BIG DATA

LABDATA

FUNDAÇÃO INSTITUTO DE ADMINISTRAÇÃO

Introdução ao Big Data

Tema da Aula: **Text Mining**

Prof.: **Dino Magri**

Coordenação:

Prof. Dr. Adolpho Walter Pimazzi Canton

Profa. Dra. Alessandra de Ávila Montini

- Contatos:
  - E-mail: professor.dinomagri@gmail.com
  - Twitter: https://twitter.com/prof_dinomagri
  - LinkedIn:http://www.linkedin.com/in/dinomagri
  - Site: http://www.dinomagri.com

Coordenação:

Prof. Dr. Adolpho Walter Pimazzi Canton

Profa. Dra. Alessandra de Ávila Montini

Currículo

- **(2014-Presente)** – Professor no curso de Extensão, Pós e MBA na Fundação Instituto de Administração (FIA) – www.fia.com.br
- **(2018-Presente)** – Pesquisa e Desenvolvimento de Big Data e Machine Learning na Beholder (http://beholder.tech)
- **(2013-2018)** – Pesquisa e Desenvolvimento no Laboratório de Arquitetura e Redes de Computadores (LARC) na Universidade de São Paulo – www.larc.usp.br
- **(2012)** – Bacharel em Ciência da Computação pela Universidade do Estado de Santa Catarina (UDESC) – www.cct.udesc.br
- **(2009/2010)** – Pesquisador e Desenvolvedor no Centro de Computação Gráfica – Guimarães – Portugal – www.ccg.pt
- **Lattes:** http://lattes.cnpq.br/5673884504184733

Material das aulas

- Caso esteja utilizando seu próprio computador, realize o download de todos os arquivos e salve na **Área de Trabalho** para facilitar o acesso.
  - Lembre-se de instalar os softwares necessários conforme descrito no documento de Instalação (**`InstalaçãoPython3v1.2.pdf`**).

- Nos computadores da FIA os arquivos já estão disponíveis, bem como a instalação dos softwares necessários.

Conteúdo da Aula

- Objetivo
- Mineração de Texto
- Processamento Linguagem Natural
- Classificação do Texto
- Desenvolvimento
- Referências

Conteúdo da Aula

- **Objetivo**
- Mineração de Texto
- Processamento Linguagem Natural
- Classificação do Texto
- Desenvolvimento
- Referências

Conteúdo da Aula

- O objetivo dessa aula é introduzir os conceitos envolvidos na **mineração de texto** através do desenvolvimento de uma aplicação que realiza análise de sentimento.

Conteúdo da Aula

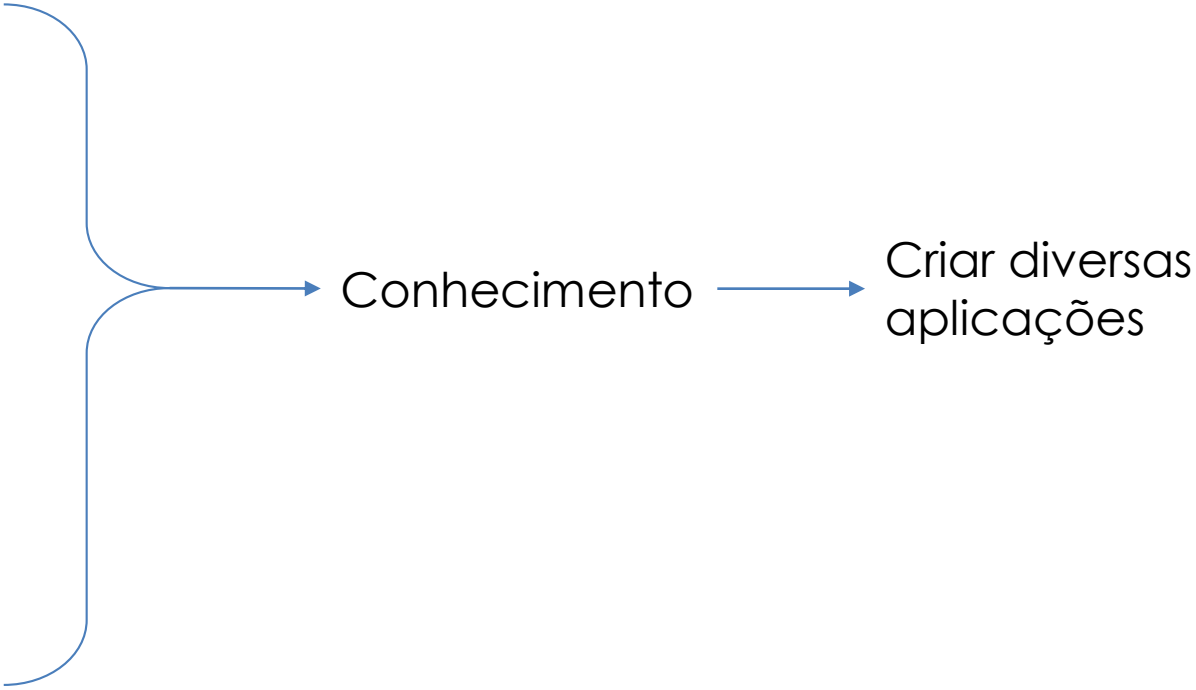
- Objetivo
- **Mineração de Texto**
- Processamento Linguagem Natural
- Classificação do Texto
- Desenvolvimento
- Referências

