MBA
USP
ESALQ

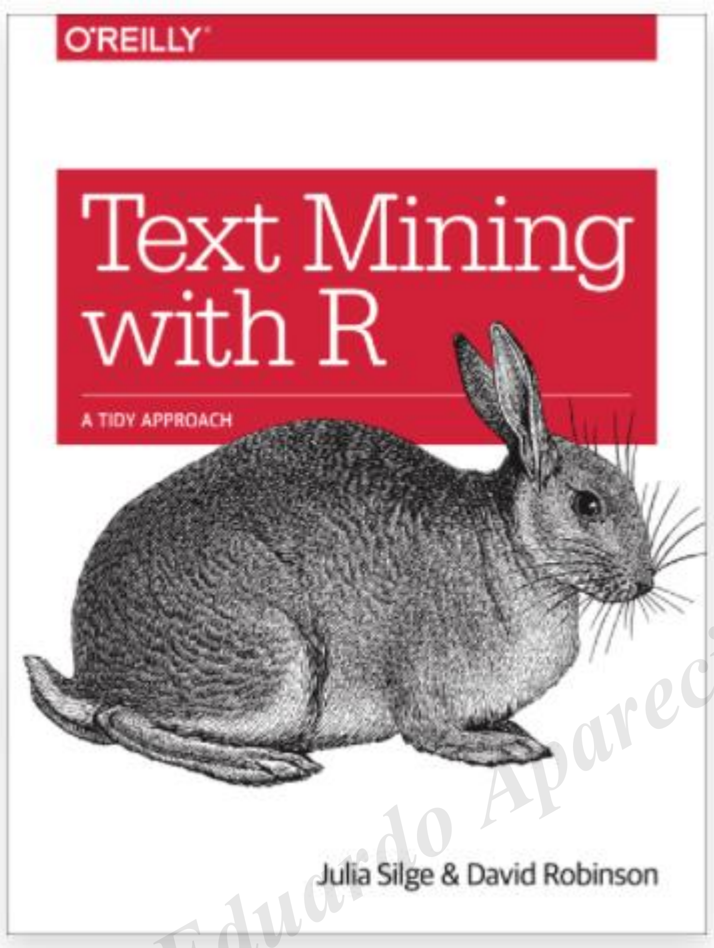
TEXT MINING, SENTIMENT ANALYSIS AND NLP

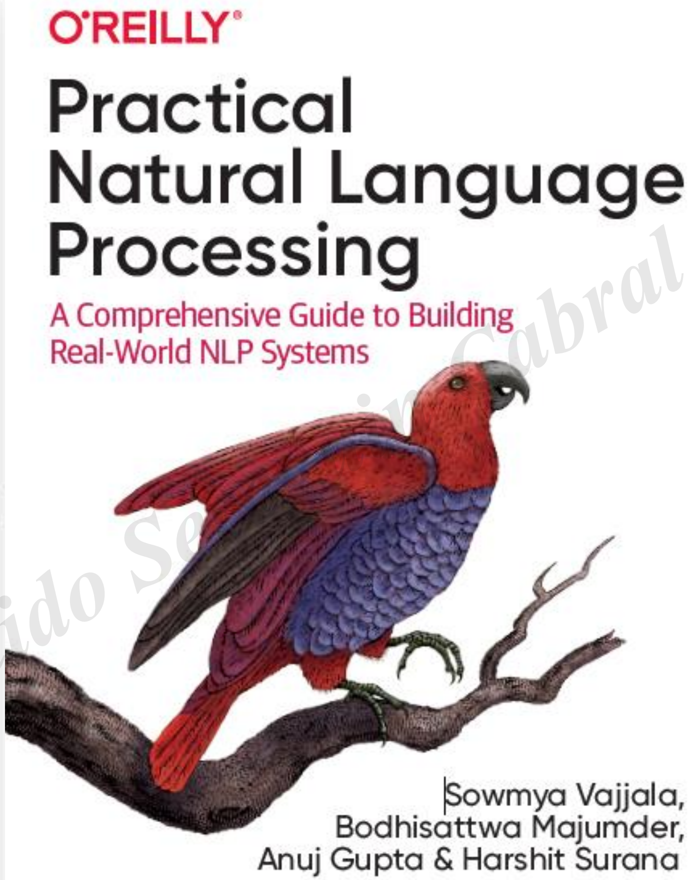
Prof. Dr. Jeronymo Marcondes

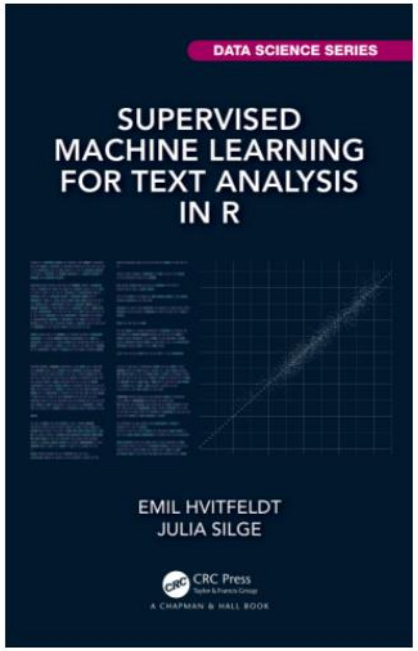
Plan of attack

- TF-IDF
- Sentiment analysis word by word
- Sentiment analysis with supervised algorithm

Plan of attack







TF-IDF

- 3 ways of representing a set of texts (in this class) :

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

TF-IDF

- What is the importance of a word in a text?
- Depending on the previous choice – different answer
- Some examples for bag of words: “for", "with", “name” etc.

TF-IDF

- Stop words or not, some words are more common – not always the best form
- Bag of words choose the most common word in word count
- This makes relevant information be lost

TF-IDF

- We can not consider only frequency (tf) but also the behavior of words throughout a set of documents: "*corpus*"

- Another approach is to observe the frequency-inverse document (idf) of a term, which decreases the weight of words commonly used and increases the weight of words that are not very used in a collection of documents.

TF-IDF

- Zipf's law:

Zipf's law states that in the data set of a language, the frequency of a word is inversely proportional to its position in the global list of words after classified by its frequency in a descent form.

