



Analytics e Inteligência Artificial

Aula 5, 6 e 7

Processamento de Texto e Análise de Sentimentos



FIA Business School

NOSSOS DIFERENCIAIS | QUEM SOMOS

2



BUSINESS SCHOOL

Graduação, pós-graduação, MBA, Pós- MBA, Mestrado Profissional, Curso In Company e EAD



CONSULTING

Consultoria personalizada que oferece soluções baseadas em seu problema de negócio



RESEARCH

Atualização dos conhecimentos e do material didático oferecidos nas atividades de ensino



Líder em Educação Executiva, referência de ensino nos cursos de graduação, pós-graduação e MBA, tendo excelência nos programas de educação. Uma das principais **escolas de negócio do mundo**, possuindo convênios internacionais com Universidades nos EUA, Europa e Ásia. +8.000 **projetos de consultorias** em organizações públicas e privadas.



Único curso de graduação em administração a receber as notas máximas



A primeira escola brasileira a ser finalista da maior competição de MBA do mundo



Única *Business School* brasileira a figurar no *ranking* LATAM



Signatária do Pacto Global da ONU



Membro fundador da ANAMBA - Associação Nacional MBAs



Credenciada pela AMBA - Association of MBAs



Credenciada ao Executive MBA Council



Filiada a AACSB - Association to Advance Collegiate Schools of Business



Filiada a EFMD - European Foundation for Management Development



Referência em cursos de MBA nas principais mídias de circulação



O **Laboratório de Análise de Dados** – LABDATA é um Centro de Excelência que atua nas áreas de ensino, pesquisa e consultoria em análise de informação utilizando técnicas de **Big Data, Analytics e Inteligência Artificial**.



Profª Drª Alessandra Montini

O LABDATA é um dos pioneiros no lançamento dos cursos de *Big Data* e *Analytics* no Brasil

Os diretores foram professores de grandes especialistas do mercado

+10 anos de atuação

+1000 alunos formados

Docentes

- Sólida formação acadêmica: doutores e mestres em sua maioria
- Larga experiência de mercado na resolução de cases
- Participação em Congressos Nacionais e Internacionais
- Professor assistente que acompanha o aluno durante todo o curso

Estrutura

- 100% das aulas realizadas em laboratórios
- Computadores para uso individual durante as aulas
- 5 laboratórios de alta qualidade (investimento +R\$2MM)
- 2 Unidades próximas a estação de metrô (com estacionamento)



Profª Dra.
Alessandra Montini

Diretora do LABDATA-FIA, apaixonada por dados e pela arte de lecionar. Têm muito orgulho de ter criado na FIA cinco laboratórios para as aulas de Big Data e inteligência Artificial. Possui mais de 20 anos de trajetória nas áreas de Data Mining, Big Data, Inteligência Artificial e Analytics. Cientista de dados com carreira realizada na Universidade de São Paulo. Graduada e mestra em estatística aplicada pelo IME-USP e doutora pela FEA-USP. Com muita dedicação chegou ao cargo de professora e pesquisadora na FEA-USP, ganhou mais de 30 prêmios de excelência acadêmica pela FEA-USP e mais de 30 prêmios de excelência acadêmica como professora dos cursos de MBA da FIA. Orienta alunos de mestrado e de doutorado na FEA-USP. Membro do Conselho Curador da FIA, Coordenadora de Grupos de Pesquisa no CNPQ, Parecerista da FAPESP e Colunista de grandes Portais de Tecnologia.

 [linkedin.com/in/alessandramontini/](https://www.linkedin.com/in/alessandramontini/)



