Faculdade IMPACTA TECNOLOGIA

Pós Graduação em BI com Big Data

Web Mining

Introdução à Web mining: Conceitos e Tecnologias

Prof. MSc. Fernando Sousa

Bibliografia

- Bibliografia básica para esta aula
  - Weiss SM, Indurkhya N, Zhang T. Fundamentals of Predictive Text Mining. 2010 edition. London ; New York: Springer; 2010. 226 p. - Cap. 1
  - Liu B. Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data. 1st ed. 2007. Corr. 2nd printing edition. Berlin ; New York: Springer; 2009. 552 p – Cap 6.

Bibliografia

- Bibliografia complementar
  - Manning CD, Raghavan P, Schütze H. Introduction to Information Retrieval. 1 edition. New York: Cambridge University Press; 2008. 506 p.
  - Salton G, McGill MJ. Introduction to Modern Information Retrieval. New York, NY, USA: McGraw-Hill, Inc.; 1986.
  - Gabardo AC. Análise de Redes Sociais: Uma Visão Computacional. Novatec Editora; 2015. 24 p.

Bibliografia

- Bibliografia complementar
  - Sousa FS. Análise Comparativa de Métodos de Recuperação de Informação para Categorização de Conteúdos Web Relacionados à Saúde [Dissertação de Mestrado]. [São Paulo]: Universidade Federal de São Paulo (UNIFESP); 2011.
  - Araujo, GD. Análise de sentimento de mensagens do Twitter em português brasileiro relacionadas a temas de saúde [Dissertação de Mestrado]. [São Paulo]: Universidade Federal de São Paulo (UNIFESP); 2014.

Apresentação

- Prof. MSc. Fernando Sousa
  - Graduado em Informática Biomédica – USP
  - Mestre em Ciências, Área de Gestão e Informática em Saúde - UNIFESP
  - Analista de Dados
- Contato:
  - professor.fsousa@gmail.com

Referências Web

- Intro to Computer Science & Programming Course [Internet]. Disponível em: https://www.udacity.com/course/cs101
- Home - RapidMiner Documentation [Internet]. [cited 2015 Aug 15]. Disponível em: http://docs.rapidminer.com/

Motivação

Fonte: https://smbp.uwaterloo.ca/2015/03/lots-of-data-little-time-what-social-media-metrics-really-matter/

Motivação

- Desde que foi criada, a Web é utilizada para armazenar dados, principalmente semiestruturados e não estruturados
- Na última década houve um grande crescimento da quantidade de dados existentes na Internet, principalmente devido a criação de redes sociais
  - Estima-se que em 2013 existiam 672 exabytes de dados acessíveis e 1 Yottabyte de dados armazenados
  - Big Data!

1 Yottabyte == 1.000.000.000.000.000.000.000.000 bytes == 10^{24} bytes

1 Exabyte== 1.000.000.000.000.000.000 bytes == 10^{18} bytes

Motivação

- Usuários, seus dados e comportamento estão na Web
  - Onde estou / fui / vou ?
  - Como estou me sentindo ?
  - Quem está comigo?
  - Quem são meus amigos / família?
  - O que está acontecendo?
  - O que estou pensando?
  - O que comprei / procurei ?
- Dados são variados
  - Fotos, vídeos, texto, notícias, páginas pessoais, áudio...

Motivação

- Ainda existe a possibilidade de relacionar estes dados
  - Web mining é a primeira etapa para trabalhar com Web Semântica
- Enorme potencial para extrair informação e descoberta de conhecimento
  - Estratégias de marketing direcionada
  - Sugestões de locais para visitar
  - Sugerir amigos
  - Direcionar postagens de acordo com o humor