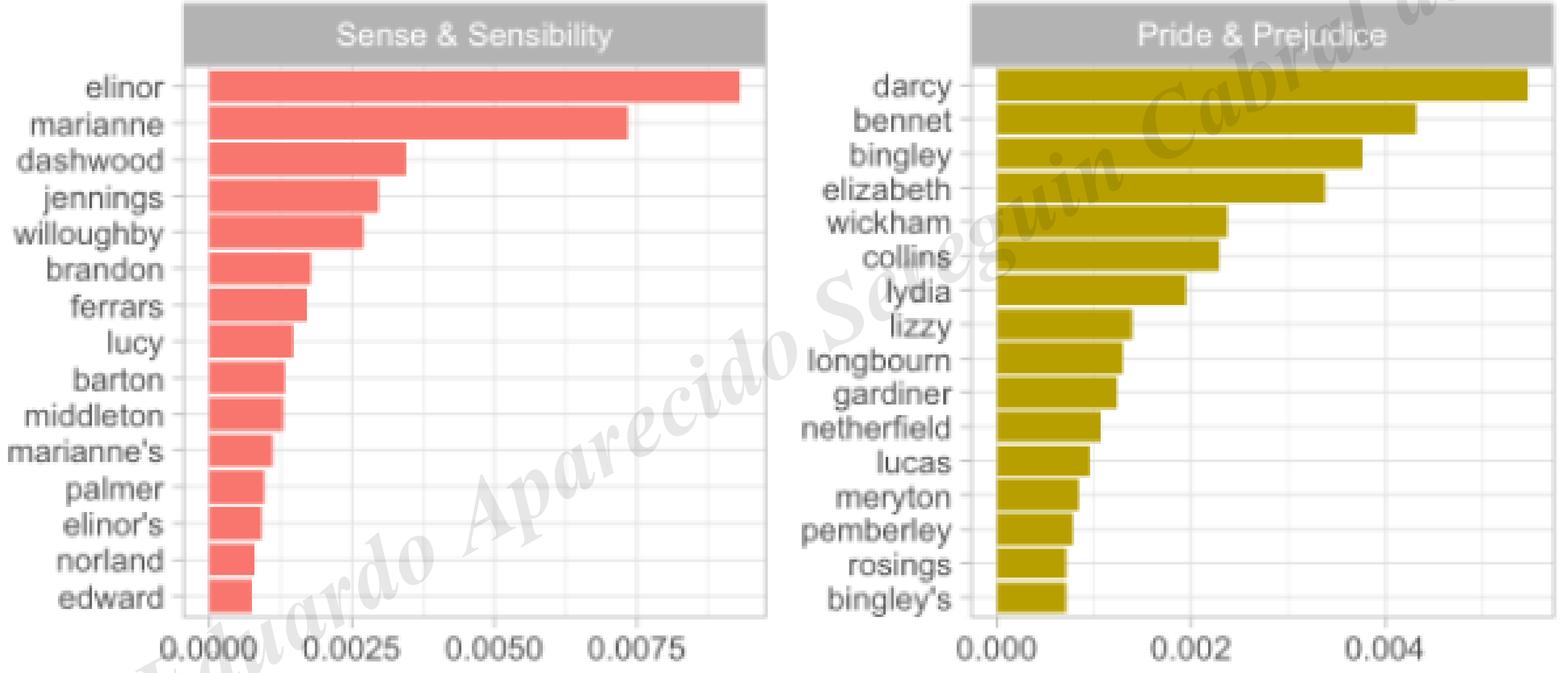
Source: https://www.wolfram.com/

TF-IDF

- TF-IDF aims to verify how important a word is in a document

- Intuitively, the word has to appear a lot in a certain document, but its frequency in other documents cannot be that great

TF-IDF



Source: Text Mining with R: a tidy approach

Text classification

- One of the most common goals of NLP
- To insert a text in a category.