Mineração de Texto

- Como vimos, existem diversas formas de obter dados.
- As evoluções que ocorreram nas áreas de Big Data, Mídias Sociais, Mobile e Computação em nuvem, possibilitam um aumento exponencial no volume dos dados.
- O grande desafio é **extrair dados** não estruturados para realizar análises e tomar as decisões necessárias para cada tipo de negócio.

Mineração de Texto

- Internet
- Blogs
- Noticias
- E-mails
- Livros
- Mídias Sociais



Mineração de Texto

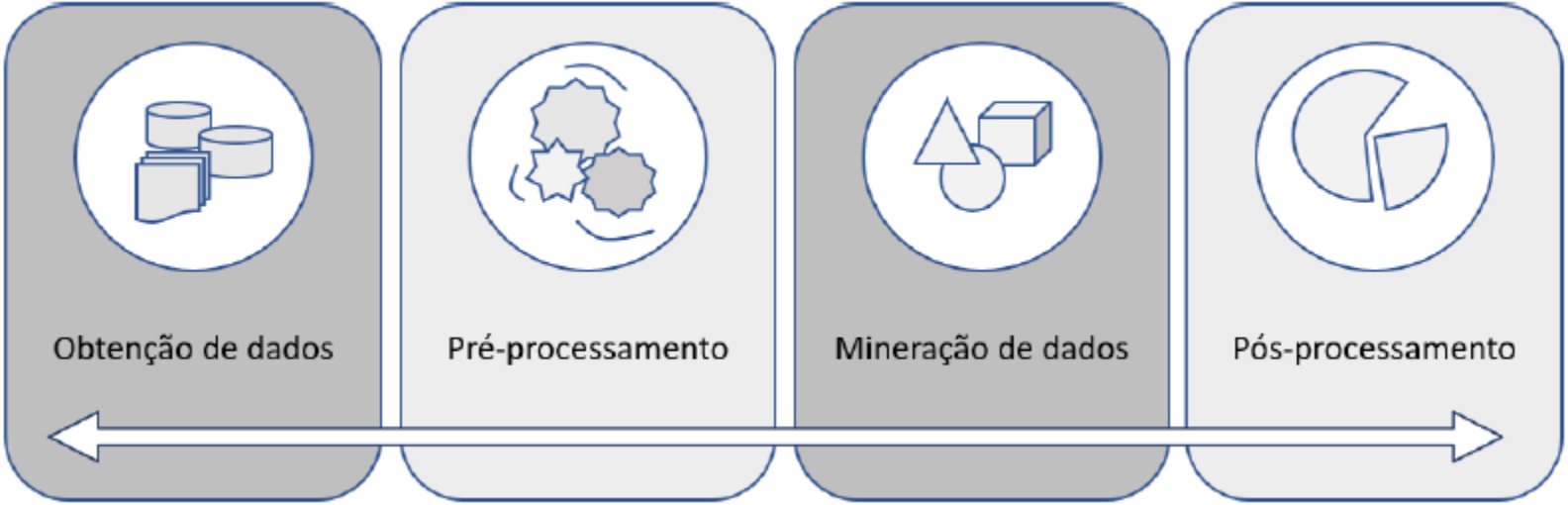
- Tradução de texto
- Identificação dos atos de fala
- Extração de informação
- Categorização de texto
- Atendimento por Bots
- Análise de mídias sociais
- Filtragem de Spam
- Análise de sentimento
- ...

