

Materiais :

Q: Onde eu encontro os códigos-fonte das aulas ministradas na disciplina?

R: Os materiais de todas as aulas da disciplina podem ser acessados por meio do ícone de folha a4 que fica ao lado do nome da disciplina no menu à esquerda.

O professor Renan fez um grande apanhado da área de Ciência de Dados mesclando teoria e prática. Durante as aulas foi utilizada a linguagem para análise de dados R e os códigos estão disponíveis também no menu a4 que fica ao lado do nome da disciplina no menu à esquerda, "funções auxiliares". Cabe lembrar que os arquivos CSV utilizados pelo professor Renan devem ser baixados na página do *Kaggle* disponível na página 39 do PDF do professor.

Q: Professor, sobre o Livro: *The Elements of Statistical Learning*, temos ele em PDF? Onde é possível acessar? Tem na versão em português ou só em inglês?

R: Esse livro é possível achar em sua versão para o idioma inglês, pelo site da Stanford (pago).

Q: O Link da máquina de seleções de tomate é privativo.

R: Este vídeo mostra a mesma ideia <https://www.youtube.com/watch?v=aYQ\ 5c6m8Is>

Q: Não estou conseguindo acesso ao papaer *Advances in natural language processing* pois parece estar bloqueado exigindo algum tipo de subscrição.

R: Caso não consiga acesso via ícone de folha a4 da disciplina no Sala Virtual, o artigo pode ser acessado via GitHub.

Ferramentas :

Q: A linguagem R será abordada com maior ênfase durante a continuação do curso?

R: Durante o curso, há uma disciplina chamada Estatística para Ciência de Dados onde a linguagem será explorada com maiores detalhes. É uma disciplina voltada para análise estatística dos dados e exploração de conceitos estatísticos.

O curso tem maior ênfase na utilização da linguagem Python, conforme cronograma estabelecido pelo curso de Especialização.

Q: Sou novo com a linguagem de programação Python, por onde posso começar?

R: Hoje em dia há muitos cursos sobre Python disponíveis na web, existem opções gratuitas quanto pagos e das mais diversas complexidades. No Youtube existem diversas videoaulas de demonstração, assim como a *Udemy* possui alguns cursos bem acessíveis. Durante o curso teremos uma disciplina muito interessante, apenas para processamento de dados usando Python com o professor Juliano. Também teremos uma disciplina de Python para Ciência de Dados com o professor Mangan.

Q: Durante a aula do professor Renan, foram apresentadas no slide 11 ferramentas de manipulação e extração de dados, quais dessas são ideais para fazer isso sem programação em Python?

R: Nessas condições de, não ser via programação em Python talvez o slide não apresente nenhuma. A não ser que estes dados estejam em uma base de dados já estruturada, assim,

é possível utilizar o SQL, Presto, *RStudio* combinado com expressões regulares etc. Segue alguns outros exemplos de ferramentas, algumas gratuitas e outras pagas:

1 - *Knime Analytics*

Essa é uma plataforma que ajuda a manipular, analisar e modelar dados por meio de programação visual. Essa ferramenta conta com vários módulos e possui uma ampla variedade de ferramentas integradas. Essa plataforma é *Open Source* e é semestralmente atualizada.

2 - *Orange*

É uma plataforma de análise em que os usuários podem extrair dados via programação visual ou scripts Python em uma janela de terminal.

Também conta com componentes para *Machine Learning* e funcionalidades como mineração de dados de fontes externas para execução de processamento de linguagem natural, mineração de texto, bioinformática e análise de rede.

3 - [Tableau Public](#)

Ela permite analisar e visualizar dados, possibilitando que os usuários efetuem a publicação de dados interativos na web.

Esse software também faz a extração de dados de outras ferramentas, como, por exemplo:

- Planilhas Google,
- Microsoft Excel,
- arquivos CSV,
- arquivos JSON,
- arquivos estatísticos,
- arquivos espaciais,
- conectores de dados da Web e OData.

Conceitos e Exercícios:

Q: O professor Michael algumas vezes mencionou os termos "*conhecimento intencional e conhecimento extensional*". Qual a diferença entre os dois? Poderia dar alguns exemplos?

R: Como diz o professor Michael próximo aos minutos 28 e 29, quando se tem uma grande base de dados e a partir desses, pegamos várias instâncias desses dados como EXEMPLOS, estamos utilizando um conhecimento extensional, pois a partir de dados que já sabemos e temos experiência, geramos o conhecimento.

No conhecimento extensional, nós geramos o conhecimento a partir de uma longa base de dados. Como por exemplo o exercício do professor Renan, a partir dos dados sobre prazo de entrega, nós podemos gerar conhecimento sobre novos dados e inferir que quando a entrega está atrasada ela pode gerar uma avaliação ruim.

Já a respeito do conhecimento intencional, nós podemos pré-definir comportamentos ou regras para gerar os sistemas. Um exemplo claro sobre o conhecimento intencional são sistemas de Processamento de linguagem natural, onde podemos usar regras do nosso cotidiano, para inferir algo sobre as entidades.

Por exemplo, sabemos que uma pessoa que é filho de alguém, ela deve ter um pai e uma mãe, logo essa é uma regra e assim podemos identificar esses papéis em uma sentença.

Q: Na ambiguidade sintática, há frases que podem ser um problema de falta de informação, pois nem um humano conseguiria entender, o agente precisaria fazer uma pergunta caso o contexto não resolva certo? Como que o agente formularia essa pergunta para completar a informação?

R: A ambiguidade é inerente a língua. As duas interpretações são cabíveis do ponto de vista sintático. Só com a sentença, o agente é incapaz de resolver. Pode usar uma heurística simples como escolher sempre uma das interpretações, o que é passível de erro. Também é possível construir um modelo capaz de analisar o contexto e assim, desambiguar a tarefa. Nesse último caso, a solução tem maior complexidade, pois pode envolver semântica.

