MBA USP ESALQ

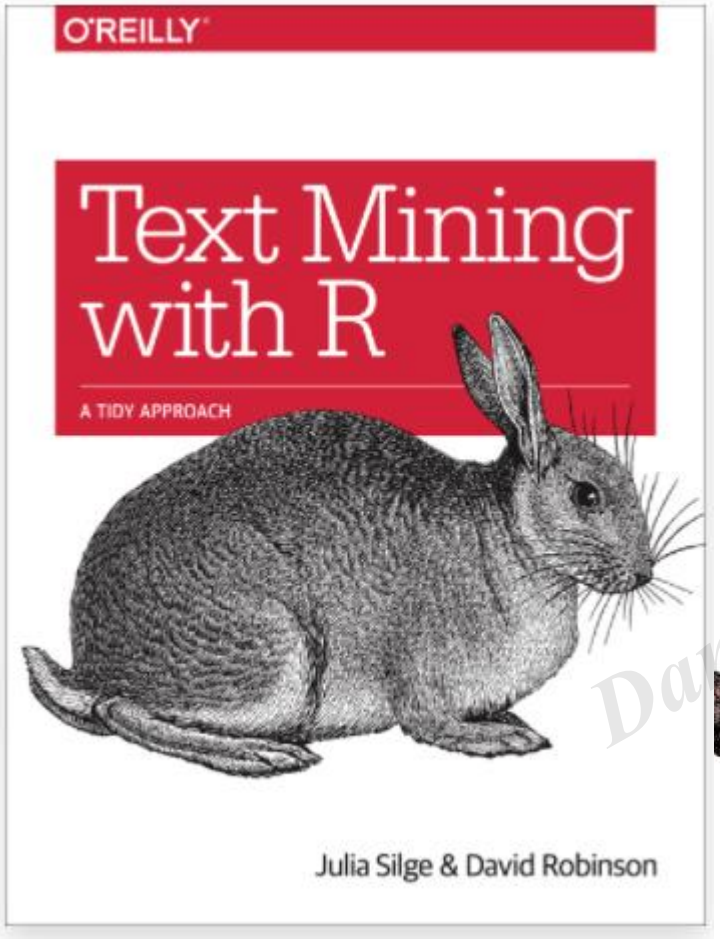
TEXT MINING, SENTIMENT ANALYSIS E NLP

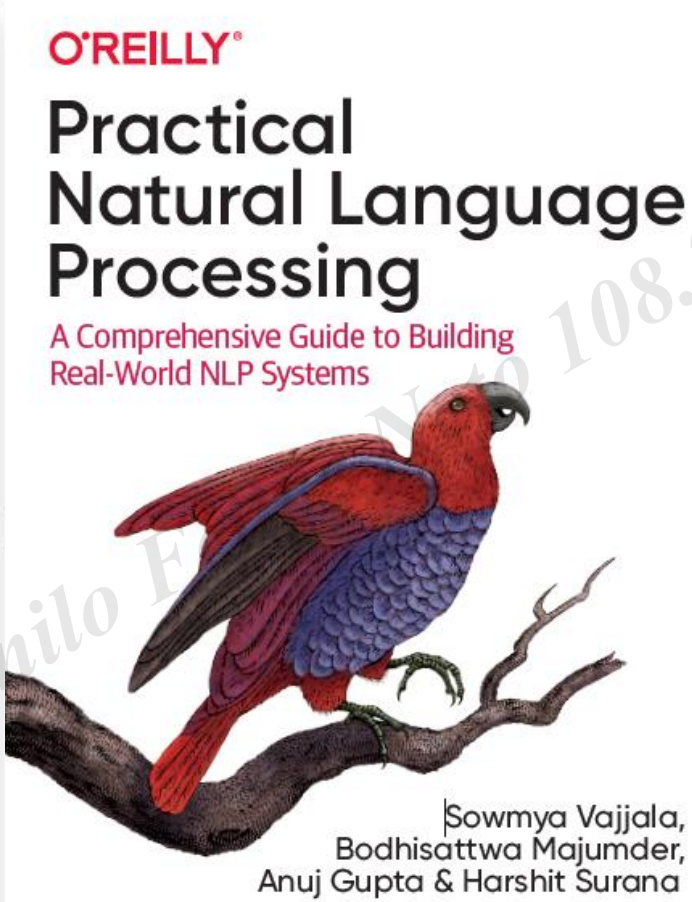
Prof. Dr. Jeronymo Marcondes

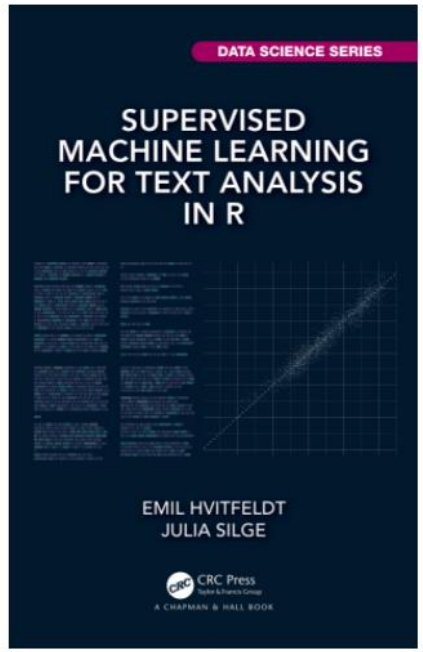
Plano de ataque

- TF-IDF
- Análise de sentimentos palavra a palavra
- Análise de sentimentos com algoritmo supervisionado

Plano de ataque







TF-IDF

- 3 formas de se representar um conjunto de textos (nessa aula):

1. Bag of words
2. Bag of n-grams
3. TF-IDF

TF-IDF

- Qual a importância de uma palavra em um texto?
- A depender da escolha anterior – resposta diferente
- Alguns exemplos para bag of words: “para”, “com”, “nome”, etc

TF-IDF

- Stop words ou não, algumas palavras são mais comuns – nem sempre a melhor forma
- Bag of words escolhe a palavra mais comum no word count
- Isso faz com que percamos informação relevante

TF-IDF

- Não podemos levar em conta apenas a frequência (tf), mas também o comportamento das palavras ao longo de um conjunto de documentos: *"corpus"*

- Outra abordagem é observar a frequência de documento inversa (idf) de um termo, o que diminui o peso de palavras comumente usadas e aumenta o peso de palavras que não são muito usadas em uma coleção de documentos.

TF-IDF

- Lei de ZIPF:

A lei de Zipf afirma que no conjunto de dados de uma linguagem, a frequência de uma palavra é inversamente proporcional a sua posição na lista global de palavras depois de classificadas por sua frequência de forma descendente.