Mineração de Texto

- Logo, o **processo de descoberta de conhecimento** no conjunto de dados é muito importante.
  - *KDD – Knowledge Discovery in Databases*

- É um processo automático, sistemático e analítico que possibilita realizar a mineração nos dados.

Mineração de Texto



Fonte: Silva (2016)

Mineração de Texto

- Mineração de texto (Text Mining) é um processo que constitui a **descoberta de padrões**, **associações** e mudanças para a produção do conhecimento.
- A mineração de texto é uma **área ampla** com diversos conceitos e possui diferentes abordagens e técnicas.

Mineração de Texto

- Iremos realizar a mineração de texto com foco na **análise de sentimento** (positivo, negativo ou neutro) em comentários coletados no Youtube.
- As principais etapas envolvidas:
  - Coleta dos dados (texto)
  - Processamento do texto
  - Criação do classificador
  - Utilização do classificador gerado nos dados do Youtube.

Conteúdo da Aula

- Objetivo
- Mineração de Texto
- **Processamento Linguagem Natural**
- Classificação do Texto
- Desenvolvimento
- Referências

Processamento de Linguagem Natural

- Linguagem natural é aquela utilizada para comunicação entre pessoas.
- Em contraste com as linguagens artificiais (linguagem de programação e notações matemáticas), **as linguagens evoluem à medida que passam de geração para geração**, sendo difícil defini-las por meio de regras explicitas.

Processamento de Linguagem Natural

- O estudo do processamento de linguagem natural (PLN) tem como objetivo criar técnicas computacionais para que seja possível compreender os textos.
- Desta forma, é necessário abstrair e estruturar a língua a partir das regras dos idiomas em que os textos estão escritos.
- Existem **diversas técnicas** que podem ser utilizadas para abstrair e estruturar um texto, elas devem ser analisadas no contexto de cada aplicação, observando seus objetivos.

Processamento de Linguagem Natural

- Para realizar a análise de sentimento, as técnicas de PLN comumente utilizadas são:
  - Normalização
  - Eliminação dos termos irrelevantes
  - Redução do termo ao radical
  - Representação dos documentos

Processamento de Linguagem Natural

- Para melhor entendimento das etapas envolvidas no pré-processamento, considere os seguintes documentos (frases):

Doc 01	Ambiente agradável e tranquilo. Comida e música com qualidade. Adoramos o filé à parmegiana!!!!!
Doc 02	O filé à parmegiana da cidade. Ambiente agradável e qualidade no atendimento.
Doc 03	O filé à parmegiana com fritas é uma delícia.

Processamento de Linguagem Natural

- **Normalização**
- Eliminação dos termos irrelevantes
- Redução do termo ao radical
- Representação dos atributos

Processamento de Linguagem Natural

- **Normalização**
  - Retira-se a pontuação dos textos, acentos, caracteres especiais.
  - Realiza-se a padronização de todas as palavras em minúsculo.
  - Separa a frase por palavras

Processamento de Linguagem Natural

- **Normalização**
  - As regras de normalização dependem do contexto da análise desejada.

Doc 01	'ambiente' 'agradavel' 'e' 'tranquilo' 'comida' 'e' 'musica' 'com' 'qualidade' 'adoramos' 'o' 'file' 'a' 'parmegiana'
Doc 02	'o' 'file' 'a' 'parmegiana' 'da' 'cidade' 'ambiente' 'agradavel' 'e' 'qualidade' 'no' 'atendimento'
Doc 03	'o' 'file' 'a' 'parmegiana' 'com' 'fritas' 'e' 'uma' 'delicia'

Processamento de Linguagem Natural

- Normalização
- **Eliminação dos termos irrelevantes**
- Redução do termo ao radical
- Representação dos atributos

Processamento de Linguagem Natural

- **Eliminação dos termos irrelevantes**
  - É importante eliminar palavras muito frequentes em todos os documentos como **artigos**, **preposições** e **conjunções**; pois não possuem significado relevante para a construção da análise e podem aumentar consideravelmente a estrutura de indexação do conjunto final dos dados.
  - Essas palavras, também são conhecidas como Stop-words.
  - Podem representar uma redução de 30 a 50% do tamanho do conjunto final.