Motivação

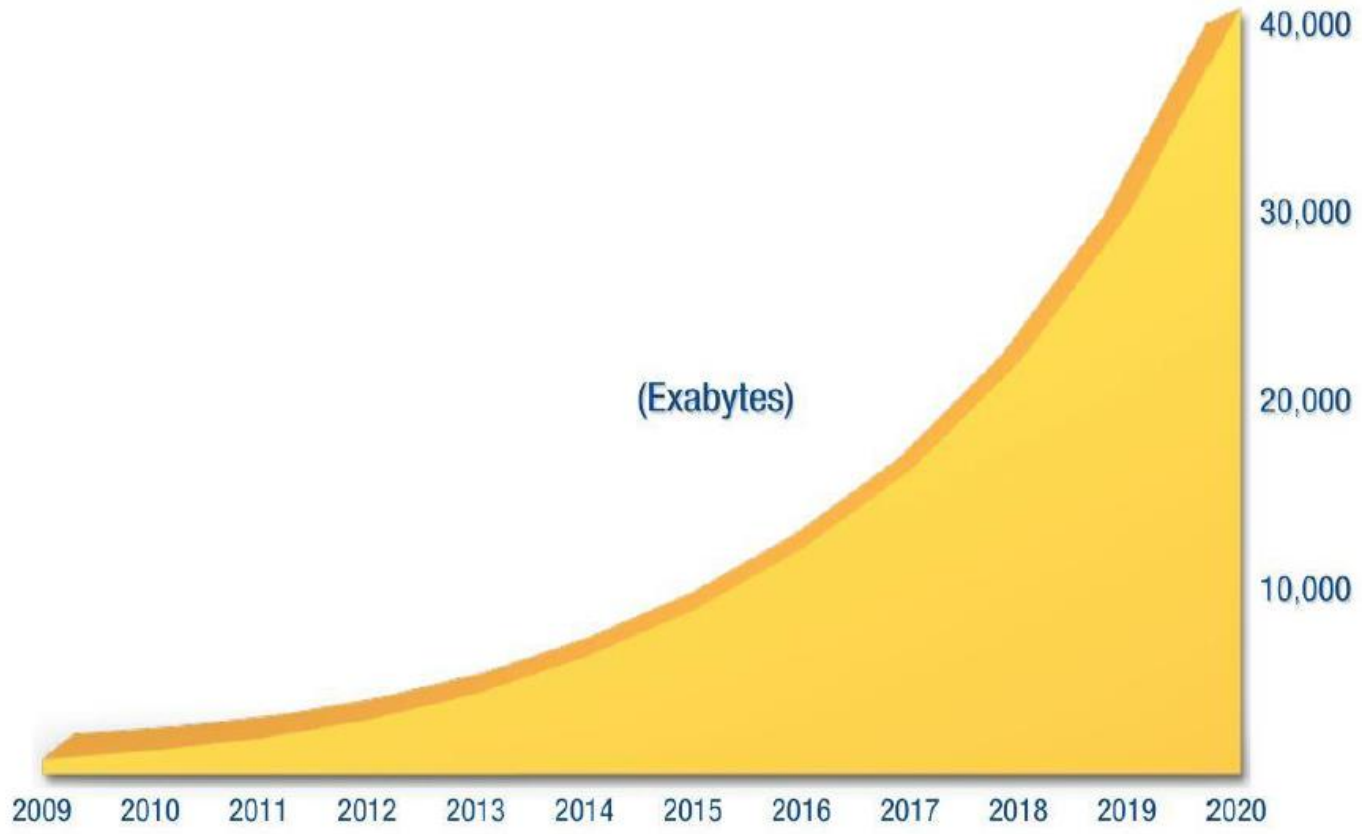
- Um dos problemas é a forma como os dados são apresentados na Web: não estruturados
  - Pessoas comuns criam dados para a web
  - Não existe padronização
- Técnicas de web mining auxiliam no processamento destes dados
  - Encontrar palavras mais frequentes
  - Analisar redes sociais
  - Analisar imagens
  - Buscadores web

Dados digitais

- A maior parte dos dados e transações executadas hoje são digitais
  - Documentos eletrônicos, fotos, vídeos, comércio eletrônico, bate-papo, redes sociais...
- Muitos dados estão disponíveis para análise
  - Big Data