Prof. Dr.
Adolpho Walter Canton

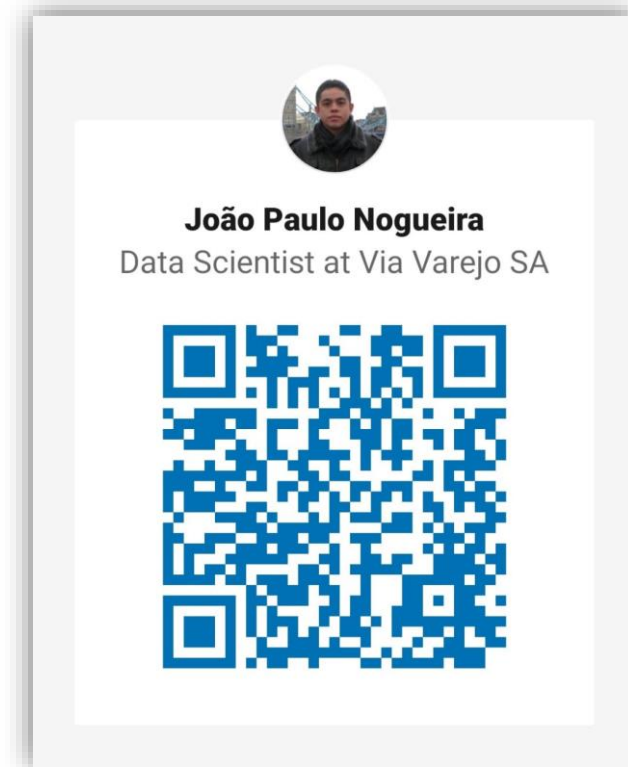
Diretor do LABDATA-FIA. Consultor em Projetos de *Analytics*, *Big Data* e Inteligência Artificial. Professor FEA – USP. PhD em Estatística Aplicada pela *University of North Carolina at Chapel Hill*, Estados Unidos.



Currículo – Prof. João Nogueira

FORMAÇÃO ACADÊMICA | EXPERIÊNCIA PROFISSIONAL

- **(2019-Presente)** – Professor nos cursos de Extensão, Pós e MBA em Big Data e Data Mining na Fundação Instituto de Administração (FIA) – www.fia.com.br
- **(2018-Presente)** – Cientista de Dados na Via Varejo – <https://viavarejo.com.br>
- **(2016-Presente)** – Doutorando em Física Computacional e Estatística pelo Departamento de Física na Universidade Federal do Ceará – <https://fisica.ufc.br>
- **(2014-2016)** – Mestre em Física da Matéria Condensada pelo Departamento de Física na Universidade Federal do Ceará - <https://fisica.ufc.br>
- **(2012-2013)** – Estudante Intercambista na Universidade de Coimbra – Portugal – <https://www.uc.pt>
- **(2010-2014)** – Bacharel em Física pela Universidade Federal do Ceará – <http://www.ufc.br>
- **Contatos:**
 - E-mail: joaonogueira@fisica.ufc.br



Currículo – Prof. Dr. Helio Cavalcante

FORMAÇÃO ACADÊMICA | EXPERIÊNCIA PROFISSIONAL

- **(2020-Presente)** – Professor nos cursos de Extensão, Pós e MBA em Big Data e Data Mining na Fundação Instituto de Administração – FIA
- **(2020-Presente)** – Especialista Cientista de Dados na Via Varejo
- **(2018-2021)** – Professor na Faculdade Metropolitanas Unidas – FMU
- **(2019-2020)** – Consultor em Inteligência Artificial na Atento
- **(2018-2019)** – Especialista em Inteligência Artificial na Almaxwave
- **(2010-2012)** – Professor Convidado na Universidade Federal de Alagoas – UFAL
- **(2012-2017)** – Doutor em Ciência e Tecnologia da Informação pela Universidade de Coimbra – UC em Portugal – PT
- **(2009-2010)** – Mestrado em Modelagem Computacional do Conhecimento pela Universidade Federal de Alagoas – UFAL
- **(2005-2008)** – Bacharel e Licenciatura em Análise e Desenvolvimento de Sistema pelo Centro de Estudos Superiores de Maceió – CESMAC
- **Contatos:**
 - E-mail: helio.hx@gmail.com
 - Lattes: <http://lattes.cnpq.br/9214503462764328>



Helio Cavalcante

Data Science Specialist na Via
Varejo SA



Currículo – Prof.Dino Magri

FORMAÇÃO ACADÊMICA | EXPERIÊNCIA PROFISSIONAL

7

- **(2014-Presente)** – Professor nos cursos de Extensão, Pós e MBA em Big Data e Data Mining 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
- **Contatos:**
 - E-mail: professor.dinomagri@gmail.com
 - Site: <http://www.dinomagri.com>
 - Twitter: <https://twitter.com/dinomagri>
 - Lattes: <http://lattes.cnpq.br/5673884504184733>



Conteúdo da Aula

-
-
- 1. O que é uma API?
- 2. Técnicas para Text Mining
- 3. spaCy
-



1. O que é uma API?



O que é uma API?