Processamento de Linguagem Natural

- **Eliminação dos termos irrelevantes (stop-words)**
  - É um processo que deve ser executado de forma iterativa e interativa.

Doc 01	'ambiente' 'agradavel' 'tranquilo' 'comida' 'musica' 'qualidade' 'adoramos' 'file' 'parmegiana'
Doc 02	'file' 'parmegiana' 'cidade' 'ambiente' 'agradavel' 'qualidade' 'atendimento'
Doc 03	'file' 'parmegiana' 'fritas' 'delicia'

Processamento de Linguagem Natural

- Normalização
- Eliminação dos termos irrelevantes
- **Redução do termo ao radical**
- Representação dos atributos

Processamento de Linguagem Natural

- **Redução do termo ao radical**
  - Lematização (*Stemming*) é uma técnica utilizada para reduzir a palavra em seu radical de acordo com as regras do idioma.
  - Por exemplo, a palavra `amor` pode variar em `amado`, `amou`, `amará`, `amante`, entre outras. Essa variação é dada pela identificação e retirada do seu afixo, **`"am"`**.
  - Essa tarefa reduz as variações de uma palavra para um conceito comum, aumentando a força do atributo e reduzindo o tamanho da estrutura de indexação.

Processamento de Linguagem Natural

- **Redução do termo ao radical**
  - Uma técnica possível de *stemming*, **é a retirada dos sufixos das palavras**. Essa abordagem utiliza os seguintes passos:
    - Passo 1: são removidos sufixos terminados pela letra s;
    - Passo 2: são removidos alguns sufixos de adjetivos e substantivos, além de formas verbais no pretérito e no particípio passado;
    - Passo 3: são retirados os principais sufixos formadores de plural de substantivos e adjetivos;
    - Passo 4: são tratadas palavras que são exceção às regras sufixais.
  - O processo de lematização para a língua portuguesa é o **RSLP** (Removedor de Sufixos da Língua Portuguesa), que utiliza o algoritmo de Porter.

Processamento de Linguagem Natural

- **Redução do termo ao radical**
  - É importante analisar quais benefícios se adquiri utilizando essa etapa em diferentes algoritmos.

Doc 01	'ambient' 'agrad' 'tranquil' 'comid' 'music' 'qualidad' 'ador' 'fil' 'parmegian'
Doc 02	'fil' 'parmegian' 'cidad' 'ambient' 'agrad' 'qualid' 'atend'
Doc 03	'fil' 'parmegian' 'frit' 'delici'

Processamento de Linguagem Natural

- Normalização
- Eliminação dos termos irrelevantes
- Redução do termo ao radical
- **Representação dos atributos**

Processamento de Linguagem Natural

- **Representação dos atributos**
  - O conjunto de dados final é criado utilizando uma representação vetorial dos atributos.
  - Basicamente atribui-se o valor 1 para representar a presença do atributo no documento e 0 para ausência.

Processamento de Linguagem Natural

- **Representação dos documentos**

	ador	agrad	ambient	atend	cidad	comid	delici	fil	frit	music	parmegian	qualidad	tranquil
Doc 01	1	1	1	0	0	1	0	1	0	1	1	1	1
Doc 02	0	1	1	1	1	0	0	1	0	0	1	1	0
Doc 03	0	0	0	0	0	0	1	1	1	0	1	0	0

Conteúdo da Aula

- Objetivo
- Mineração de Texto
- Processamento Linguagem Natural
- **Classificação do Texto**
- Desenvolvimento
- Referências

Classificação do Texto

- Uma vez que temos o conjunto de dados podemos realizar sua classificação em Positivo, Negativo e Neutro.
- Para realizar a classificação existem diversos métodos:
  - Manual
  - Baseado em regras
  - **Estatístico/Probabilístico**

Classificação do Texto

- Iremos utilizar o algoritmo **Naive Bayes** para criar um modelo e realizar a classificação do texto.
- É um algoritmo probabilístico de classificação que se baseia no teorema de Bayes, cujo intuito é reconhecer padrões e realizar predições.
- É um classificador denominado ingênuo, pois assume que os atributos são independentes entre si.
- Tem custo **computacional significativo** para determinar a hipótese ótima.