Dados digitais

- Quantidade de dados criados

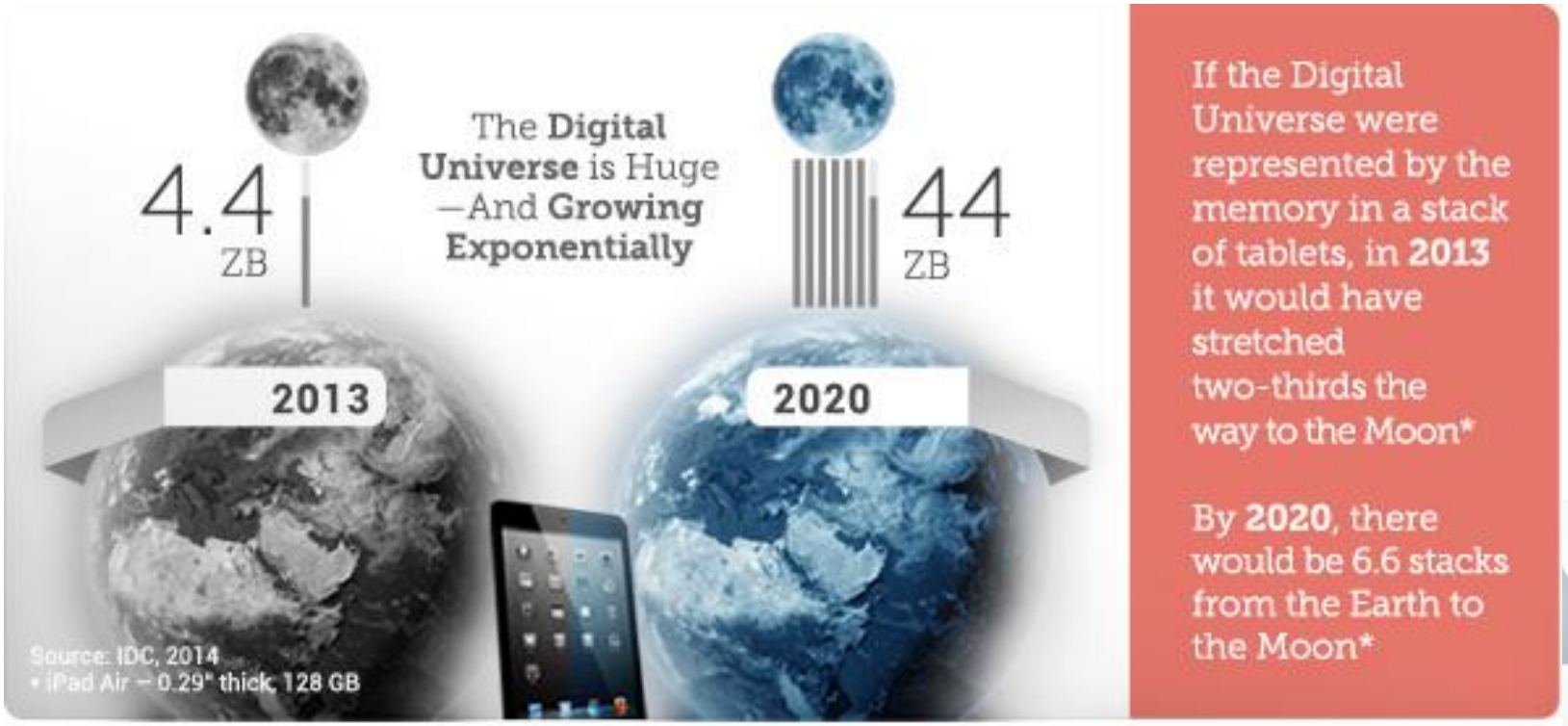


Source: IDC's Digital Universe Study, sponsored by EMC, December 2012

https://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf

Dados digitais

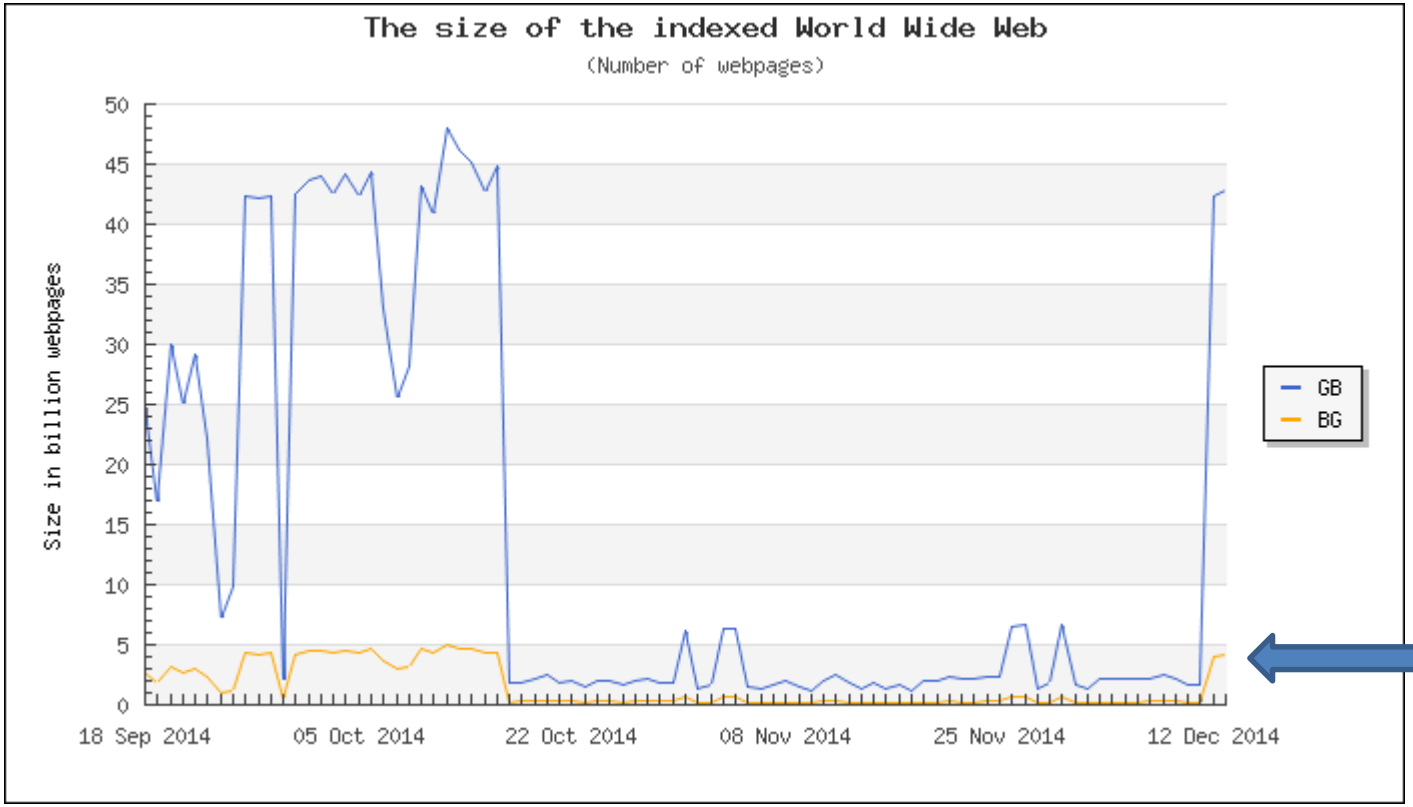
- Quantidade de dados criados



https://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm

Dados digitais

- Web: maior fonte de dados abertos do mundo



4,35 bilhões

http://www.worldwidewebsize.com/ - Acessado em dez/2014

Dados digitais

- De onde vem estes dados?
  - 2/3 (2.9 ZB) é gerado e consumido por usuários da web
    - Redes sociais, TV digital, troca de imagens/vídeos, serviços na nuvem...
  - Destes 2.9 ZB, 85% são administrados por empresas. Ou seja, são dados privados

https://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf
https://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm

Dados digitais

- Mas nem todos os dados são possíveis de analisar
  - Em 2013, apenas 22% dos dados digitais são candidatos para análise
  - Mas apenas 5% disso foi realmente analisados

https://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf
https://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm

Data Mining

- Os dados são analisados através de técnicas de mineração de dados (data mining)
  - Dados estruturados
  - Dados não estruturados
  - Texto
  - Web
- Minerar dados é encontrar padrões com valor agregado nestes dados disponíveis
  - Dados → Informação → Conhecimento
- Métodos de mineração de dados aprendem com dados do passado para fazer predições

Faculdade IMPACTA TECNOLOGIA

Data Mining

- Aplicações de data mining utilizam conjuntos de dados estruturados
  - Pode haver uma preparação dos dados para extrair um conjunto estruturado com dados numéricos e categóricos
- Dados estruturados são basicamente numéricos ou categóricos.
- O conjunto de dados estruturado é representado em forma de tabelas de BD ou planilhas

Data Mining

- Elementos de um Conjunto de dados (Data Set)

Sexo	Batimento cardíaco	Peso	Código da doença
M	175	65	3
F	141	72	1
...	...	...	...
F	161	59	2

Baseado em: Weiss, Sholom M., Nitin Indurkhya, and Tong Zhang. Fundamentals of predictive text mining. Vol. 41. Springer, 2010. p2.

Data Mining

- Elementos de um Conjunto de dados (Data Set)

Exemplo ou registro: um registro de experiência do passado, composto por um conjunto de atributos

Sexo	Batimento cardíaco	Peso	Código da doença
M	175	65	3
F	141	72	1
...	...	...	...
F	161	59	2

Baseado em: Weiss, Sholom M., Nitin Indurkhya, and Tong Zhang. Fundamentals of predictive text mining. Vol. 41. Springer, 2010. p2.

Data Mining

- Elementos de um Conjunto de dados (Data Set)

Atributo: uma característica de um exemplo

Sexo	Batimento cardíaco	Peso	Código da doença
M	175	65	3
F	141	72	1
…	…	…	…
F	161	59	2

Baseado em: Weiss, Sholom M., Nitin Indurkhya, and Tong Zhang. Fundamentals of predictive text mining. Vol. 41. Springer, 2010. p2.

Data Mining

- Elementos de um Conjunto de dados (Data Set)

Classe ou categoria: atributo alvo na predição

Sexo	Batimento cardíaco	Peso	Código da doença
M	175	65	3
F	141	72	1
...	...	...	...
F	161	59	2

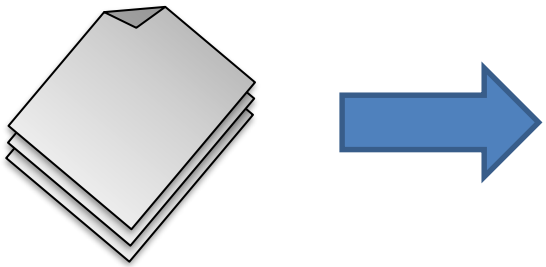
Baseado em: Weiss, Sholom M., Nitin Indurkhya, and Tong Zhang. Fundamentals of predictive text mining. Vol. 41. Springer, 2010. p2.

Data Mining

- A partir destes exemplos, com os atributos e classes, algoritmos de data mining são utilizados para encontrar padrões e criar um modelo para fazer predição ou agrupamentos:
  - Naive Bayes
  - Árvores de decisão
  - KNN
- Estes algoritmos também são utilizados em text mining. Mas antes, os textos devem ser tratados...

Text Mining

- Em princípio, textos são dados não estruturados
  - Textos são conjuntos de documentos não estruturados sem regras para sua criação
- Como transformar um texto não estruturado em um dado estruturado para utilizar os algoritmos de data mining?



Faculdade IMPACTA TECNOLOGIA

Text Mining

- Textos são diferentes de conjuntos estruturados,
  - Dados estruturados são formados por números e categorias
  - Textos são formados por palavras e termos
  - Os atributos/ características dos textos não estão explícitos
  - Estas características devem ser extraídas durante o pré-processamento
- Textos podem conter características complexas, como gramática e significado das palavras
- Entretanto, técnicas simples de extração de características de textos podem ser utilizadas
  - Pré-processamento

Text Mining

- A abordagem mais comum para transformar textos em dados numéricos é a contagem de ocorrência das palavras ou termo do documento

	$\mathbf{Termo_1}$	$\mathbf{Termo_2}$	...	$\mathbf{Termo_n}$	Categoria
$\mathbf{Texto_1}$	$f_{1,1}$	$f_{1,2}$	...	$f_{1,n}$	C_1
$\mathbf{Texto_2}$	$f_{2,1}$	$f_{2,2}$	...	$f_{2,n}$	C_1
...	...	...	...	...	...
$\mathbf{Texto_m}$	$f_{m,1}$	$f_{m,2}$	...	$f_{m,n}$	C_2

- O texto não estruturado agora está representado de forma estruturada, e pode ser utilizado em um método de data mining

Text Mining

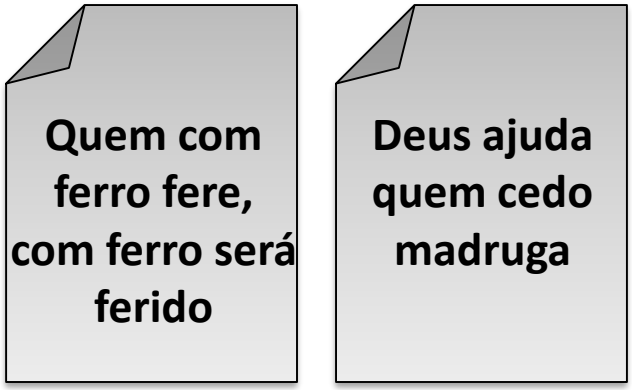
- Por exemplo, no texto abaixo, escrito em um documento digital, pode-se contar quantas vezes cada palavra ou termo aparece no texto, formando um registro de um data set estruturado