- The challenge of text classification is "learn" this categorization from a collection of examples for each of these categories and predict the categories for new examples.

Text classification

- The classification of text is a machine learning technique that assigns a set of predefined categories to the open text.
- Examples:

1. Detection of abusive speech
2. Spam filter
3. Label in topics

Sentiment Analysis

- One of the main forms of categorization: Sentiment Analysis

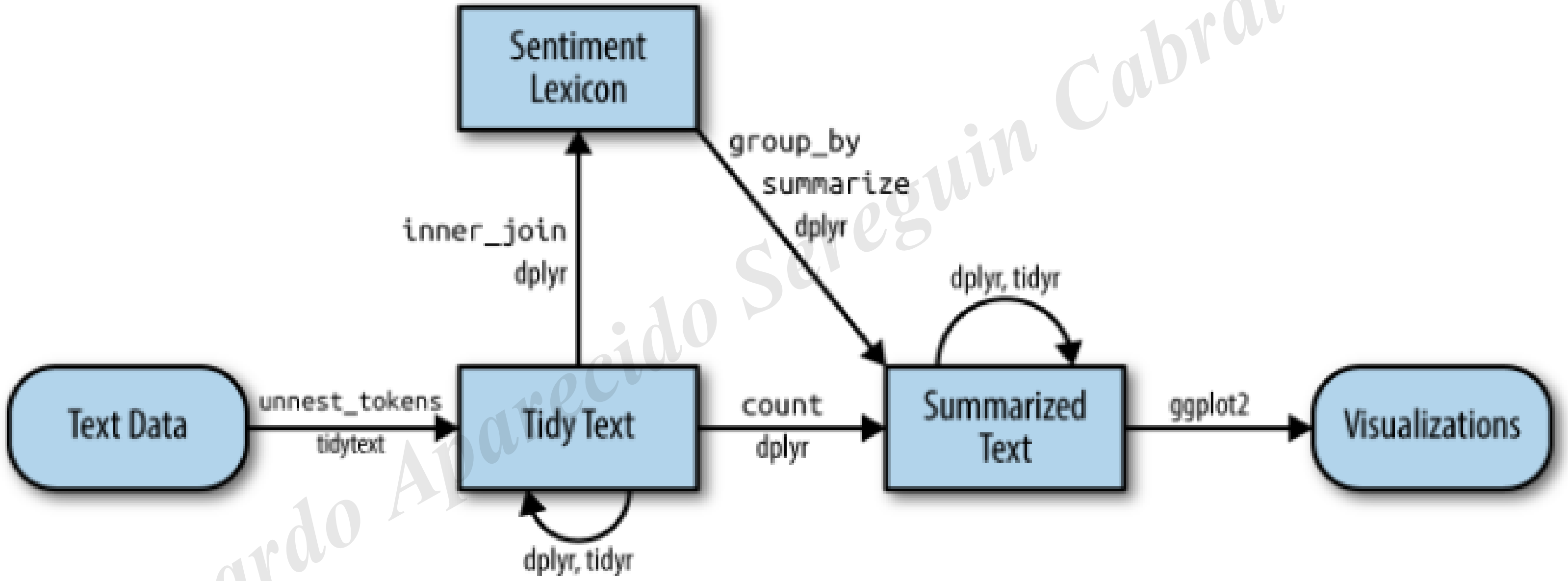
- What is the sentiment involved in a text?