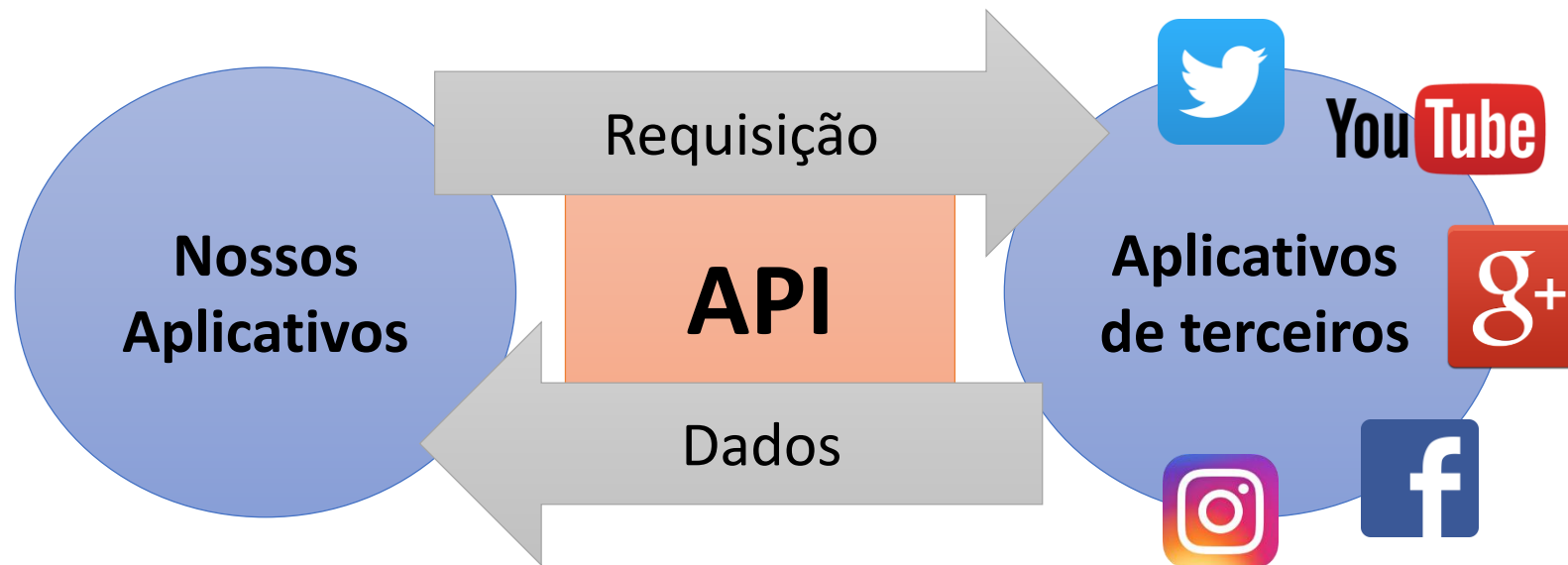


O Que é uma API?

- É um **conjunto de padrões** para acesso a uma plataforma baseada na Web.
- É uma interface de comunicação que **possibilita que diferentes tipos de software comuniquem entre si.**
- Uma API deve fornecer:
 - Flexibilidade, Segurança, Simplicidade no Uso, Escalabilidade, Portabilidade.



O Que é uma API?



<https://developer.twitter.com/en/docs>



APIs do Twitter

- Além disso, iremos utilizar **OAuth** para conectar usuários ao Twitter e enviar requisições seguras e autorizadas para as APIs do Twitter.
- Para acessar as APIs do Twitter precisamos ter acesso a desenvolvedor.
- É importante ter realizado os passos descritos na documentação: **[Criando uma Conta developer no Twitter.pdf](#)**



APIs do Twitter

- Precisamos recuperar as credenciais de acesso. Para isso:
 - Acesse <https://developer.twitter.com>
 - Clique em **Sign In** do lado direito superior
 - Realize o login.
 - Acesse diretamente o endereço:
 - <https://developer.twitter.com/en/apps>
 - Ou clique no nome de usuário que aparece e depois em Apps.