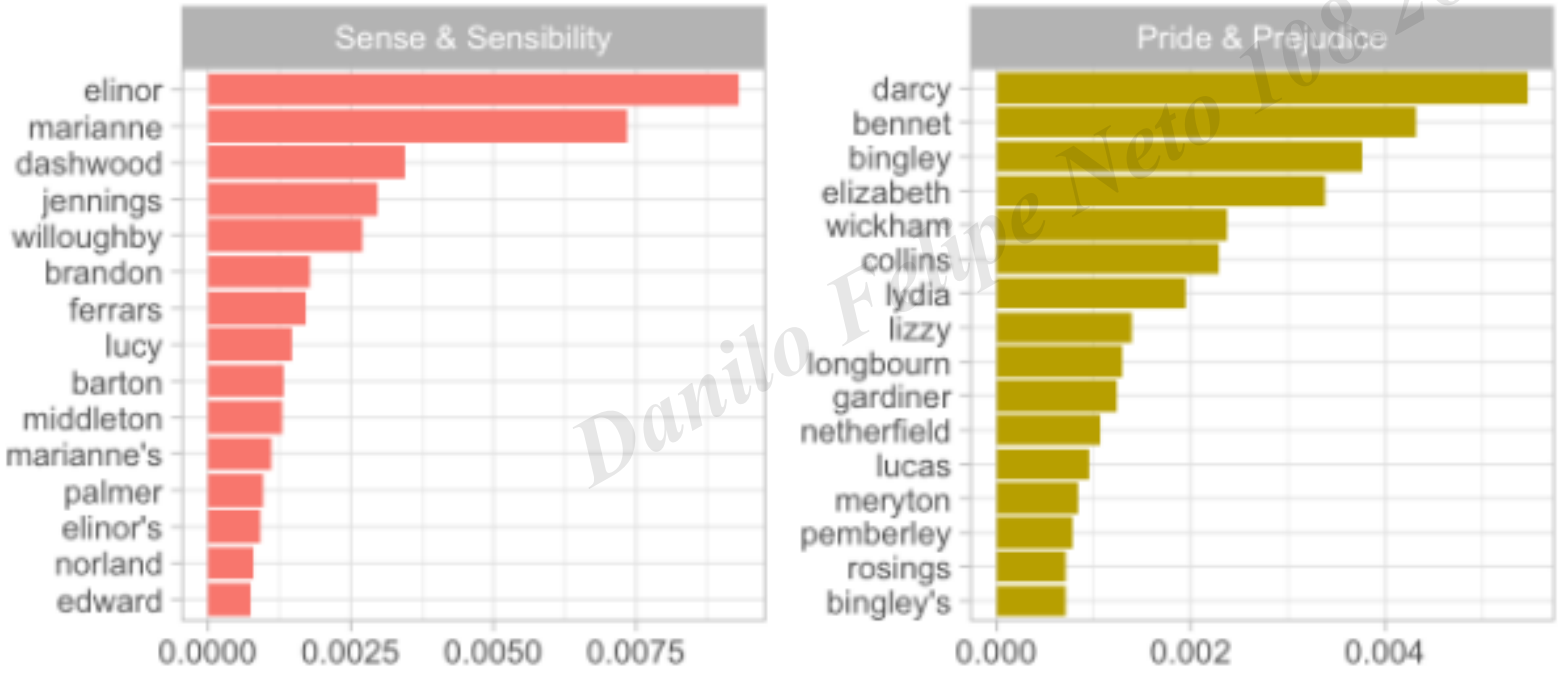
Fonte: https://www.wolfram.com/

TF-IDF

- TF-IDF visa verificar o quanto uma palavra é importante em um documento

- Intuitivamente, a palavra tem que aparecer muito em um determinado documento, mas sua frequência nos demais documentos não pode ser tão grande

TF-IDF



Fonte: Text Mining with R: a tidy approach

Classificação de texto

- Um dos objetivos mais comuns de NLP
- Colocar um texto em uma categoria.
- O desafio da classificação de textos é “aprender” essa categorização a partir de uma