UFRGS - Instituto de Informática - Redes Neurais e Sistemas Fuzzy
Gabriel de Souza Seibel - 00262513 - 20/09/2017

Relatório do Projeto
- Impeachment Vote Predictor -

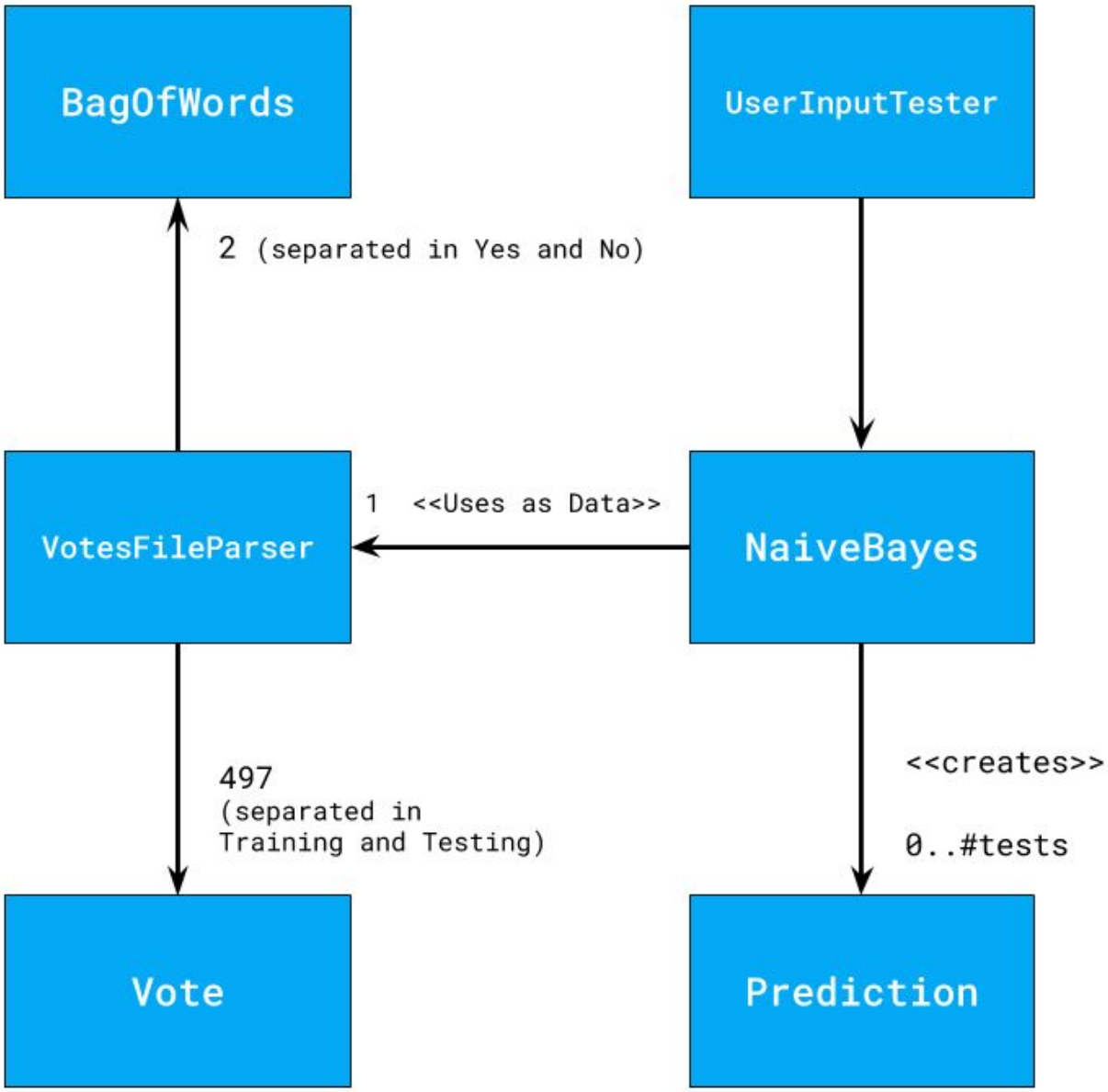
Contextualização

A proposta deste trabalho era de implementar um preditor de votos. A partir de dois arquivos (fornecidos pelos professores) contendo votos reais sobre o impeachment da ex-presidente Dilma Rousseff, um modelo Naive Bayes deveria ser treinado. Durante a preparação dos dados, seriam separados alguns votos para servirem de teste, assim ao final do treinamento poderíamos estimar a acurácia do modelo.

Arquitetura e Metodologia

A segregação de votos para treinamento e teste se deu da seguinte maneira: para cada arquivo, obtinha-se todos os votos nele contidos e cada um passava por um sorteio pseudo-aleatório com 10% de chance - se sorteado, era separado para testes.