APIs do Twitter

- Clique em **Keys and tokens!**
- Pronto! Já temos as credenciais de acesso:
 - Consumer Key (API Key)
 - Consumer Secret (API Secret)
 - Access Token
 - Access Token Secret



APIs do Twitter

Tweepy

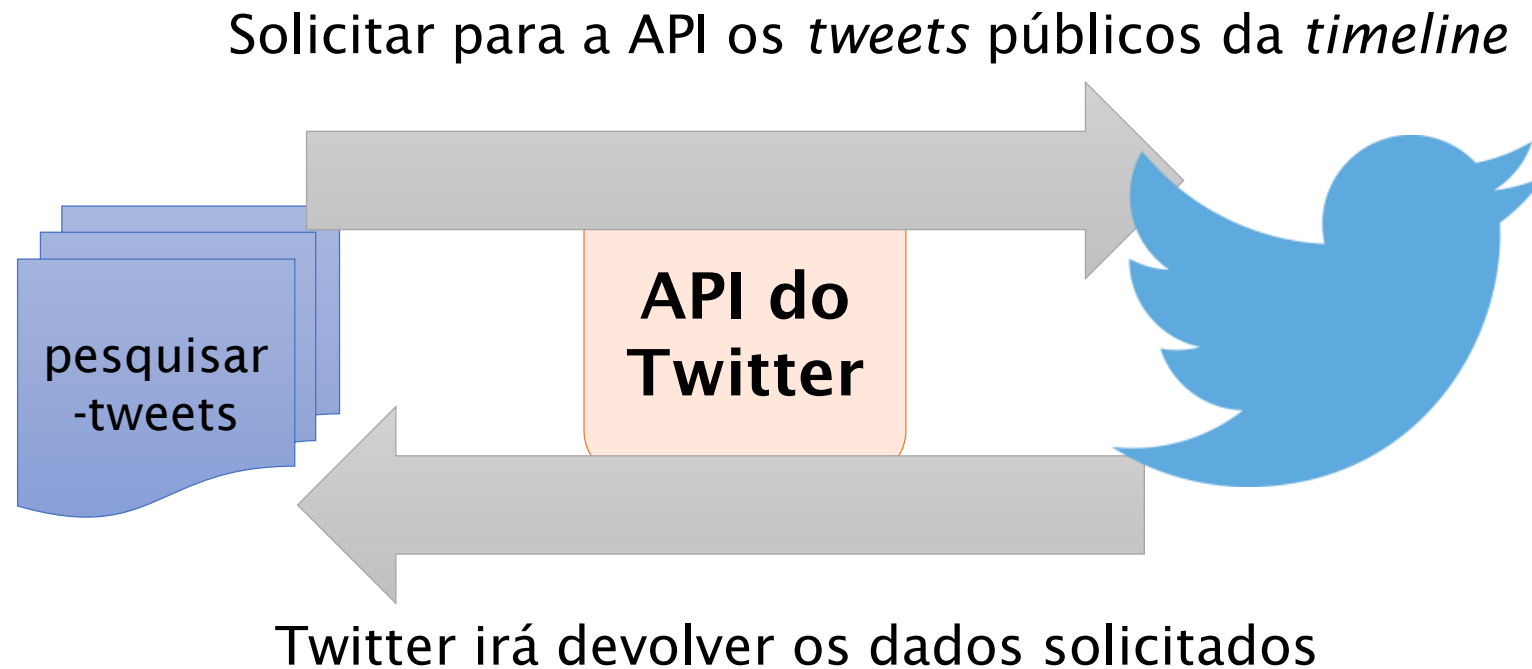
An easy-to-use Python library for accessing the Twitter API.

- Permite atualizar o status
- Permite receber informações do perfil
- Permite monitorar os tweets e realizar alguma ação quando um evento acontecer.
- Permite receber os tweets públicos.
- Seguir todos os usuários que seguirem o usuário autenticado.
- E muito mais!