coleção de exemplos para cada uma dessas categorias e prever as categorias para novos.

Classificação de texto

- A classificação de texto é uma técnica de aprendizado de máquina que atribui um conjunto de categorias predefinidas ao texto aberto.
- Exemplos:

1. Detecção de falas abusivas
2. Spam Filter
3. Label em tópicos

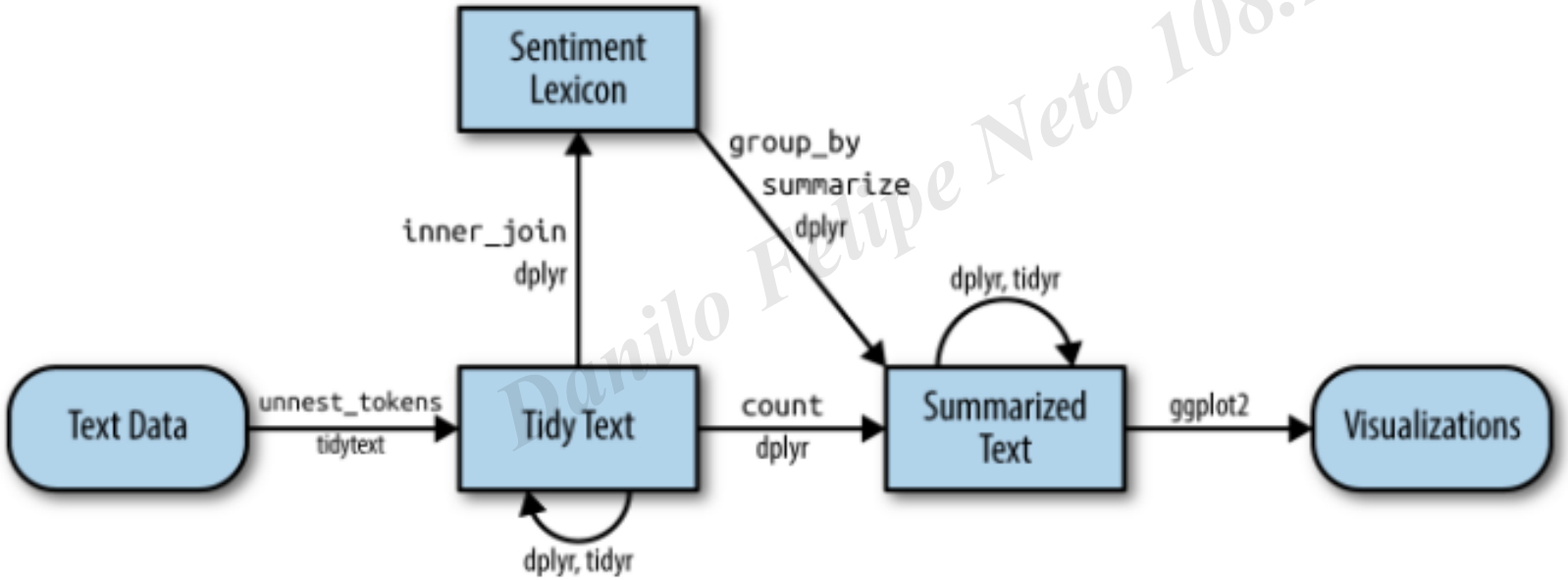
Análise de Sentimentos

- Uma das principais formas de categorização: Análise de Sentimentos
- Qual o sentimento envolvido em um texto?
- Exemplo: Críticas de um produto em um site.