Classificação do Texto

- Na teoria da probabilidade, o teorema de Bayes mostra a relação entre uma probabilidade condicional e a sua inversa, ou seja, **a probabilidade de uma hipótese, dada a observação de uma evidência; e a probabilidade da evidência, dada pela hipótese**.
- Por exemplo: qual a probabilidade de chover (hipótese), dado que está ventando e com muitas nuvens (evidência)?

Classificação do Texto

- O classificador **Naive Bayes** é capaz de reconhecer os padrões presentes nos dados de entrada e é utilizado para fazer predições em novos dados.
- Considere um conjunto de dados que possui as seguintes frases:

Frase original	Frase pré-processada	Classe
Me sinto completamente amado	sint complet am	Positivo
Eu estou muito bem hoje	bem hoj	Positivo
Eu sinto amor por você	sint am	Positivo
Isso me deixa apavorada	deix apavor	Negativo
Este lugar é apavorante	lug apavor	Negativo

Classificação do Texto

- As probabilidades são calculadas:

Emoção	sint		complet		am		bem		hoj		deix		apavor		lug	
	S	N	S	N	S	N	S	N	S	N	S	N	S	N	S	N
	2	3	1	4	2	3	1	4	1	4	1	4	2	3	1	4
Positivo 3/5	2/2	1/3	1/1	2/4	2/2	1/3	1/1	2/4	1/1	2/4	0	3/4	0	3/3	0	3/4
Negativo 2/5	0	2/3	0	2/4	0	2/3	0	2/4	0	2/4	1/1	1/4	2/2	0	1/1	1/4

Classificação do Texto

- Como a frase "**`Me sinto completamente bem neste lugar`**" será classificada?
  - Frase pré-processada: `sint complet bem lug`
  - Desta forma, teremos:
    - `sint=S, complet=S, am=N, bem=S, hoje=N, deix=N, apavor=N, lug=S`
  - Verifica-se a probabilidade de ser **Positivo** e a probabilidade de ser **Negativo**.

Classificação do Texto

- As probabilidades são calculadas:

Emoção	sint		complet		am		bem		hoj		deix		apavor		lug	
	S	N	S	N	S	N	S	N	S	N	S	N	S	N	S	N
	2	3	1	4	2	3	1	4	1	4	1	4	2	3	1	4
Positivo 3/5	2/2	1/3	1/1	2/4	2/2	1/3	1/1	2/4	1/1	2/4	0	3/4	0	3/3	0	3/4
Negativo 2/5	0	2/3	0	2/4	0	2/3	0	2/4	0	2/4	1/1	1/4	2/2	0	1/1	1/4

Classificação do Texto

- Desta forma, podemos calcular as probabilidades:
  - P(positivo) = 3/5 * 2/2 * 1/1 * 1/3 * 1/1 * 2/4 * 3/4 *3/3 *** 0**
  - P(negativo) = 2/5 *** 0 * 0** * 2/3 *** 0** * 2/4 * 2/4 * 1/4 *** 0** * 1/1

Classificação do Texto

- Caso algum atributo nunca ocorra para uma classe, podemos utilizar o **Correção de Laplace** que soma 1 a contagem de todas as combinações da classe e valor de atributo. Por exemplo:

Emoção	sint		complet		am		bem		hoj		deix		apavor		lug	
	S	N	S	N	S	N	S	N	S	N	S	N	S	N	S	N
	2 (3)	3	1	4	2	3	1	4	1	4	1	4	2	3	1	4
Positivo 3/5 (3/6)	2/2 (2/3)	1/3	1/1	2/4	2/2	1/3	1/1	2/4	1/1	2/4	0	3/4	0	3/3	0	3/4
Negativo 2/5 (3/6)	0 (1/3)	2/3	0	2/4	0	2/3	0	2/4	0	2/4	1/1	1/4	2/2	0	1/1	1/4

Classificação do Texto

- Como o conjunto de dados que estamos utilizando nesse exemplo é muito pequeno, **vamos ignorar esses valores**, pois quando adicionamos um novo documento para corrigir o valor 0, ele passa a representar 1/3, sendo uma probabilidade muito próxima da outra classe (2/3), o que impactaria significativamente o resultado final da probabilidade que estamos calculando.
  - **Isso não ocorre com *corpus* muito grande**, pois o valor da probabilidade é relativamente pequeno.