<https://tweepy.readthedocs.io/en/latest/>



Recuperar Tweets



Recuperar Tweets

- Para criar esse programa, temos que:
 - Importar o módulo
 - Definir as chaves e tokens de acesso
 - Criar a autorização
 - Definir o acesso
 - Criar a API com base no acesso
 - Utilizar o método **search** para procurar por tweets

 Abra o arquivo "Aula 05 - Parte 1 - Conector com API Twitter.ipynb"



2. Técnicas para Text Mining



Técnicas para Text Mining

- Retirar pontuação
 - [, - ? ! \$ @ ; :]
- Tokenizar

Estou
aprendendo
processamento
de
linguagem
natural
,
curso
em
Maceió

```
['Estou',  
'aprendendo',  
'processamento',  
'de',  
'linguagem',  
'natural,',  
'curso',  
'em',  
'Maceió']
```



Técnicas para Text Mining

- Stopwords
 - Na computação, uma palavra vazia (ou stop word, em inglês) é uma palavra que é removida antes ou após o processamento de um texto em linguagem natural.
- Texto em caixa alta ou baixa
 - OI COMO VAI VOCÊ?
 - Oi como vai o Helio?



Técnicas para Text Mining

- Lematização
 - A lematização é o processo, efetivamente, de deflexionar uma palavra para determinar o seu lema (as flexões chamam-se lexemas)
 - Exemplo: Cantando, Cantado, Cantarolando = Cantar



Técnicas para Text Mining

- Stemização
 - Em morfologia linguística e recuperação de informação a stemização (do inglês, stemming) é o processo de reduzir palavras flexionadas (ou às vezes derivadas) ao seu tronco (stem), base ou raiz, geralmente uma forma da palavra escrita.
 - Exemplo: **Cant**ando, **Cant**ado, **Cant**arolando = Canta



Técnicas para Text Mining

- Marcação POS (part-of-speech)
 - Tags Part-of-Speech (PoS) - Conjunto de etiquetas que assinalam a relação que uma palavra tem com as que a envolvem. Geralmente considerada como sendo uma gramática.
 - POS (part-of-speech) atribui para as palavras partes da fala, como substantivos, adjetivos, verbos
 - Importante para a detecção de entidades no texto, pois primeiro é necessário saber o que o texto contém
 - Lista de tokens: <https://spacy.io/api/annotation#pos-tagging>
 - Português: <https://www.sketchengine.eu/portuguese-freeling-part-of-speech-tagset/>



Técnicas para Text Mining

- Marcação POS (part-of-speech)

```
#verificando o documento  
documento = pln('Estou estando processamento de linguagem natural, curso da FIA em São Paulo')
```

```
Estou AUX  
estando VERB  
processamento NOUN  
de ADP  
linguagem NOUN  
natural ADJ  
, PUNCT  
curso NOUN  
da DET  
FIA PROPN  
em ADP  
São PROPN  
Paulo PROPN
```



Técnicas para Text Mining

- Entidades no Texto
 - O reconhecimento de entidades mencionadas (REM) é um ramo do processamento de linguagem natural que procura extrair e classificar as entidades mencionadas em um texto escrito em linguagem natural.
 - NER (Named-Entity Recognition)
 - Encontrar e classificar entidades no texto, dependendo da base de dados que foi utilizada para o treinamento (pessoa, localização, empresa, numéricos)
 - Usado em chatbots para saber o assunto falado
 - Siglas: <https://spacy.io/api/annotation#named-entities>
 - <https://spacy.io/models/pt>



Técnicas para Text Mining

- Entidades no Texto

A **IBM** **ORG** é uma empresa dos **Estados Unidos** **LOC** voltada para a área de informática. Sua sede no **Brasil** **LOC** fica em **São Paulo** **LOC** e a receita em 2018 foi de aproximadamente 320 bilhões de reais

Bill Gates **PER** nasceu em **Seattle** **LOC** em 28/10/1955 e foi o criador da **Microsoft** **ORG**



Técnicas para Text Mining

Bag of Words

- É um processo que transforma um conjunto de texto em números.
- Basicamente ele cria uma coluna para cada token, que indica a presença do token na frase (1) ou a ausência (0). Procedimento é similar ao One Hot Encoder.

