1. Naive Bayes Classifier.

O Bayes classifier é o classificador ideal que escolhe uma classe observando qual das classes, tem é mais provável ocorrer para certos parâmetros, ou seja tem conhecimento à posteriori pois sabe a distribuição probabilística das classes em antemão (MAP). Este classificador baseia-se no teorema de Bayes para definir qual a classe e este é dado pela seguinte equação:

Onde corresponde à distribuição à posteriori das classes. É importante referir que no Bayes Classifier o é apenas um fator de escala pois é comum para todos os e por isso não necessita de ser calculado (a menos que se queiram comparar vários vetores de features). O Naive Bayes Classifier é em tudo igual ao Bayes Classifier mas com uma diferença importante, não são sabidas as distribuições de probabilidade de dado , e por isso é pressuposto que as features são condicionalmente independentes (Covariancia=0) ou seja c

Apartir desta hipótese, seria de esperar que o Naive Bayes Classifier não obtivesse bons resultados no entanto não é isso que ocorre como se pode observar ao longo deste laboratório

4.

Confirmar probabilidades iguais 50 em cada classe de 150 elementos/done/was done before

Como se pode observar o erro que foi obtido é inferior a 1.5% ; Podemos confirmar que a aproximação que foi feita de as distribuições das classes se a próxima uma distribuição de gaussiana; Naive assume independência entre as features com esta suposição no calculo das matrizes da variância e pondo o valor da covariância a 0, resultando a suposição de serem independentes vão ser feitas 7 predições erradas, enquanto que quando foram calculados os valores da covariância os erro de predições foram apenas 2.

Análise preliminar das frases:

“**O curso dura cinco anos.**”:

A linguagem reconhecido nesta frase é a correta embora o valor da margem de classificação seja baixo, sendo que a classe que obteu a classificação mais próxima foi a linguagem espanhola, esta semelhança tem a sua origem em que as duas línguas têm raízes em comum muito perto entre si, por isso para que seja possível distinguir entre si é preciso que existam ocorrências de palavras exclusivas de língua para língua para que se consigam “afastar” as classificações da frase, os valores de classificação também se vão distancia consoante o tamanho do texto a analisar. Neste caso a maioria dos trigramas são comuns nas duas linguagens, sendo difícil distinguir entre o português e o espanhol.

PT

-178.98584

0.66143016

“**El mercado está muy lejos.**” :

Neste caso a linguagem reconhecida pelo classificador corresponde à linguagem real, e o valor da margem de classificador é consideravelmente elevado, ou seja, com este texto o classificador tinha informação para “afastar a concorrência” sendo que o texto a analisar possui característica comuns na linguagem espanhol que são raros nas outras linguagens, fazendo diminuir a pontuação das outras linguagens quando analisaram trigramas característicos da linguagem espanhola:

Como caso de estudo, por exemplo o trigrama “muy”, nos ficheiros utilizados para o treino do classificador aparece 355233 vezes no espanhol, 3103 no francês, 1218 no português e 10 no inglês. Não tomar estes resultados como uma representação total do texto a ser analisado, mas consegue-se encontrar uma preferência evidente com a língua espanhola.

ES

-201.39303

19.288685

“**Eu vou à loja.**”:

Aqui a linguagem reconhecida não corresponde à real sendo que o classificador considerou que estava perante um texto na língua francesa quando na realidade se encontrava perante a portuguesa. No entanto podemos ver que a língua com melhor cotação a seguir à francesa foi a portuguesa, no entanto o valor da margem de classificação não é baixo. Sendo que neste caso estas duas línguas também têm origem no latim as suas raízes também são comuns sendo que existem palavras/(silabas). Parte da possibilidade de o classificador tenha errado pode vir do caso do facto de o texto a ser analisado ser relativamente pequeno(contém apenas 12 trigramas), este conjunto de condições causa que o classificador não irá conseguir encontrar trigramas comuns numa linguagem que seja raro noutra, sendo que após várias ocorrências da situação agora definida o classificador irá isolar o candidato mais adequado ao texto a ser analisado. Como a maioria dos trigramas são comuns na duas linguagens se ocorre algum trigrama muito comum numa linguagem específica poderá fazer “tip the scale” para uma das linguagens. Como é o caso do trigrama “ à ” que ocorre no francês 12 milhões de vezes e no português 500 mil vezes.

FR

-120.26244

3.6242765

“**the word é is very short.**”:

Certo

Se não pusermos os ponto final o classificador considera o texto como, sendo que com a adição do ponto final o classificar vai passar a analisar o trigrama “rt.”, que vai tomar valores muito baixos no “dicionário portugues”(A DATA NÃO CORRESPONDE COM A SUPOSIÇÃO) português dizer que se este for o caso, não é comum palavras em português que acabem em t e,.

O trigrama “ é ” é muito comum no português 1966239 e não existe no inglês. Sendo que o score do português deste trigrama “é i” = 55737 e em inglês = 57. Sendo que estes trigramas vão causar que a classificação geral do texto não irá conseguir isolar a língua inglesa como a língua correspondente do texto introduzido.

Sendo que neste texto a margem do classificador é muito baixo, indicande que é preciso um sample size superior para que se apresentem resultados mais fiáveis.

EN

-213.55727

0.34573567

“**I go to the beach … I go to the beach.**“ a frase está repetida 10 vezes

Neste caso como o texto que temos para analisar é consideravelmente superior logo vai apresentar uma maior margem de classificação , no entanto não nos encontramos perante um bom exemplo pois como o conteúdo vai estar repetido, se a frase original, neste caso “I go to the beach. I ” Possuir uma margem de erro pequena pode querer indicar que está mal classificada logo, quando repetida poderá levar a um margem superior, podendo estar errado. No entanto a frase original neste caso possui evidentemente um terno que é característico da língua inglesa com um valor elevado de presença e evidente em comparação com as outras línguas às quais estamos a comparar

Perante a situação que nos encontramos a repetição de uma frase não vai causar um aumento exponencial do valor do score pois este não esta a “tomar atenção à ocorrência de vários trigramas na mesma frase, congunto de trigramas não tem infuencia nenhuma para o score total, por exemplo co existência de I e the na mesma frase (pedaços da frases reconhecidos como comuns na língua inglesa) não vão gerar um score suplementar se aparecem mais vezes sendo apenas contabilizados o valor da apparence no dicionário usado para testes.”

???????? o “the” e “ i “ é mais comum nas outras línguas que no ingles???????

Como é o caso que é apresentado encima em que a linguagem real era portuguesa mas foi identificado como francês, quando repetido 10X ainda considera fr no entanto se repetirmos por volta de 20-30 X o classificador começa a assumir que se encontra perante pt

EN

-1447.8983

259.87589

I’m bored gonna stop now

SRY WILL COMEBACK LATER