Quem com ferro fere, com ferro será ferido

	Quem	com	ferro	fere	será	ferido
$Texto_1$	1	2	2	1	1	1

Text Mining

- Obviamente um uma aplicação real haverá muitos textos para serem analisados. Então ao adicionar mais um texto na base, o conjunto de dados ficaria assim:



	Quem	com	ferro	fere	será	ferido	Deus	ajuda	cedo	madruga
$Texto_1$	1	2	2	1	1	1	0	0	0	0
$Texto_2$	1	0	0	0	0	0	1	1	1	1

Text Mining

- Pode-se fazer uma associação entre os conceitos de data mining e web mining

Data Mining	Text Mining
Atributo	Palavras / Termos
Valores dos atributos	Ocorrência da palavra/ termo
Exemplo/Registro	Documento

Text Mining

- Pode-se fazer uma associação entre os conceitos de data mining e web mining

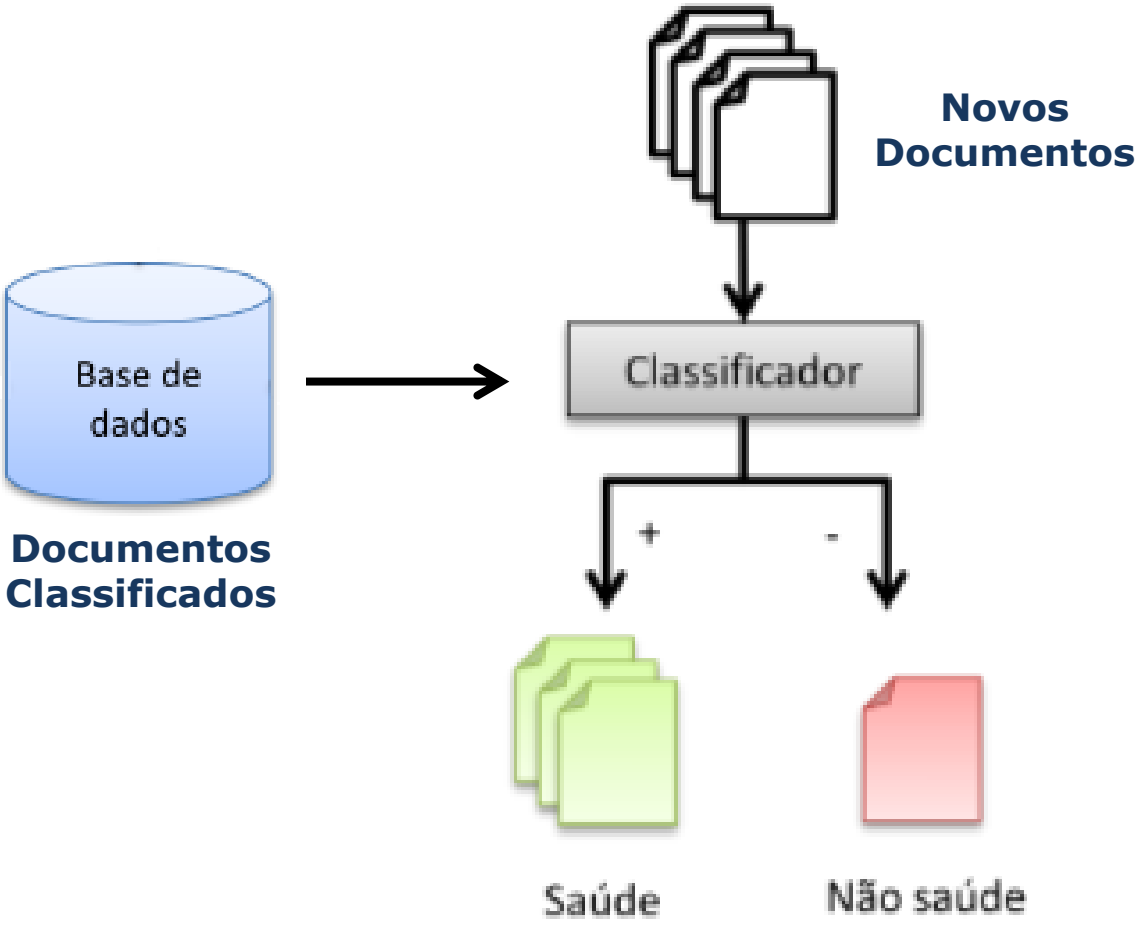
	Sexo	Batimento cardíaco	Peso	Código da doença
$Exemplo_1$	M	175	65	3
$Exemplo_2$	F	141	72	1
$Exemplo_3$	F	161	59	2

	Quem	com	ferro	fere	será	ferido	Deus	ajuda	cedo	madruga
$Texto_1$	1	2	2	1	1	1	0	0	0	0
$Texto_2$	1	0	0	0	0	0	1	1	1	1

Aplicações Text Mining

- Classificação de textos / Supervisionado
  - Existe um conjunto de textos já classificado com documentos classificados como “saúde” e documentos classificados como “não saúde”
  - Dado um novo texto, ainda não classificado, classifique-o como sendo de saúde ou não, baseado nas características do conjunto já classificado
  - Características do conjunto classificado (frequência das palavras, gramática, semântica, etc.) serão utilizadas para classificar um documento desconhecido