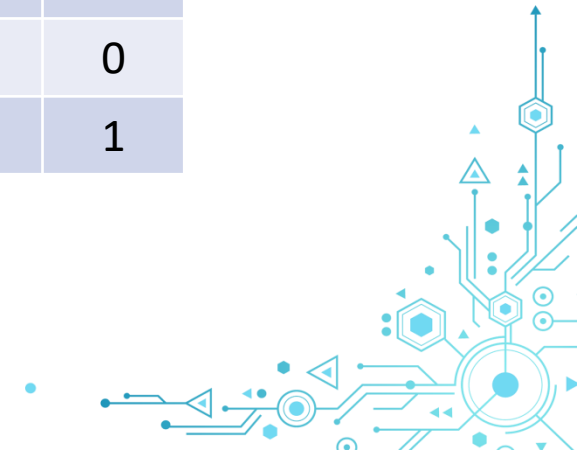


Técnicas para Text Mining

Bag of Words

Documento	Texto
01	Eu quero ir para casa
02	Você gosta de jantar
03	Maria quer Jantar

Documento	Eu	quero	ir	para	casa	Você	gosta	De	jantar	Maria	quer
01	1	1	1	1	1	0	0	0	0	0	0
02	0	0	0	0	0	1	1	1	1	0	0
03	0	0	0	0	0	0	0	0	1	1	1



3. spaCy



spaCy



- Construída por Matthew Honnibal em 2015
- NLP avançado escrita em Python e Cython
- Considerado o parser sintático mais rápido do mundo
- Possui internamente modelos de redes neurais em vários idiomas
- Uso em larga escala (auxiliado com o Cython)
- Processamento de textos para deep learning
- Link: <https://spacy.io/>



spaCy

- Quem utiliza?



Uber

Quora

Retriever

STITCH FIX



AND MANY MORE

FEATURED ON

recode

The
Washington
Post

BBC

Microsoft

VentureBeat

ThoughtWorks





Vamos entender como utilizar o spaCy e como ele nos ajuda a implementar vários conceitos de processamento de dados textuais.

 Abra o arquivo "Aula 05 - Parte 2 - Introdução ao spaCy.ipynb"



Conteúdo da Aula



1. Análise Exploratória de Texto



2. Análise de Sentimentos



1. Análise Exploratória de Texto

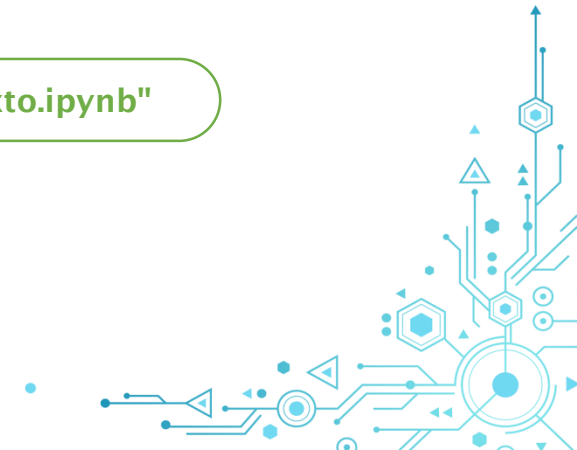


Análise Exploratória de Texto

- Vamos aplicar o nosso conhecimento em uma base extraída do Twitter.
- Importante lembrar que a base que iremos utilizar foi manualmente classificada o sentimento.



