FORMAÇÃO CIENTISTA DE DADOS

MINERAÇÃO DE TEXTO

MINERAÇÃO DE TEXTO

- **DADOS ESTRUTURADOS VS DADOS NÃO ESTRUTURADOS**

No.	1: outlook	2: temperature	3: humidity	4: windy	5: **play**
	Nominal	Nominal	Nominal	Nominal	Nominal
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

**board of directors approved a two-for-one stock split of its
common shares for shareholders of record as of April 1, 1987.
The company also said its board voted to recommend to
shareholders at the annual meeting April 23 an increase in the
authorized capital stock from five mln to 25 mln shares.**

APLICAÇÕES

- ANALISE DE SENTIMENTO
- CLASSIFICAÇÃO DE DOCUMENTOS
- DETECÇÃO DE FRAUDES
- ANÚNCIOS CONTEXTUALIZADOS
- FILTRO DE SPAM

PACOTE tm

- CORPUS: CONJUNTO DE DOCUMENTOS
  - VCORPUS: VOLÁTIL
  - PCORPUS: PERSISTENTE
- FONTES DE TEXTO:
  - DATAFRAMESOURCE, DIRSOURCE, URISOURCE, VECTORSOURCE, XMLSOURCE
- FORMATO DO TEXTO:
  - READPDF, READPLAIN, REDXML, READTABULAR...

PACOTE tm

- TM_MAP: FUNÇÕES DE TRANSFORMAÇÃO
- STOPWORDS: PALAVRAS SEM VALOR SEMÂNTICO

AULA PRÁTICA

- CRIAR UM CORPUS COM MAIS DE 1500 DOCUMENTOS
- GERAR UMA NUVEM DE PALAVRAS
- GERAR TABELA DE TERMOS MAIS FREQUENTES