Aplicações Text Mining



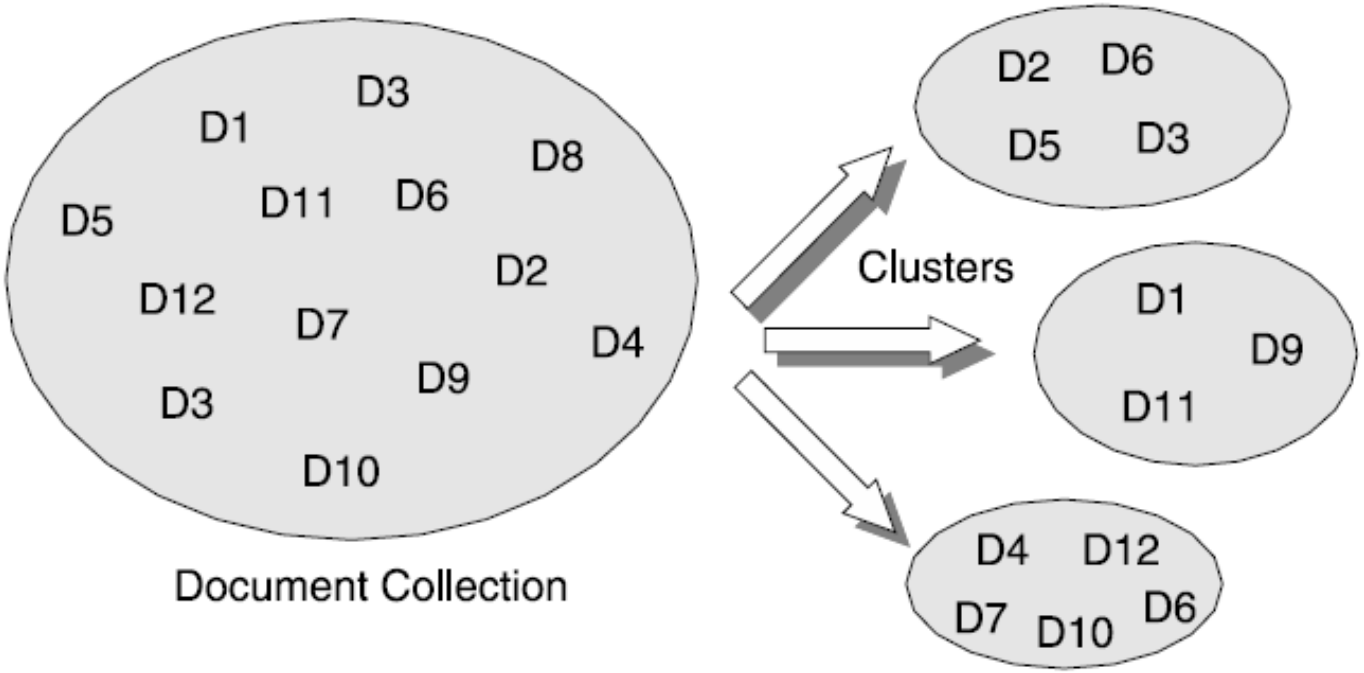
Adaptado de Sousa FS. Análise Comparativa de Métodos de Recuperação de Informação para Categorização de Conteúdos Web Relacionados à Saúde [Dissertação de Mestrado]. [São Paulo]: Universidade Federal de São Paulo (UNIFESP); 2011.

Aplicações Text Mining

- Cluster / Não Supervisionado
  - Existe um conjuntos de textos não classificados
  - A tarefa agora é agrupar os textos em conjuntos de acordo com a similaridade entre eles
    - Tamanho
    - Conteúdo
    - Semântica
    - Palavras e termos