Análise de Sentimentos

- Abordagens de análise de sentimentos
- Análise de Sentimentos baseada em palavras
- Abordagem baseada em Machine Learning.

Abordagem Heurística



Datasets de Sentimentos

- AFINN, bing, nrc.
- Baseados em definição de sentimentos por palavras = unigramas.
- Contém as palavras e os respectivos “scores” de cada uma.

Datasets de Sentimentos

- Métodos baseados em dicionário, como os que estamos discutindo, encontram o sentimento total de um pedaço de texto somando as pontuações de sentimento individuais para cada palavra no texto.

- Sentimento de um texto = valor líquido da soma dos sentimentos de cada palavra.

Procedimento

1. Unnest tokens
2. Datasets de Sentimentos
3. Inner Join

Procedimento

```
#> # A tibble: 303 × 2
#>    word        n
#>    <chr>   <int>
#>  1 good      359
#>  2 young     192
#>  3 friend    166
#>  4 hope      143
#>  5 happy     125
#>  6 love      117
#>  7 deal       92
#>  8 found      92
#>  9 present    89
#> 10 kind       82
#> # … with 293 more rows
```

```
library(tidytext)


get_sentiments("afinn")
```

```
#> # A tibble: 2,477 × 2
#>    word       value
#>    <chr>      <dbl>
#>  1 abandon       -2
#>  2 abandoned     -2
#>  3 abandons      -2
#>  4 abducted      -2
#>  5 abduction     -2
#>  6 abductions    -2
#>  7 abhor         -3
#>  8 abhorred      -3
#>  9 abhorrent     -3
#> 10 abhors        -3
#> # … with 2,467 more rows
```

Limitações

- Falta de contexto
- Ordem não importa
- Dificuldade de generalização – não há “aprendizado”

Pipeline NLP

- Construir modelo de ML
- Diferentes modelos
- Iremos abordar: Naive Bayes e Support Vector Machine

Métodos ML para NLP

- Qual o objetivo?
- O que procuramos fazer?
- Usos

Naive Bayes

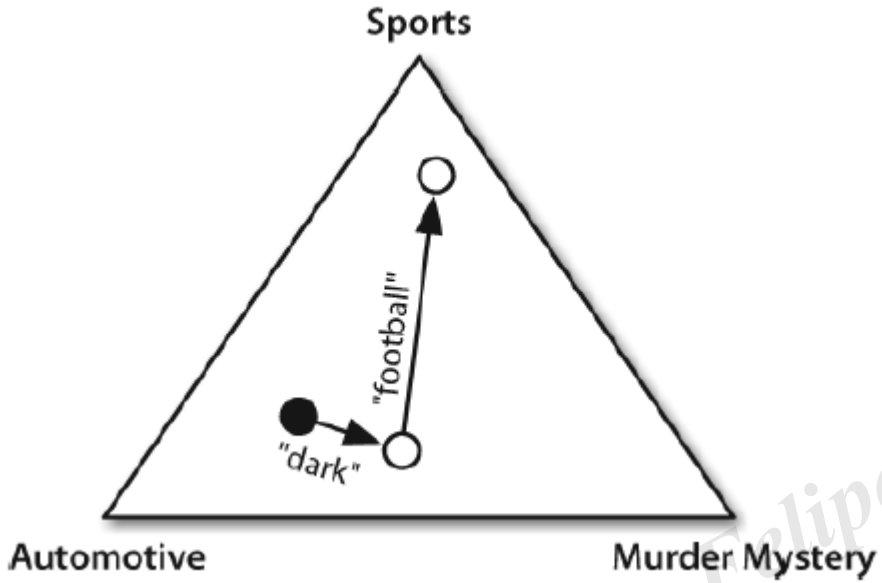
- Baseado no teorema de Bayes
- Suponha dois eventos A e B.
- $P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$

Naive Bayes

- Naive Bayes é um modelo probabilístico baseado no Teorema de Bayes que pode ser utilizado para classificar texto com base nos dados de treinamento.