Q: Em relação ao *Cloud Computing* e à infraestrutura para Big Data:

R: O conceito central do *Hadoop*, plataforma para Big Data, é a localidade dos dados visto que a partir deste conceito, é possível ter um nível maior de processamento. O processamento, de fato, é importante no *Hadoop* (*Map-Reduce* como mecanismo, concorrência/paralelismo na manipulação dos dados), mas mesmo é consequência da localidade.

Q: Qual profissional é o responsável por gerenciar o HDFS?

R: O HDFS é uma parte do cluster, a sua disponibilidade, configuração e demais questões ficariam em teoria, sob responsabilidade de um time de infra (Devops) ou do próprio time de engenharia de dados.

Quanto à ingestão\gravação de dados no HDFS\Data Lake e afins, em empresas maiores, o setor de engenharia de dados seria o responsável. Mais para frente durante o curso, terá uma ótima disciplina de Gerência de Infraestrutura para Big Data com o professor Tiago Ferreto que vai abordar alguns desses assuntos como HDFS, Hive, Spark entre outros.

Q: O que são símbolos funcionais?

R: Na lógica de predicados (ou de primeira ordem), os predicados são aplicados a 0 ou mais termos. Por exemplo, *e_alto(joao)* se aplica a um termo (um símbolo representando uma pessoa, o João), *e_mais_alto_que(joao, maria)* se aplica a dois termos.

Um termo representa um objeto (um elemento qualquer) do universo de discurso. Termos podem ser constantes (*joao, maria*), variáveis ou funtores (símbolos funcionais aplicados a seus elementos, outros termos).

Símbolos funcionais representam funções universo de discurso. Por exemplo, *idade_de(joao) = 20*. *idade_de* é um símbolo funcional

Q: Não entendi no exemplo do pneu furado a definição dos efeitos da Ação (*LeaveOverNight*). Por que os pneus deixam de existir?

R: Nesse contexto, isolando o problema do pneu furado, a Ação *LeaveOverNight* é não fazer exatamente nada. O efeito dela é que nenhum pneu se modifica, o reserva permanece no porta-malas e o furado no eixo.

Q: Lendo o artigo recomendado, *Data Science: Challenge and Directions*, quais os conceitos de *IID data problem* e *non-IID data Problem* ?

R: Sobre o conceito de IID, ele é mais um conceito de estatística:

Na teoria de probabilidade e estatística, uma sequência de variáveis aleatórias é independente e identicamente distribuída (IID) se cada variável aleatória tem a mesma distribuição de probabilidade que as outras e são todas mutuamente independentes.

Frequentemente, o pressuposto IID surge no contexto das sequências de variáveis aleatórias. Então, "independentes e identicamente distribuídas" em parte implica que um elemento na sequência é independente das variáveis aleatórias que vieram antes dele.

Exemplos:

Uma sequência de valores observados de giros de uma roleta viciada ou não viciada é IID. Uma implicação disto é que se a bola cai no vermelho, por exemplo, 20 vezes seguidas, não é mais, nem menos provável que a bola caia no preto no próximo giro do que em qualquer outro.

Uma sequência de lances de dados viciados ou não viciados é IID.

Uma sequência de cara ou coroa com uma moeda viciada ou não viciada é IID.

Uma sequência IID é diferente de uma sequência de Markov, em que a distribuição de probabilidade para a n -ésima variável aleatória é uma função da variável aleatória anterior na sequência (para uma sequência de Markov de primeira ordem).

Q: Na seguinte questão:

Considerando que o problema seja gerar um resumo de um conjunto de notícias de jornal publicadas, durante a semana, sobre o mesmo assunto.

Sabendo também que o *dataset* disponibilizado para treinamento contém o texto, a data, o título e a seção da notícia, mas não indica o assunto.

Sabendo ainda que o sistema, quando em uso, receberá apenas o título, a data e o texto da notícia, a solução mais indicada é escolher algoritmos de *Machine Learning* capazes de executar as tarefas de:

A resposta indicada como correta foi a seguinte:

Classificação para aprender a categorizar os textos de acordo com as seções,

Agrupamento para organizar os textos por assunto e Sumarização para gerar os resumos dos textos existentes em cada grupo.

Porém, no material de aula, a utilização de agrupamento está associada a dados não classificados. Seria possível, e correto, a utilização de agrupamento após classificação dos dados?

R: Sim é possível e não deixa de ser correto a utilização de 2 ou mais modelos sequenciais, as chamadas abordagens em pipelines. Muitas vezes você irá se deparar com cenários que apenas um modelo não irá ser o suficiente para resolver o seu problema, tendo que executar um modelo após o outro.

Sobre a questão em específico:

É importante ressaltar que a forma mais indicada para solução é sempre a mais assertiva.

Desprezar informações e dados significativos disponíveis não são bons caminhos quando se trata de algoritmos de *Machine Learning*. Abordagens supervisionadas geram resultados melhores pois dispõem de mais informação: o rótulo. Sendo assim, desprezar a informação quanto à seção é um erro.

Logo, agrupamento não deve ser a primeira tarefa a ser executada. O mais correto é aproveitar as seções para aprender a classificar as notícias. Como a informação sobre o assunto não está disponível e nem sempre o título é informativo o suficiente e nem define classes, a tarefa seguinte deve ser agrupamento para gerar grupos com as notícias mais parecidas.

E, por fim, aplicar algoritmos de sumarização nos grupos gerados, filtrados por data.

**FAQ gerado com base em comentários até o dia 02/03/2022.*