Aplicações Text Mining

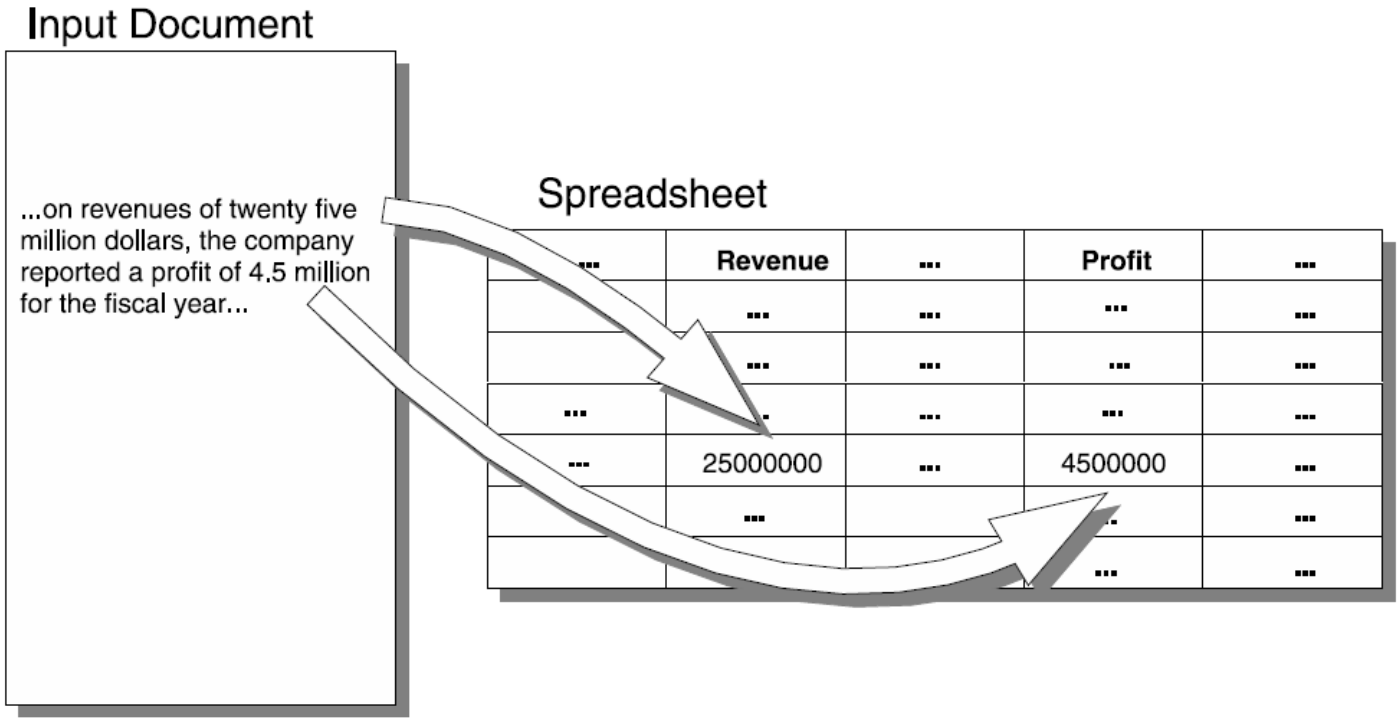
- Cluster



Weiss SM, Indurkhya N, Zhang T. Fundamentals of Predictive Text Mining. 2010 edition. London ; New York: Springer; 2010. 226 p. - Cap. 1

Aplicações Text Mining

- Extração de Informação
  - Extrair dados relevantes de um texto (dado não estruturado) para, por exemplo, popular um conjunto de dados



Weiss SM, Indurkhya N, Zhang T. Fundamentals of Predictive Text Mining. 2010 edition. London ; New York: Springer; 2010. 226 p. - Cap. 1

Web

- A World Wide Web, ou simplesmente Web, foi concebida para que documentos de hipertexto (HTML) fossem armazenados e trocados através da Internet (rede mundial de computadores)
  - Foi criada no final da década de 80 e tornou-se pública e de acesso gratuito em 1993
  - No início, era formada basicamente por documentos de hipertexto (HTML) e outros tipos de documentos, como imagens e vídeos , conectados entre si através de links
  - Os conteúdos eram providos basicamente por pessoas especializadas com conhecimento técnico de TI
  - Posteriormente, este formato ficou conhecido como Web 1.0

Web

- Com o tempo, usuários comuns começaram a atuar como provedores de conteúdo para a web
  - Contribuem para a elaboração do conteúdo existente na Web
  - Contribuem para o crescimento dos dados disponíveis
- Ficou conhecida como Web 2.0
  - Apenas conceitual; não há alteração de padrões ou especificações técnicas
  - Web social e colaborativa
    - Redes sociais
    - Aplicativos móveis

Web

- Com usuários colaborando com o conteúdo, houve um grande crescimento na quantidade de dados disponíveis
- Paralelo a isso, a evolução das técnicas computacionais e de poder de processamento possibilitou que começassem as primeiras aplicações de web mining
  - Buscadores de conteúdo na web
    - Google: 1996
- Começou-se a criar uma fonte rica de dados publicamente disponíveis
  - Que informações importantes pode-se extrair destes dados?
  - Qual o relacionamento entre eles?

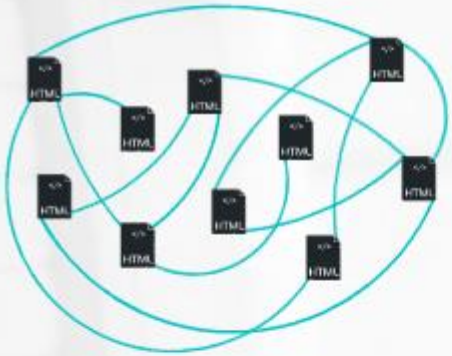
Web

- A partir do começo dos anos 2000 um novo conceito surgiu: Web Semântica, ou Web 3.0
  - Semântica: significado
  - Agora o desafio é atribuir significado aos dados da web
  - Não só documentos, imagens e vídeos estão disponíveis e conectados, mas qualquer tipo de dado
    - Pessoas, locais, dados, músicas, livros, doenças, empresas, meta-dados, etc.
    - Link de “coisas”: Web das Coisas (Web of Things)
  - Os dados agora tem alguns padrões de representação, consulta e relacionamento
  - Web mining dá à Web Semântica

A evolução da WEB

1989 - 1993 | 1999-2004 | 2001-2006

World Wide Web



HTTP

HTML

Link de documentos

Conteúdo estático

Textos, imagens e vídeos

Define apenas conteúdo sintático

Construída basicamente por profissionais de computação

WEB 2.0

Conceitual

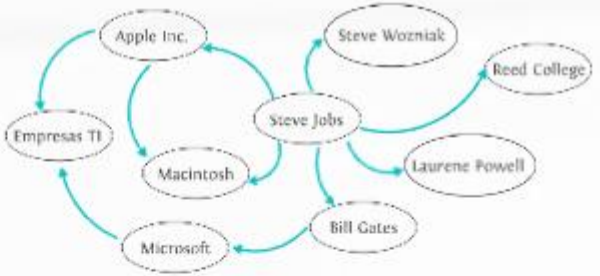
Sem alteração de padrões ou especificações técnicas

Mudança apenas na forma como o conteúdo é criado

Web social e colaborativa

O usuário atua como um provedor de conteúdo

Web Semântica / WEB 3.0



Link de "coisas"

Pessoas, locais, dados, arquivos, vídeos, música, livros, doenças, empresas, etc...