- Ele estima a probabilidade condicional de uma determinada label ser gerada por uma feature: calcula a probabilidade de ocorrência de cada label sozinha e depois avalia como cada feature pode contribuir para determinados valores.

Naive Bayes

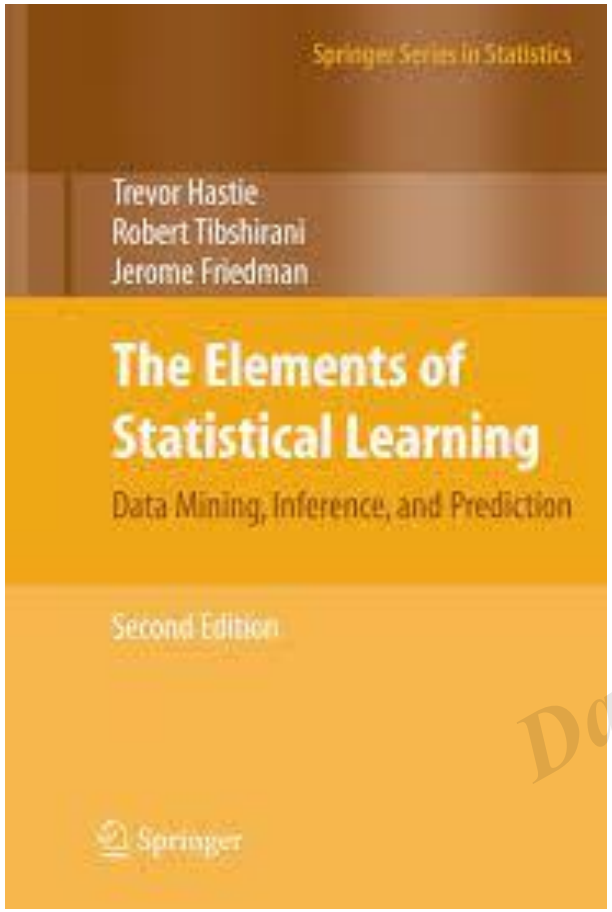


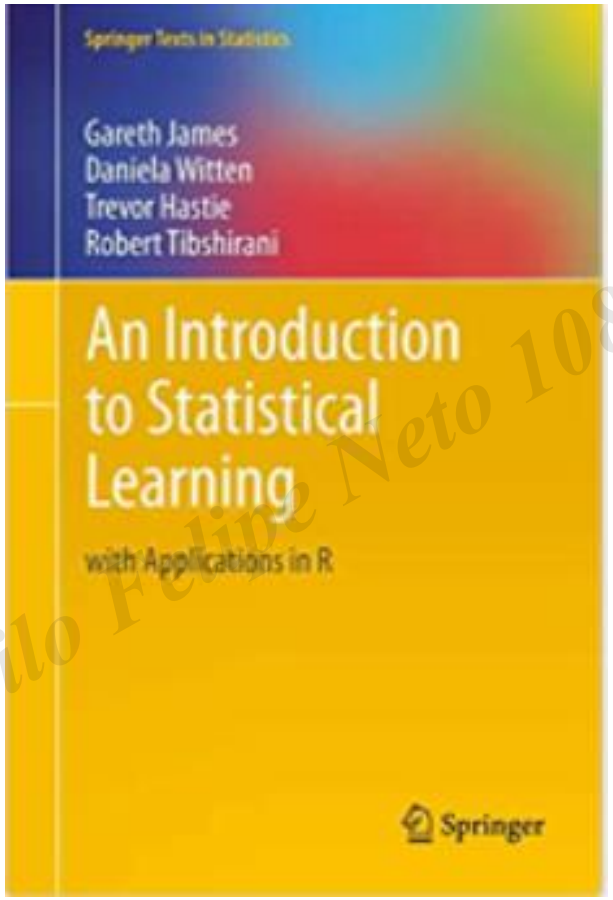
Fonte: Natural Language Processing with Python

O mais comum são labels de automobilismo, portanto começa ali.

Aparecem as palavras “dark” (indicador fraco de mistério) e “football” (indicador forte de esportes).