- Example: Critices of a product in a website.

Sentiment Analysis

- Sentiment analysis approaches
- Sentiment Analysis based on words
- Approach based on Machine Learning.

Heuristic approach



Sentiment Datasets

- AFINN, bing, nrc.
- Based on definition of sentiments by words = unigram.
- It contains the words and the respective "scores" of each one.

Sentiment Datasets

- Methods based on a dictionary, such as those that we are discussing, they find the total sentiment of a text part by adding the scores of individual sentiment for each word in the text.

- Sentiment of a text = net value of the sentiments sum of each word.

Procedure

1. Unnest tokens
2. Sentiment Datasets
3. Inner Join

Procedure

```
#> # 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
```

Limitations

- Lack of context
- The order does not matter
- Difficulty of generalization – there is no "learning"

NLP Pipeline

- Build ML model
- Different models
- We will approach: Naive Bayes and Support Vector Machine

ML Methods for NLP

- What is the objective?
- What do we try to do?
- Uses

Naive Bayes

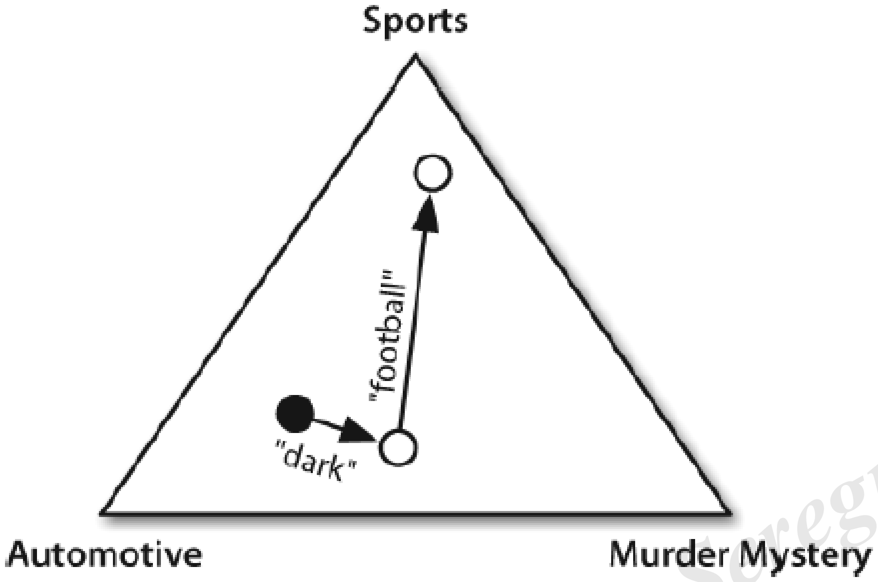
- Based on the Bayes' theorem.
- Suppose two events A and B.
- $P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$

Naive Bayes

- Naive Bayes is a probabilistic model based on the Bayes' Theorem that can be used to classify text based on training data.

- It estimates the conditional probability of a certain label to be generated by a feature: it calculates the probability of occurrence of each alone label, and, then, it evaluates how each feature can contribute to certain values.