- Desta forma, podemos calcular as probabilidades:
  - P(positivo) = 3/5 * 2/2 * 1/1 * 1/3 * 1/1 * 2/4 * 3/4 *3/3 = 0,075
  - P(**positivo**) = (0,075 / 0,092) * 100 = **81,52%**

  - P(negativo) = 2/5 * 2/3 * 2/4 * 2/4 * 1/4 * 1/1 = 0,017
  - P(**negativo**) = (0,017 / 0,092) * 100 = **18,48%**

Classificação do Texto

- Logo, a frase "`Me sinto completamente bem neste lugar`" será classificada, como **Positivo com 81,52%.**

Referência: **Estudo empírico do impacto do tamanho do corpus no desempenho do classificador Naive Bayes** (Paola Cunha, 2018).

Conteúdo da Aula

- Objetivo
- Mineração de Texto
- Processamento Linguagem Natural
- Classificação do Texto
- **Desenvolvimento**
- Referências

Desenvolvimento

- Utilizaremos a biblioteca NLTK do Python.
- É uma plataforma para criação de programas Python para trabalhar com dados de linguagem humana.
- Ele fornece interfaces fáceis de usar para mais de 50 *corpora* e recursos léxicos, além é claro de conter um conjunto de ferramentas para processar o texto, como:
  - Classificação, tokenização, derivação, marcação, análise e raciocínio semântico.
- Os autores originais foram Steven Bird, Edward Loper, Ewan Klein.
- A primeira versão foi publicada em **2001**.

Desenvolvimento

- Originalmente projetada para ensinar os conceitos envolvido no processamento de linguagem natural.
- Tem sido adotada na área de **pesquisa & desenvolvimento** de diversas empresas.
- É uma biblioteca com foco na **simplicidade**, **consistência** e **modularidade**.

Desenvolvimento

- Para instalar, abra o CMD ou Terminal e digite:

```
pip install nltk
```

Desenvolvimento

- Como nosso objetivo é realizar a classificação de texto em positivo, negativo e neutro precisamos de uma base histórica com frases previamente classificadas.
- Iremos utilizar uma base de dados contendo tweets relacionados à Copa do Mundo de 2018, devidamente classificadas.
- Essa base será utilizada para criarmos o nosso **classificador Naive Bayes**.

</> Abra o arquivo **"aula9-parte1-nltk.ipynb"**

Desenvolvimento

- Depois que o classificador foi criado e testado, iremos utiliza-lo para realizar a classificação dos comentários do YouTube capturados de vídeos com o tema da copa do mundo.

</> Abra o arquivo **"aula9-parte2-analise-sentimento.ipynb"**

Conteúdo da Aula

- Objetivo
- Mineração de Texto
- Processamento Linguagem Natural
- Classificação do Texto
- Desenvolvimento
- **Referências**

Referências Bibliográficas

- **Natural Language Processing with Python** - Steven Bird, Ewan Klein e Edward Loper - O'Reilly, 2016.

- **Python Text Processing with NLTK 2.0 Cookbook** - Jacob Perkins - Packt, 2010.

- **Recuperação de Informação** – Ricardo Baeza-Yates, Berthier Ribeiro-Neto, Bookman, 2013.

Referências Bibliográficas

- **Introdução à Mineração de dados** – Leandro Augusto da Silva, Sarajane Marques Peres, Clodis Boscarioli, Elsevier, SBC, 2016.

- **Data Science from Scratch** – Joel Grus – O'Reilly, 2015.

- **Python for Data Analysis** – Wes McKinney – USA: O'Reilly, 2013.

- As referências de links utilizados podem ser visualizados em http://urls.dinomagri.com/refs