Abra o arquivo " Aula 06 - Parte 1 - Análise e Visualização de Texto.ipynb"



2. Análise de Sentimentos



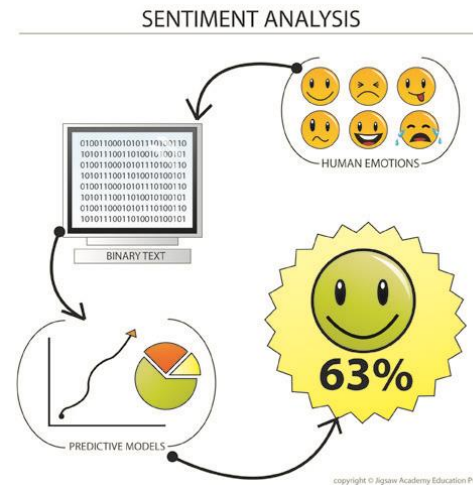
Análise de Sentimentos

- Análise de sentimentos é a atribuição de identificar, extrair, quantificar e estudar estados afetivos e informações subjetivas.
- Realizada por meio de processamento de linguagem natural, análise de texto, linguística computacional e biometria e tem a finalidade de criar conhecimento a partir destes dados.



Análise de Sentimentos

- Pode ser empregada de diversas formas: em aplicações de voz do cliente, por análise de respostas de um questionário e em comentários de avaliações.



Análise de Sentimentos

- Vamos aplicar o nosso conhecimento em uma base extraída do Twitter.
- Importante lembrar que a base que iremos utilizar foi manualmente classificada o sentimento.

 Abra o arquivo " Aula 06 - Parte 2 - Classificação de Textos com Cy.ipynb"



Conteúdo da Aula

-
-
- 1. Hands-on de análise de sentimentos com Twitter
- 2. Hands-on utilizando o Pycaret
-
-





Abra o arquivo "Aula 07 - Parte 01 - Classificação de textos do Twitter com
y.ipynb"



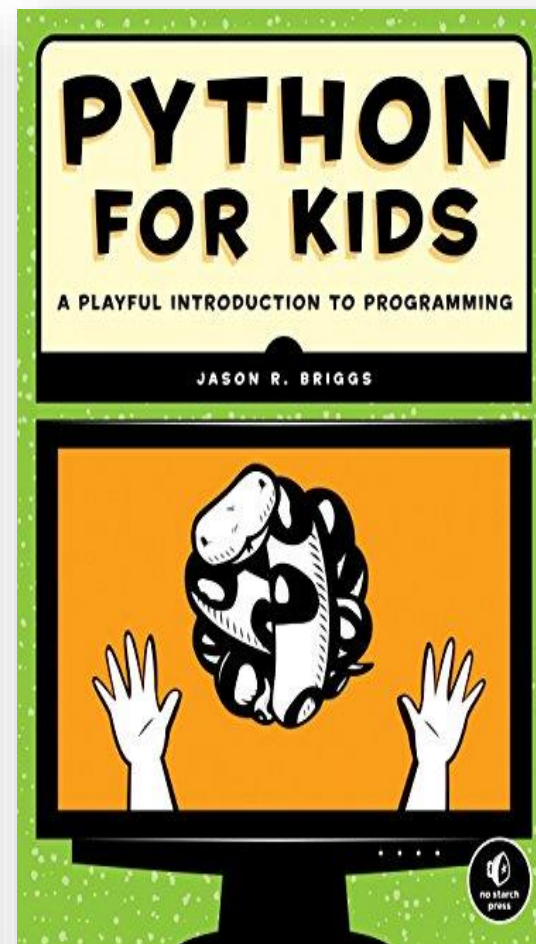
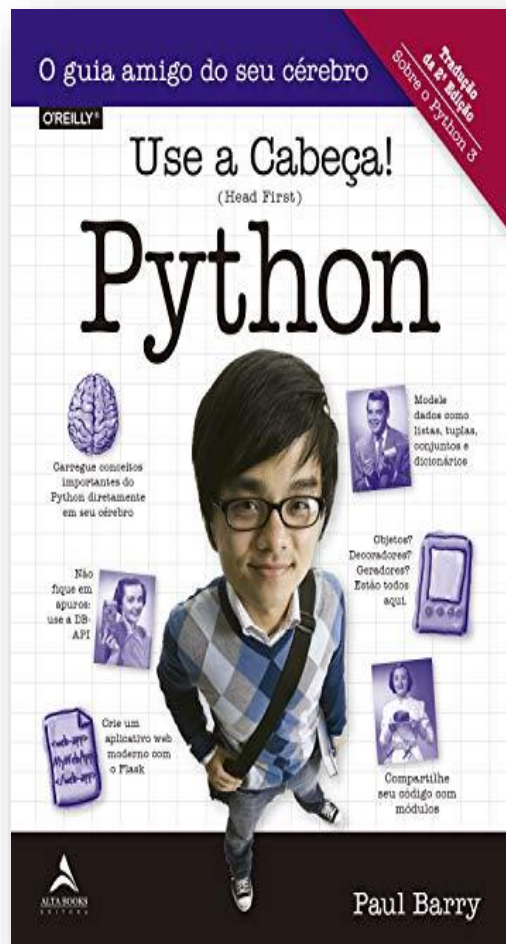
Abra o arquivo " Aula 07 – Parte 2 - Classificação de textos do Twitter com spaCy, CountVectorizer e
et.ipynb"