Naive Bayes

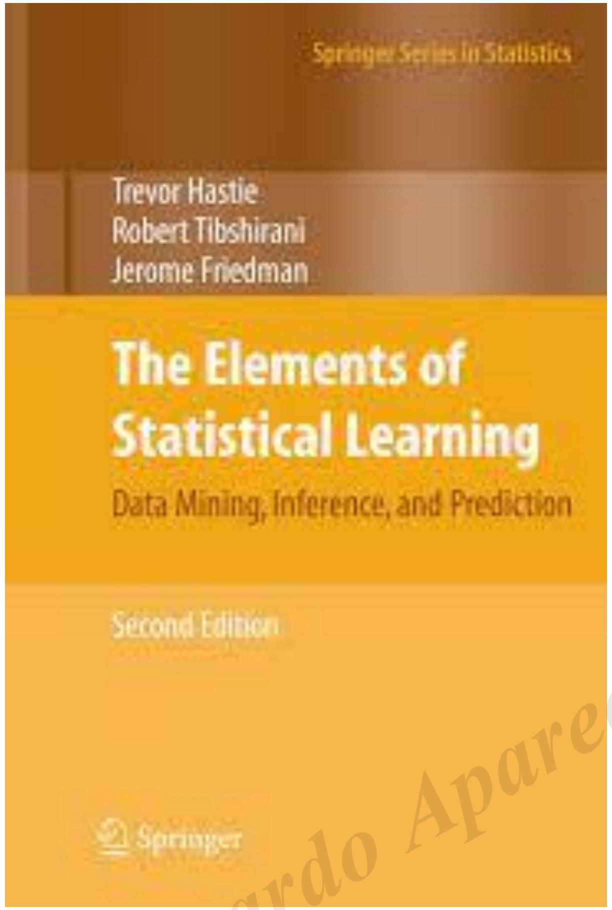


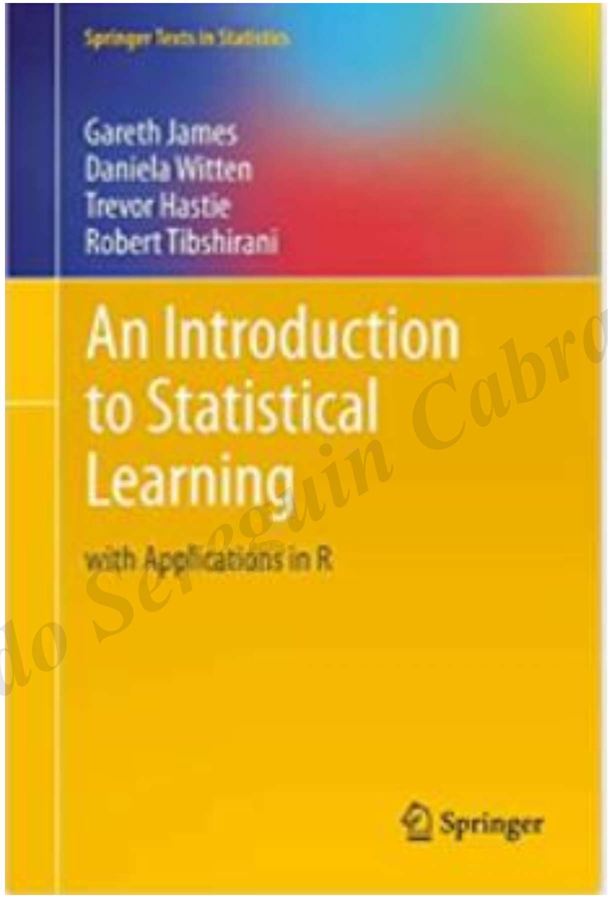
Source: Natural Language Processing with Python

The most common are automobile labels, therefore, it begins there.

The words "dark" (weak indicator of mystery) and "football" (a strong indicator of sports) appear.