Naive Bayes



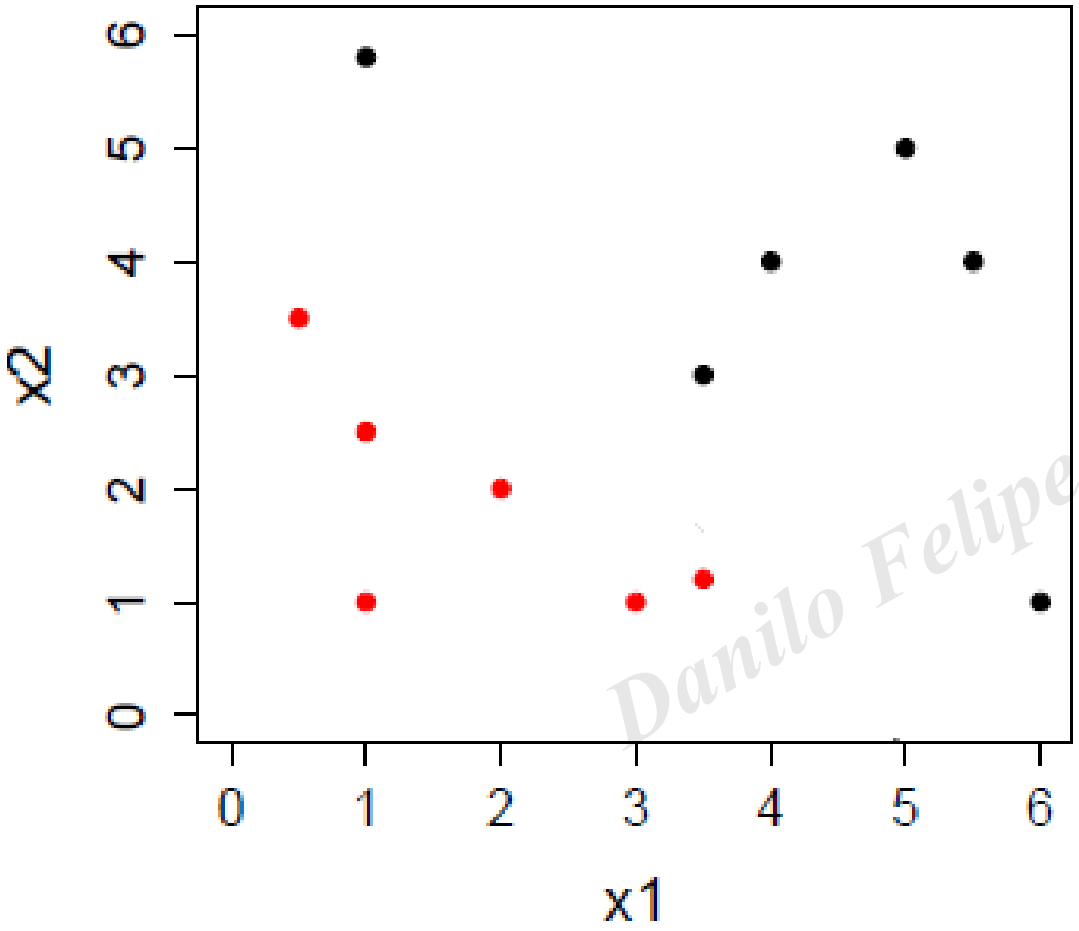


Support Vector Machine

- Busca encontrar o melhor hiperplano separado entre duas classes

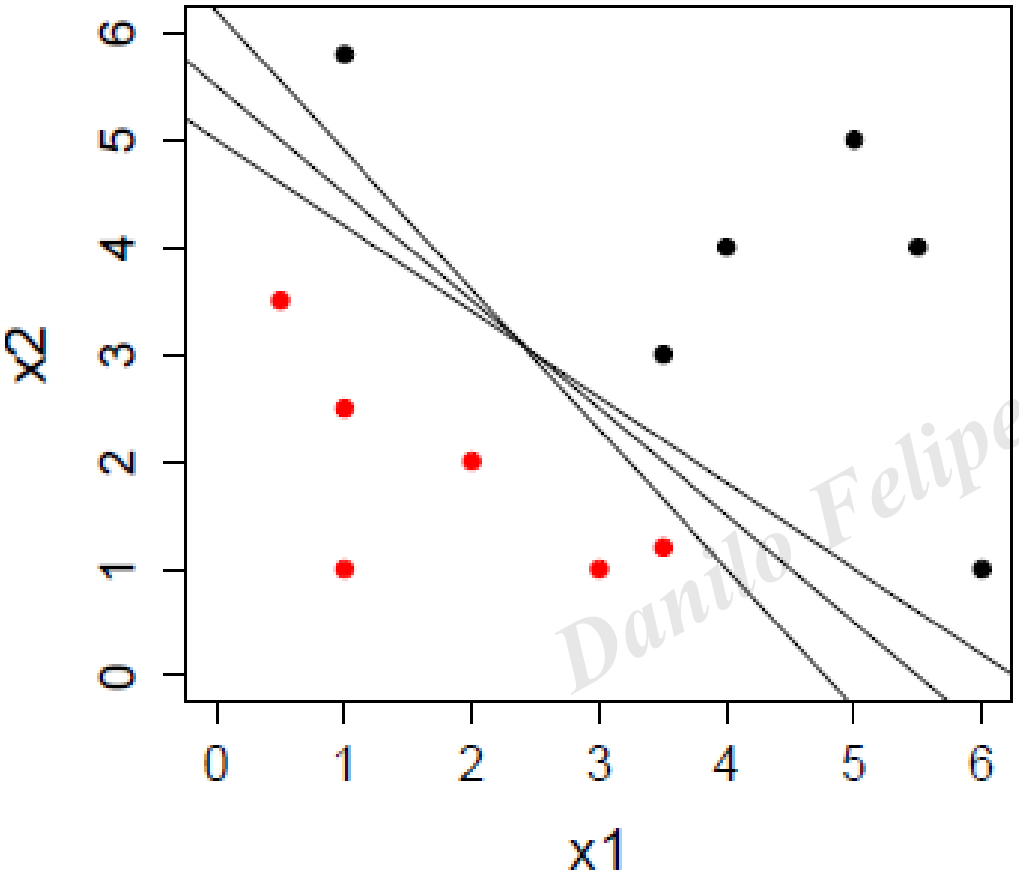
- 3 possibilidades: classificador de margem máxima, classificador de margem flexível e classificador de margem não linear.

