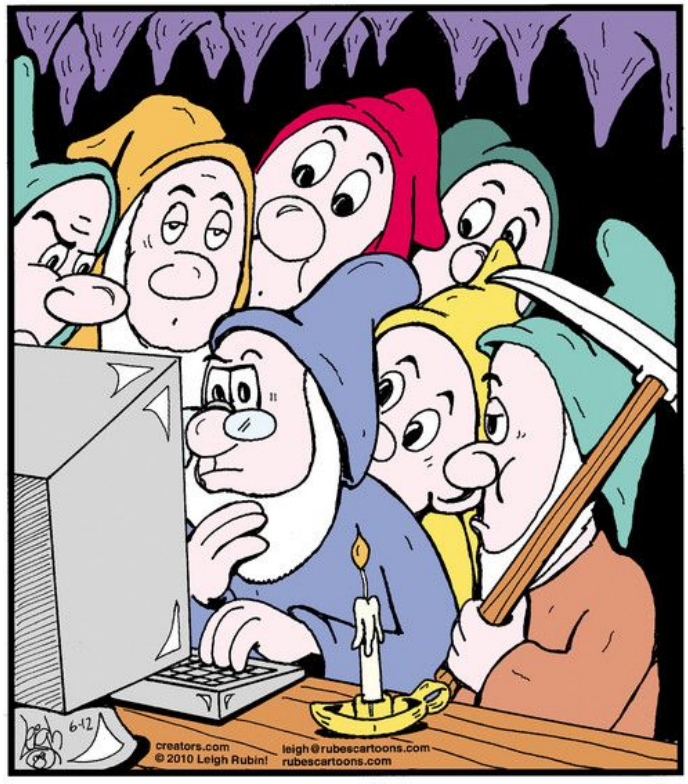
Introdução à Mineração de Dados

Luciana M. Abud

Instituto de Matemática e Estatística
Universidade de São Paulo

Introdução

- Exploração e análise de grande quantidade de dados
- Padrões e regras significativos



Data mining

Dados

- Em geral:
  - Grande volume
  - Dados estruturados
  - Pré processamento

Tarefas

- Classificação: categorizar dados não classificados
- Regressão: definir valor para variável contínua desconhecida
- Associação: determinar coocorrências
- Segmentação (*Clustering*): partição dos dados em subgrupos mais homogêneos
- Sumarização: encontrar descrição compacta para um subconjunto de dados

Aplicações

- Detecção de fraude (cartão de crédito, telecomunicação, sistemas de computadores, etc.)
- Diagnóstico médico
- Análise de sentimento em textos
- Segmentação de clientes

Mineração de Opinião

Introdução

- Redes sociais, fóruns, *tweets*, sites para avaliação de produtos e serviços, sites de notícias, etc.
- Fonte de opinião para organizações e empresas

Etapas

- Pré processamento
- Identificação
- Classificação da polaridade
- Sumarização

Pré processamento

- Representação do texto

T1: "@VirginAmerica everything was fine until you lost my bag"

T2: "@united Either is fine. However, plundering my hard-earned dollars is not fine."

  - Vetor binário

	bag	dollars	either	everything	fine	hard-earned	however	is	lost	my	not	plundering	until	was	you
T1	1	0	0	1	1	0	0	1	1	1	0	0	1	1	1
T2	0	1	1	0	1	1	1	1	0	1	1	1	0	0	0

  - Vetor de contagem de termos

	bag	dollars	either	everything	fine	hard-earned	however	is	lost	my	not	plundering	until	was	you
T1	1	0	0	1	?	0	0	1	1	?	0	0	1	1	1
T2	0	1	1	0	?	1	1	?	0	1	1	1	0	0	0

Pré processamento

- *Stop words*
- Lematização
  - am, are, is $\rightarrow$ be
  - better $\rightarrow$ good
- Stemização
  - follow, followed, follows $\rightarrow$ follow
  - university, universe, universal $\rightarrow$ univers

Identificação

- Alvo da opinião (produto, empresa, pessoa, etc.)
- Aspectos do alvo (preço, qualidade do serviço, roupa, etc.)

Classificação da polaridade

- Classificação entre positivo e negativo (ou também neutro)
- Limitações com ironias, sarcasmos, subjetividades, entre outros

Sumarização

- Criação de métricas para representar a diversidade de opiniões encontradas
- Representação em tabelas e gráficos

Classificação

- Aprendizagem de máquina
  - conjuntos de treino e de teste
- Análise sintática: dicionário de sentimentos
  - *good*, *excellent* $\rightarrow$ positivo
  - *bad*, *terrible* $\rightarrow$ negativo
- Análise estatística: avaliação de coocorrência de termos
  - ocorrência frequente junto a palavras positivas $\rightarrow$ provavelmente positivo
  - ocorrência frequente junto a palavras negativas $\rightarrow$ provavelmente negativo

Ferramentas

- Pandas: Python Data Analysis Library.
  `http://pandas.pydata.org/`
- Scikit-Learn: Machine Learning in Python
  `http://scikit-learn.org/`
- WEKA: Machine learning software to solve data mining problems.
  `https://sourceforge.net/projects/weka/`
- R: The R Project for Statistical Computing
  `https://www.r-project.org/`
  `http://www.rdatamining.com/home`

Referências

- CAMILO, Cássio Oliveira; SILVA, João Carlos da. Mineração de Dados: Conceitos, Tarefas, Métodos e Ferramentas. Goiânia: Universidade Federal de Goiás, 2009.
- Michael J. Berry and Gordon Linoff. 1997. Data Mining Techniques: For Marketing, Sales, and Customer Support. John Wiley & Sons, Inc., New York, NY, USA.
- S. Sumathi and S. N. Sivanandam. 2006. Introduction to Data Mining and its Applications (Studies in Computational Intelligence). Springer-Verlag New York, Inc., Secaucus, NJ, USA.
- Introdução à Mineração de Opiniões: Conceitos, Aplicações e Desafios. K Becker, D Tumitan. Simpósio Brasileiro de Banco de Dados, 27-52, 2013. 12, 2013.
- Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. Introduction to Information Retrieval. Cambridge University Press, New York, NY, USA.