Naive Bayes



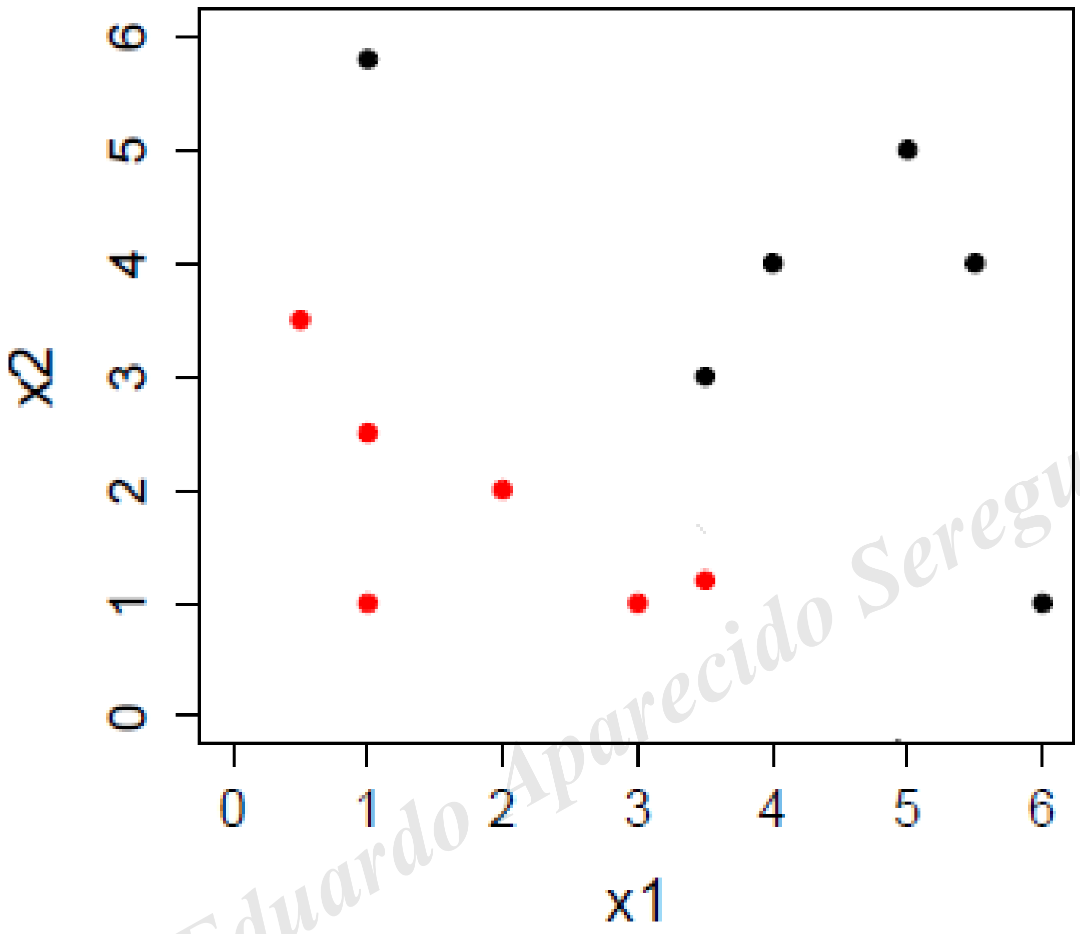


https://www.ime.unicamp.br/~dias/Intoduction%20to%20Statistical%20Learning.pdf

Support Vector Machine

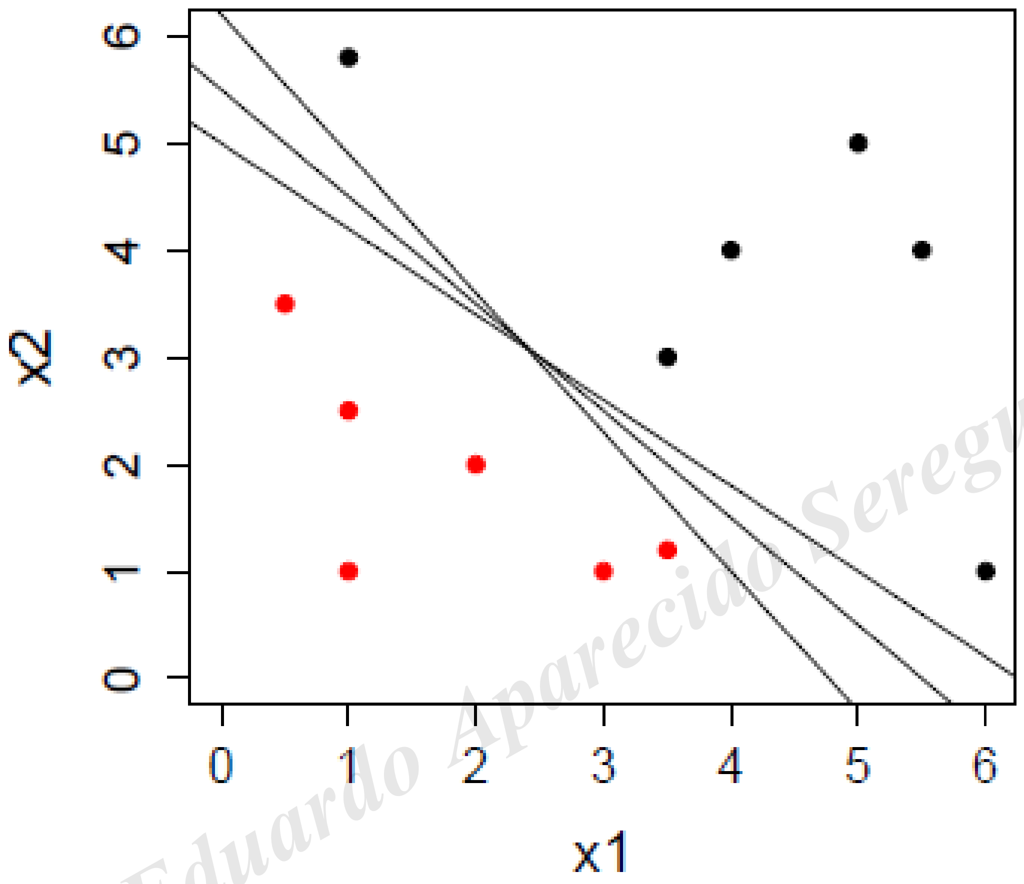
- It seeks to find the best separating hyperplane between two classes
- 3 possibilities: the maximal-margin classifier, flexible margin classifier and a non-linear margin classifier.

