Desempenho do Classificador Bayesiano utilizando a Abordagem Crossvalidation.**

O programa para a Parte 4 deve ser implementado para execução via linha de comando, tendo como argumentos os nomes dos arquivos de entrada (conjunto de dados) e de saída (resultados obtidos) e, então:

a) Carregar o conjunto de dados pré-processados e o vocabulário.

b) Aleatoriamente, dividir esse conjunto em 10 (dez) partições (folds).

c) Executar o treinamento (e teste) do classificador utilizando o método Crossvalidation.

d) Registrar, no arquivo de saída, os erros verdadeiros obtidos pelo classificador para cada uma das 10 etapas do Crossvalidation e a estimativa final para o desempenho do modelo (média e erro padrão).

O programa pode ser escrito em Java, C/C++, ou qualquer outra linguagem que gere código executável.

**Parte 5: Treinamento e Avaliação de Desempenho do Classificador Bayesiano utilizando a Abordagem Hodout e dados com *stop words* removidas.**

Para essa etapa, devem ser removidas do vocabulário as stop words contidas no arquivo disponível em: https://sites.google.com/site/kevinbouge/stop-words-lists

O programa para a Parte 5 deve ser implementado para execução via linha de comando, tendo como argumentos os nomes dos arquivos de entrada (conjunto de dados com *stop words* removidas) e de saída (resultados obtidos) e, então:

a) Carregar o conjunto de dados pré-processados (sem *stop words*) e o vocabulário.

b) Dividir, aleatoriamente, os dados em conjunto de treinamento e conjunto de teste, em que o conjunto de treinamento deve ter aproximadamente 2/3 do tamanho do conjunto de dados original.

c) Executar a Etapa I para os dados de treinamento.

d) Executar a Etapa II para os dados de teste.

**Parte 6 (Opcional – 2,0 pontos extra): Treinamento e Avaliação de Desempenho do Classificador Bayesiano utilizando a Abordagem Hodout e dados pré-processados por alguma técnica que leve em consideração a aplicação em questão (análise de sentimento).**

O programa para a Parte 5 deve ser implementado para execução via linha de comando, tendo como argumentos os nomes dos arquivos de entrada (conjunto de dados com a técnica sugerida pelo grupo) e de saída (resultados obtidos) e, então:

a) Carregar o conjunto de dados pré-processados (com a técnica sugerida pelo grupo) e o vocabulário.

b) Dividir, aleatoriamente, os dados em conjunto de treinamento e conjunto de teste, em que o conjunto de treinamento deve ter aproximadamente 2/3 do tamanho do conjunto de dados original.

c) Executar a Etapa I para os dados de treinamento.

d) Executar a Etapa II para os dados de teste.

Devem ser entregues ao final deste trabalho:

a) Todos os códigos fontes e executáveis referentes às Partes 3 e 4.

b) Um relatório que inclua:

   i) **Para a Parte 1:** a descrição completa das classes escolhidas (quais foram as classes, motivação para a escolha, número de mensagens em cada uma, etc.). Discuta como os dados foram preparados, ou seja, quais foram os passos dados para a construção do vocabulário.

   ii) **Para a Parte 2:** a descrição e discussão dos experimentos realizados.

   iii) **Para a Parte 3:** uma discussão sobre o desempenho do classificador, levando em conta o erro obtido pelo modelo para o conjunto de teste e a matriz de confusão calculada.

iv) **Para a Parte 4:** uma análise que discuta os erros obtidos pela rede para cada uma das 10 etapas do Crossvalidation e a estimativa final para o desempenho do modelo.

v) **Para a Parte 5:** uma discussão que realize a comparação entre o desempenho obtido pelo algoritmo Naive Bayes para os dados originais e o desempenho obtido para os dados com stop words removidas.

vi) **Para a Parte 6:** a descrição da técnica sugerida pelo grupo e uma discussão que realize a comparação entre o desempenho obtido pelo algoritmo Naive Bayes para os dados originais e o desempenho obtido para os dados tratados pela técnica sugerida pelo grupo.

vii) Suas impressões gerais e conclusões.