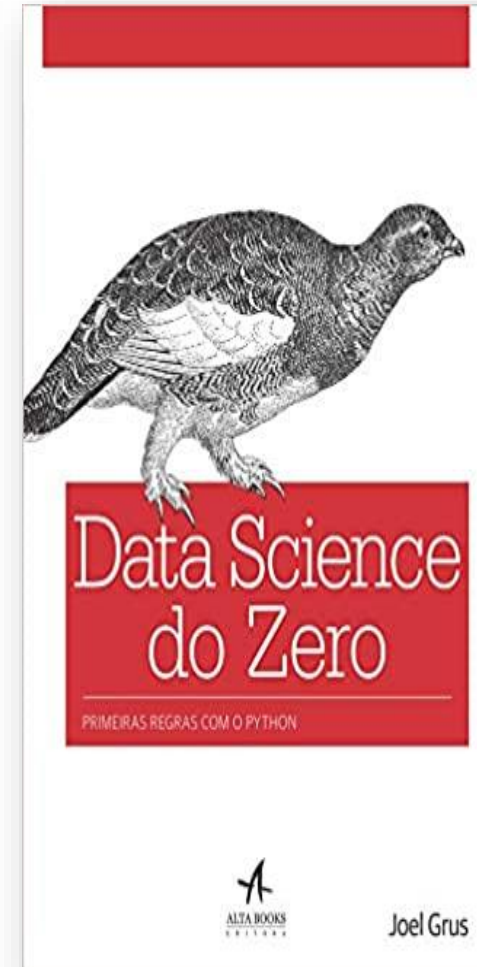
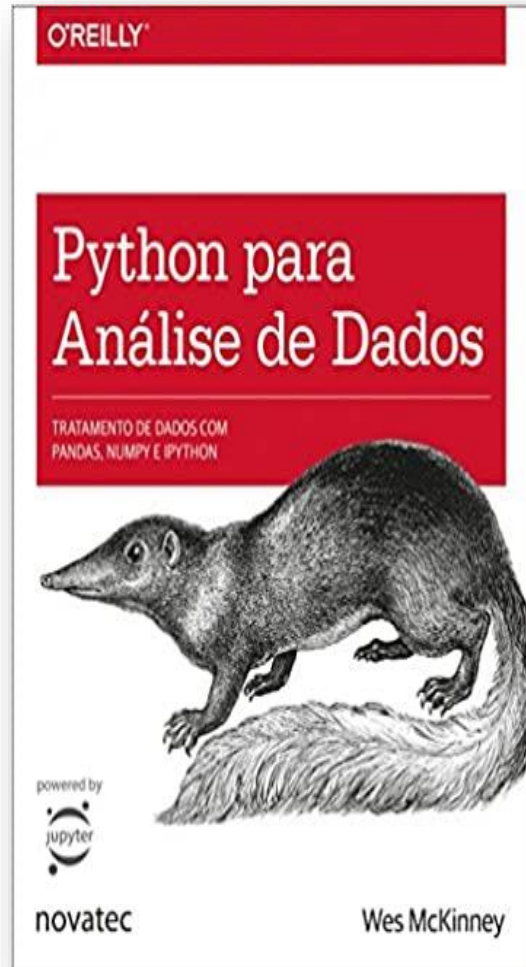
Referências Bibliográficas



Referências Bibliográficas



Referências Bibliográficas



Referências Bibliográficas