Para fins de elucidação da metodologia utilizada, segue um diagrama de classes simplificado da arquitetura do projeto:



Resultados

A acurácia do modelo foi testada em cinco variantes, sendo as primeiras 3 as requeridas no enunciado do projeto, e as últimas duas versões “bônus” para fins de análise. Abaixo, uma legenda e os resultados para cada modelo:

1. **Usando produtórios de probabilidades**
2. **Usando somatórios de logaritmos de probabilidades**
3. **Usando a correção de laplace e somatórios de logaritmos de probabilidades**
4. **Usando correção de laplace e produtórios de probabilidades**
5. **Usando o Boosted Bayes (explicada mais adiante)**

1 - Logarithms = [false] | **Laplace = [false]** | BoostedBayes [false]
Success ratio = 40046 / 55258.0 = **0.7247095443193746**
Success ratio for class yes = 38886 / 40428 = 0.9618581181359454
Success ratio for class no = 1160 / 14830 = 0.0782198246797033

2 - Logarithms = [true] | **Laplace = [false]** | BoostedBayes [false]
Success ratio = 40046 / 55258.0 = **0.7247095443193746**
Success ratio for class yes = 38886 / 40428 = 0.9618581181359454
Success ratio for class no = 1160 / 14830 = 0.0782198246797033

3 - Logarithms = [true] | **Laplace = [true]** | BoostedBayes [false]
Success ratio = 15057 / 55258.0 = **0.2724854319736509**
Success ratio for class yes = 227 / 40428 = 0.005614920352231127
Success ratio for class no = 14830 / 14830 = 1.0

4 - Logarithms = [false] | **Laplace = [true]** | BoostedBayes [false]
Success ratio = 15590 / 55258.0 = **0.282131094140215**
Success ratio for class yes = 1812 / 40428 = 0.0448204214900564
Success ratio for class no = 13778 / 14830 = 0.9290627107215105

5 - Logarithms = [true] | **Laplace = [true]** | **BoostedBayes [true]**
Success ratio = 40565 / 55258.0 = **0.7341018495059539**
Success ratio for class yes = 29307 / 40428 = 0.7249183734045711
Success ratio for class no = 11258 / 14830 = 0.7591368846931895

Foram executadas 1000 baterias de teste. Cada bateria gerava um conjunto aleatório de testes com base nos arquivos (já explicado anteriormente), e media-se a acurácia de cada modelo para o dado conjunto. Ao fim das 1000 baterias, obteve-se a média de acurácia de cada modelo.

Em primeira análise, percebe-se (como visto em aula) que o fato da variante utilizar somatório de logaritmos ou produtório simples não influencia na acurácia - as previsões são, em geral, iguais. A grande mudança de comportamentos na verdade se dá pelo uso ou não da correção de Laplace.

Como mostram os resultados, sem a correção de Laplace os modelos tendem a "chutar" quase sempre "SIM". Isso acontece porque o modelo escolhe a classe y de um voto como "SIM" se P(x|SIM) = P(x|NÃO) - assim, todo voto que contivesse uma palavra que não estivesse no ***dicionário*** (BagOfWords) de nenhuma classe (o que acontece bastante) teria P(x|SIM) = P(x|NÃO) = 0, e a classe escolhida seria "SIM".

Com a correção de Laplace, isso muda. Entretanto, nessas variantes percebe-se uma tendência de chutar "NÃO". Isso acontece pelo fato do ***dicionário*** da classe "NÃO" ser muito menor (palavras 133 votos) do que o da classe "SIM" (palavras de 364 votos). Consequentemente, o denominador do cálculo de P(x|SIM) será sempre maior do que o de P(x|NÃO), fazendo P(x|SIM) ser menor, em geral.

A versão Boosted Bayes é uma tentativa aumentar a acurácia do modelo, contornando os efeitos tendenciosos do desbalanceamento de tamanho dos dados de teste (há muito mais votos "SIM" do que votos "NÃO"). Esse objetivo foi aparentemente atingido, visto que a acurácia tanto para a classe "SIM"quanto para a classe "NÃO" ficam em torno de 0.75.

A maneira com que foi implementada essa variante é bem simples - além do denominador de P(x|y) ser sempre igual, e não haver a soma do logaritmo de P(y), existe um fator de correção que multiplica cada P(x|y) do somatório de logaritmos. Para y = SIM a correção é 1 (sem efeito), mas para y = NÃO a correção é 1.04. Foram testados vários valores, e nota-se que esse valor (para y = NÃO) influencia muito o resultado final: entre 1 (sem efeito) e 1.03 os resultados tendem para "SIM" (já que o numerador de P(x|SIM) é sempre maior); para valores maiores que 1.06, os resultados tendem para "NÃO" (valor da correção muito injusto).