Support Vector Machine



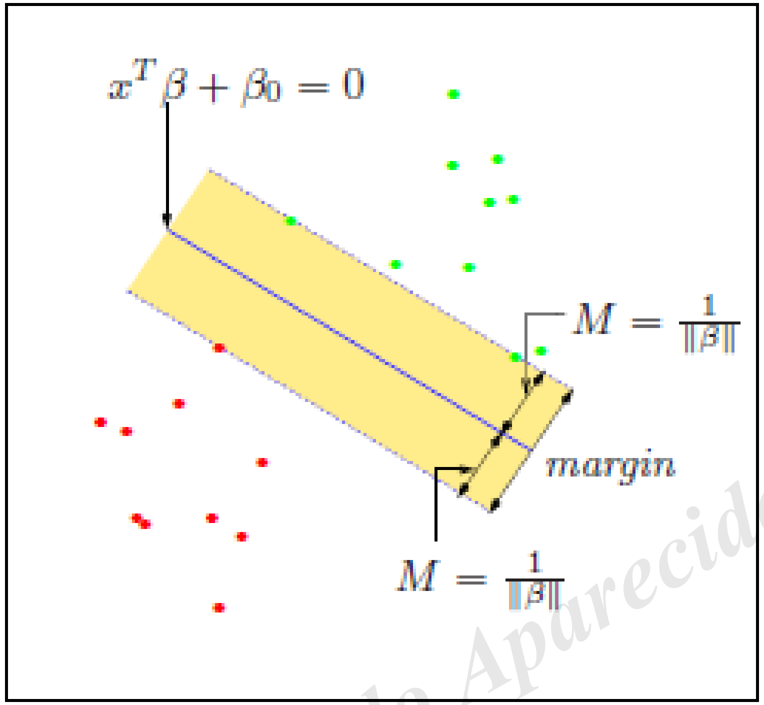
Source: Data Science - Morettin

Support Vector Machine

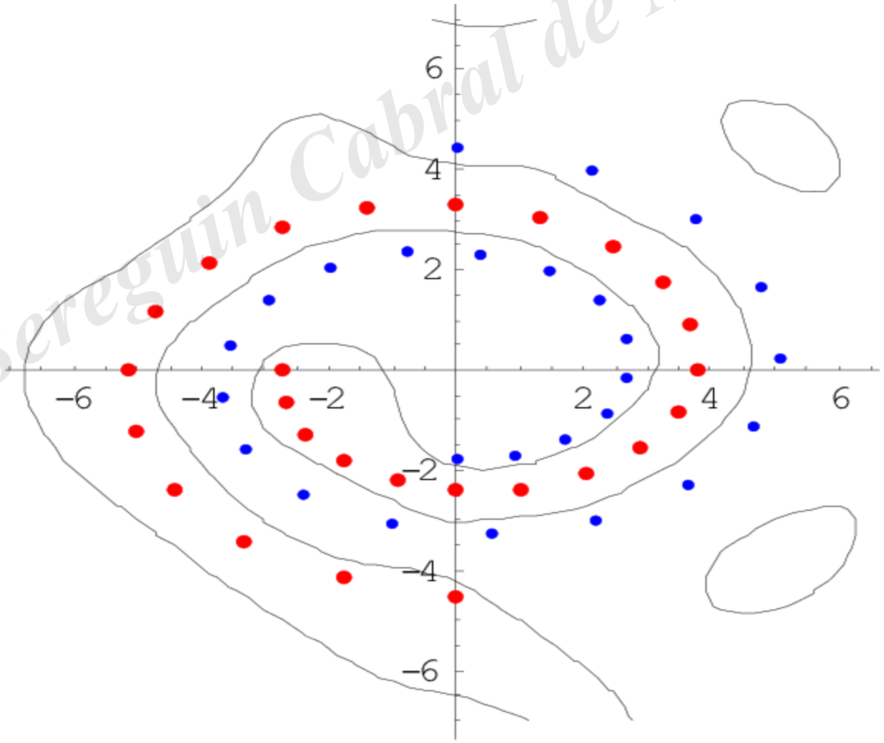


Source: Data Science - Morettin

Support Vector Machine

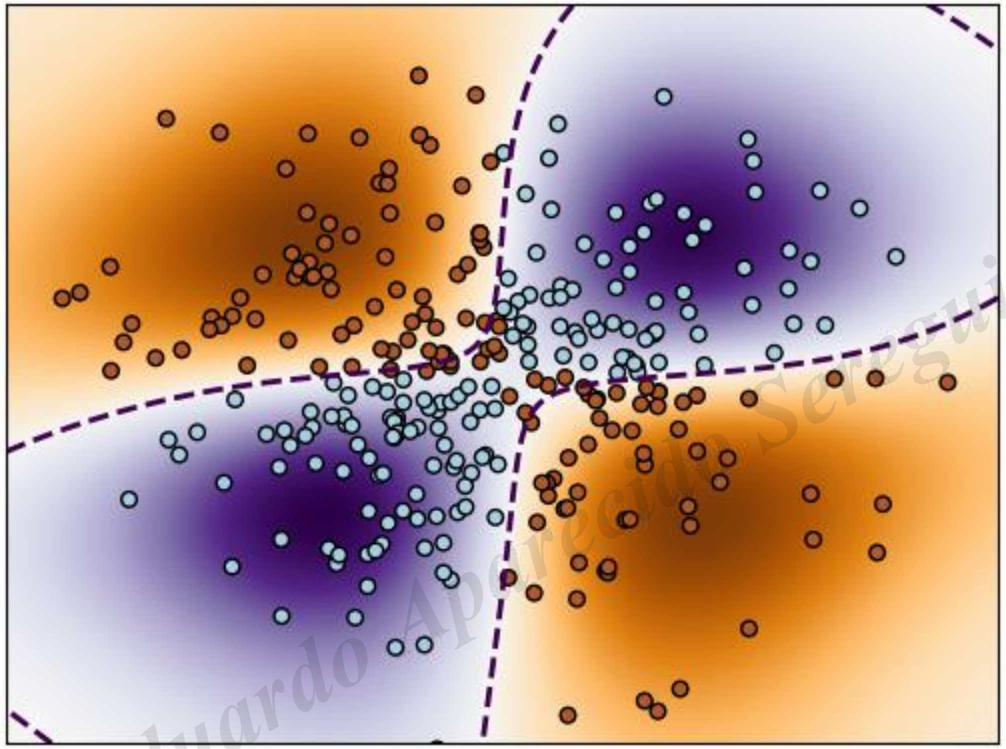


Source: Elements of Statistical Learning



Source: https://researchgate.com/

Flexible Margin



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

Support Vector Machine

- Objective: to separate classes = to classify the texts

- Objective function:

1. To maximiza the margin.
2. Subject to the fact that each point should be greater than the margin.
3. And subject to a possible term of error in flexible margin models.

Performance

- Not always the best first solution.

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.

Discussion

Future of NLP and trends.