DISCIPLINA .....: INTRODUÇÃO A BIG DATA COM APLICAÇÕES EM RELAÇÕES INTERNACIONAIS 1º BIMESTRE CGR
DEPARTAMENTO.....: CPDOC
CURSO.....: CGAP – FGV-EAESP
PROFESSOR .....: UMBERTO MIGNOZZETTI
TIPO DE DISCIPLINA .: ( ) Comum AE/AP ( ) AE (X) AP | créditos: (X) 2 ( ) 4

SEMESTRE/ANO: 1º/2017

OBJETIVOS DA DISCIPLINA

Neste curso vamos estudar as principais técnicas de manipulação e análise, aplicando nossos conhecimentos em problemas de relações internacionais. Focaremos em coleta automatizada de dados, processamento e manipulação de bases grandes, e estimação de modelos simples de machine learning, utilizando o pacote estatistico R. Aplicaremos o conhecimento em classificação de regimes em democráticos, previsão de estabilidade de alianças políticas, análise de padrões de comércio internacional, entre outros.

OBJETIVOS DE APRENDIZAGEM

1. Introduzir o aluno ao pacote estatístico R
2. Estudar processamento, coleta automatizada, e organização de grande volume de dados em R
3. Estimação de modelos preditivos simples, e ferramentas de Machine Learning acessíveis
4. Aplicações em exemplos reais de Relações Internacionais e Ciência Política, ensinando ao aluno um pouco dos temas mais debatidos na disciplina.

CONTEÚDO

Parte I – Introdução ao R

Aula 1. Introdução ao processamento de dados estatísticos em R

Aula 2. Programação e métodos avançados de processamento de dados em R

Parte II – Processamento de Dados

Aula 3. Procesamento de dados em R. O conceito de *tidy data*.

Aula 4. Coleta automatizada de dados e organização de bases de dados.

Aula 5. Reorganizando e estruturando bases de dados para análise empírica.

Parte III – Aprendendo com os Dados

Aula 6. Análise descritiva e apresentação de dados.

Aula 7. Modelos básicos de Machine Learning.

Aula 8. Processamento básico de textos + AVALIAÇÃO FINAL

CRITÉRIO DE AVALIAÇÃO

Presença em aulas (30%)

Exercícios em classe ao final das aulas (50%)

Avaliação Final (20%)

BIBLIOGRAFIA

Textos principais:

- Big Data Analytics with R (Amazon link)
- Advanced R (Amazon link)
- R for Data Science (Amazon link)

Outros textos podem ser sugeridos no decorrer do curso.

COMPROMISSO ÉTICO – PROFESSOR/ALUNO

Os alunos podem colaborar nos exercícios, mas cada aluno deverá entregar seu trabalho individualmente, por computador. Na avaliação final os alunos não podem colaborar, e a avaliação é sem consulta. Para um bom aprendizado da matéria é recomendado que cada aluno revise o conteúdo da aula anterior em casa, para fixar os conceitos, e memorizar os procedimentos e rotinas empregados no uso do software.

CONTATO E OFFICE HOURS

Professor	**Contato**	**Horário de atendimento**
Umberto Mignozzetti	umberto.mignozzetti@fgv.br	Quarta-Feira, das 14:00 às 16:00h

PROGRAMAÇÃO AULA-A-AULA

AULA 1: Introdução ao R

Nessa aula vamos discutir um pouco sobre Ciência de Dados, Big Data, Softwares utilizados nas análises, e vamos aprender o software que vamos utilizar, o R. O R é um software livre, que pode ser baixado de qualquer computador, XXXX. Vamos focar em carregar, abrir, manipular, e analisar, de modo simplificado, bases de dados no software. Aplicaremos nosso conhecimento analisando a base de dados do Quality of Government, que disponibiliza mais de 1000 variáveis sobre diversos países do mundo.

AULA 2: Programação e Métodos de Processamento em R

Essa aula aprofunda o conhecimento da aula anterior, estudando como manipular dados já existentes, criando novas informações, mais amenas à análise empírica. Vamos também discutir um pouco de conceitos básicos de ciência da computação e metodos de programação, voltados para o R. Aplicaremos os concentos estudados no Censo de 2010.

AULA 3: Processamento de dados

Vamos introduzir o conceito de tidy data, e discutir como chegar dos bancos de dados que temos, aos bancos de dados que queremos usar em nossa análise. Introduziremos o pacote de processamento **dplyr**, que é usado na manipulação de dados. Praticaremos processamento usando as bases do WITS de comércio internacional.

AULA 4: Coleta automatizada de dados

Na maior parte do tempo, tudo que precisamos está em algum lugar na internet. Saber coletar esses dados de modo rápido e eficaz é essencial para produzirmos resultados rápidos e confiáveis. Nessa aula vamos focar em baixar dados da internet e salva-los no computador. Vamos aplicar os conhecimentos baixando dados de votação e discursos na Assembléia Geral da Organização das Nações Unidas.

AULA 5: Reorganizando e estruturando as bases de dados.

Ao baixar os dados, retornamos ao problema principal da organização de dados, que é coletar dados em um formato impróprio para análise, em um banco de dados que pode ser usado para aprendermos coisas. Vamos processar os dados baixados na aula anterior, montando uma matriz de votações dos países. Nessa aula discutiremos alguns softwares que potencializam o trabalho com bases grandes, como Hadoop, MapReduce, e Spark.

AULA 6: Análise Descritiva

Vamos voltar às bases anteriores discutindo métodos eficientes de apresentar os dados e extrairmos informações razoáveis dos mesmos. Vamos discutir maneiras de montar gráficos e estatísticas descritivas que maximizem o entendimento do conteúdo e extensão dos dados.

AULA 7: Machine Learning

Vamos introduzir três técnicas de machine learning: análise de regressão, supporting vector machines (usados na classificação de emais/spams), e análise de clusters. Essas técnicas, quando bem aplicadas, auxiliam no entendimento das correlações e padrões existentes nos dados. Como bonus vamos fazer uma análise do posicionamendo ideológico dos Deputados Brasileiros.

AULA 8: Processamento Básico de Textos + Avaliação

Nessa aula vamos retomar os discursos coletados na Assembléia Geral da ONU, e vamos fazer uma análise básica de seu conteúdo, com WordClouds, contagem de palavras, e correlações entre palavras usadas nos discursos.

Na segunda metade da aula vamos ter Avaliação. A avaliação consistirá em um trabalho aplicado, que o aluno terá duas horas para executar. O trabalho irá requerer baixar conteúdos da internet, organiza-los, processar e apresentar resultados de aprendizagem nesse curto período.