Support Vector Machine



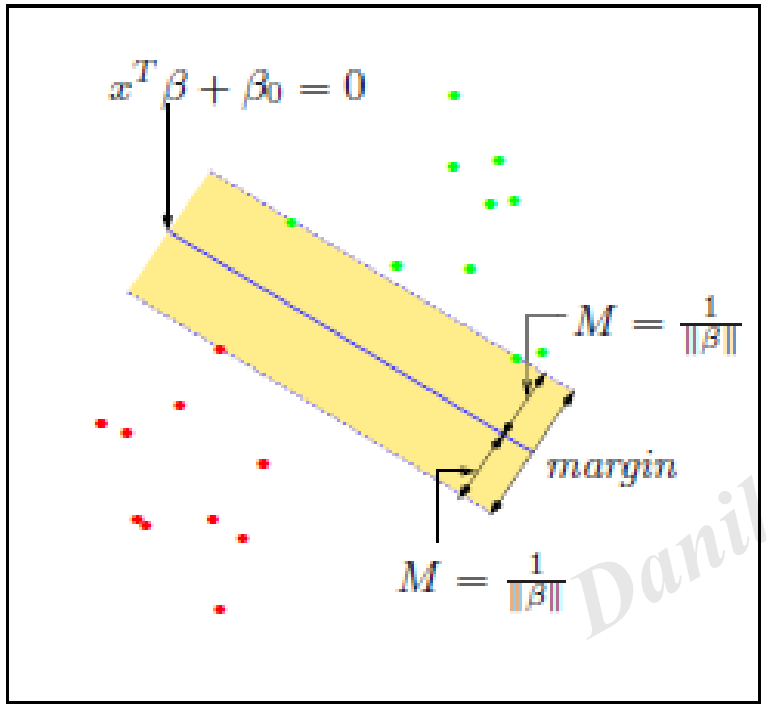
Fonte: Ciência de Dados - Morettin

Support Vector Machine

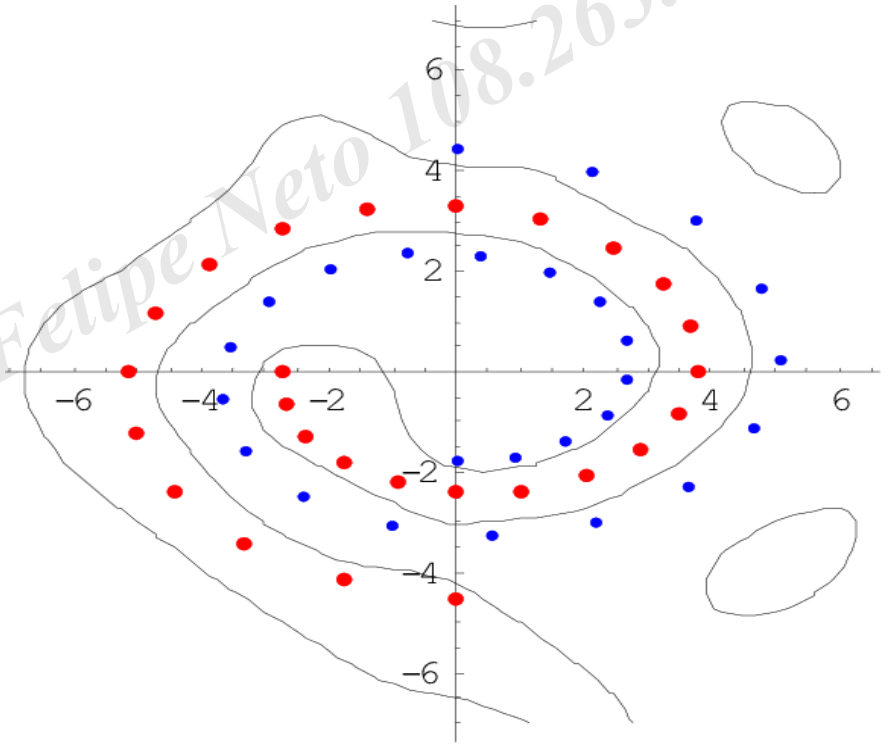


Fonte: Ciência de Dados - Morettin

Support Vector Machine

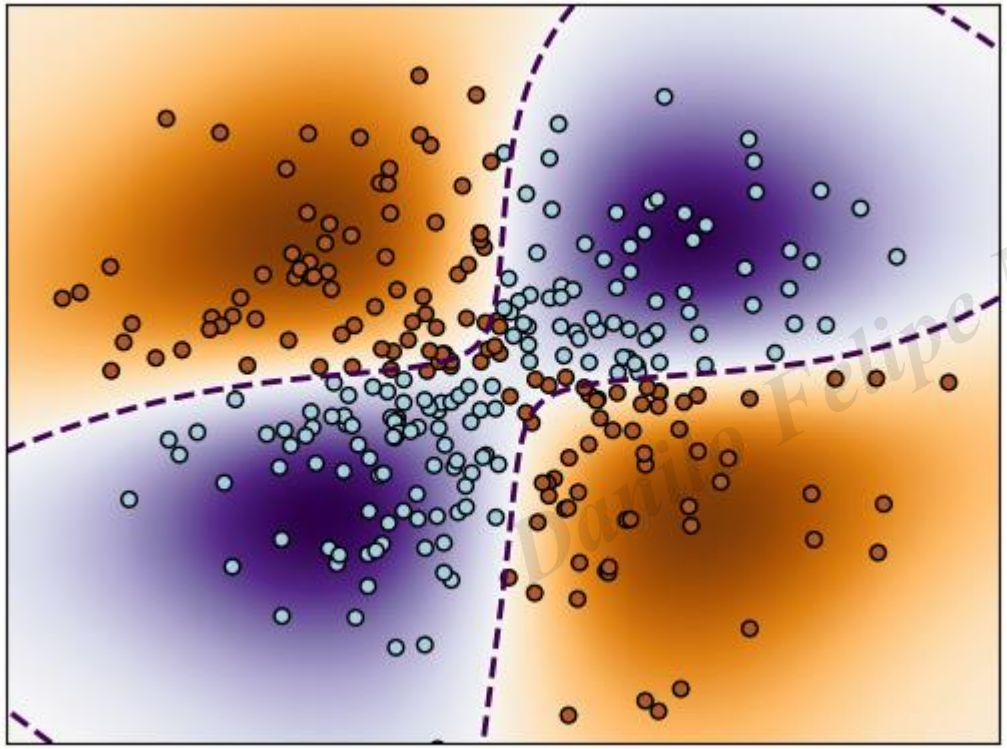


Fonte: Elements of Statistical Learning



Fonte: https://researchgate.com/

Margem Flexível



Fonte: https://scikit-learn.org/

Support Vector Machine

- Objetivo: separar as classes = classificar os textos

- Função objetivo:

1. Maximizar a margem.
2. Sujeito ao fato de que cada ponto deve ser maior que a margem.
3. E sujeito a um possível termo de erro nos modelos de margem flexível.

Performance

- Nem sempre a melhor solução de primeira.

Reason 1	Since we extracted all possible features, we ended up in a large, sparse feature vector, where most features are too rare and end up being noise. A sparse feature set also makes training hard.
Reason 2	There are very few examples of relevant articles (~20%) compared to the non-relevant articles (~80%) in the dataset. This class imbalance makes the learning process skewed toward the non-relevant articles category, as there are very few examples of "relevant" articles.
Reason 3	Perhaps we need a better learning algorithm.
Reason 4	Perhaps we need a better pre-processing and feature extraction mechanism.
Reason 5	Perhaps we should look to tuning the classifier's parameters and hyperparameters.

Discussão

Futuro de NLP e tendências.