Define novos padrões e especificações técnicas

Representação, consulta e relacionamento

Atribuir significado aos dados

Web Mining

- Web Mining , ou Mineração da Web, é uma subárea de text mining
  - Aplicar as técnicas de data mining e text mining com dados provenientes da web
  - Utilizar técnicas específicas de web mining
- Muitas das técnicas de pré-processamento, classificação, extração e recuperação de informação são utilizadas em web mining
  - Procurar por conteúdos (buscadores)
  - Classificar páginas da web
  - Extrair informação relevante

Web Mining

- Em text mining a unidade básica de informação é um texto não estruturado, disponível em uma base de dados (conjunto)
  - Conjunto de e-books de uma editora
  - Conjunto de matérias de uma revista
  - Conjuntos de artigos científicos de um autor
- Em Web Mining, a unidade básica de informação é uma página web.
- O conjunto de dados pode ser tão grande quanto toda a web, ou apenas um subconjunto dela

Web Mining

- Exemplos
  - Postagens em um blog
  - *Tweets*
  - Notícias de um jornal/revista na web
- As páginas e conteúdo web diferem ligeiramente de textos comuns
  - Existem hiperlinks
  - Semiestruturados
  - Existência de Spams (*spamming*)
  - *Hashtags*

Web Mining – Dados Semiestruturados

- O formato dos documentos presentes na web (HTML, XML, XHTML, JSON...) possibilita que seja dividida blocos com diferentes tipo de conteúdos
  - Propaganda, título, conteúdo principal, cabeçalho, rodapé
- Alguns blocos podem ser mais importantes que outros durante o processo de web mining, recuperação de informação e classificação
  - Conteúdo principal e título podem ter um papel mais importante na recuperação de informação
  - Propagandas podem ter um papel mais importante na detecção spams

http://www.mcti.gov.br/noticias/-/asset_publisher/IqV53KMvD5rY/content/relatorio-anual-do-eso-destaca-pesquisa-feita-no-observatorio-nacional

Web Mining – Dados Semiestruturados

```
<!DOCTYPE html>
▼<html class="ltr yui3-js-enabled webkit ltr js chrome chrome43 chrome43-0 win js flexbox canvas canvastext webgl no-touch
geolocation postmessage websqldatabase indexeddb hashchange history draganddrop websockets rgba hsla multiplebgs backgroundsize
borderimage borderradius boxshadow textshadow opacity cssanimations csscolumns cssgradients cssreflections csstransforms
csstransforms3d csstransitions fontface generatedcontent video audio localstorage sessionstorage webworkers applicationcache svg
inlinesvg smil svgclippaths" dir="ltr" lang="pt-BR" style>
  ▶<head>…</head>
  ▼<body class="azul yui3-skin-sam controls-visible guest-site signed-out public-page site">
    ▶<div id="barra-brasil">…</div>
    <!-- jQuery (necessary for Bootstrap's JavaScript plugins) -->
    <!-- Include all compiled plugins (below), or include individual files as needed -->
    ▼<div id="wrapper">
      ▶<header id="banner" role="banner">…</header>
      ▼<div id="content" class="container">
        ::before
        ▶<div class="responsive-secom-menu col-sm-12 col-xs-12 hidden-lg hidden-md">…</div>
        ▶<nav class="site-breadcrumbs hidden-xs hidden-sm" id="breadcrumbs">…</nav>
        ▼<div class="container-fluid">
          ::before
          ▼<div class="row">
            ::before
            <a id="irparaoconteudo"></a>
            ▶<div class="portlet-boundary portlet-boundary_103_ portlet-static portlet-static-end " id="p_p_id_103_">…</div>
            ▼<div class="MCTI_Geral" id="main-content" role="main">
              ▼<div class="portlet-layout">
                ▶<div class="portlet-column col-md-2 side-menu" id="column-1">…</div>
                ▼<div class="portlet-column col-md-10" id="column-2">
                  ▼<div class="portlet-dropzone portlet-column-content portlet-column-content-last" id="layout-column_column-2">
                    ▼<div class="portlet-boundary portlet-boundary_101_ portlet-static portlet-static-end portlet-borderless
```


http://www.mcti.gov.br/noticias/-/asset_publisher/IqV53KMvD5rY/content/relatorio-anual-do-eso-destaca-pesquisa-feita-no-observatorio-nacional

Faculdade IMPACTA TECNOLOGIA

Web Mining

- O foco de Web mining é descobrir informação e conhecimento relevante a partir dos hiperlinks da Web, conteúdos de páginas e log de uso
- Baseado nisso, Web mining pode ser categorizada em três tipos:
  - Web Structure Mining
  - Web Content Mining
  - Web Usage Mining

Web Mining

- Web Structure mining
  - Extrair informação e descobrir conhecimento através dos hiperlinks (estrutura básica da web)
- Web Content Mining
  - Extrair informação e descobrir conhecimento dos conteúdos das páginas Web
- Web Usage Mining
  - Minera os padrões de acesso dos usuários através dos logs de uso, como por exemplo os cliques dos usuários

Exemplo (Web Mining)

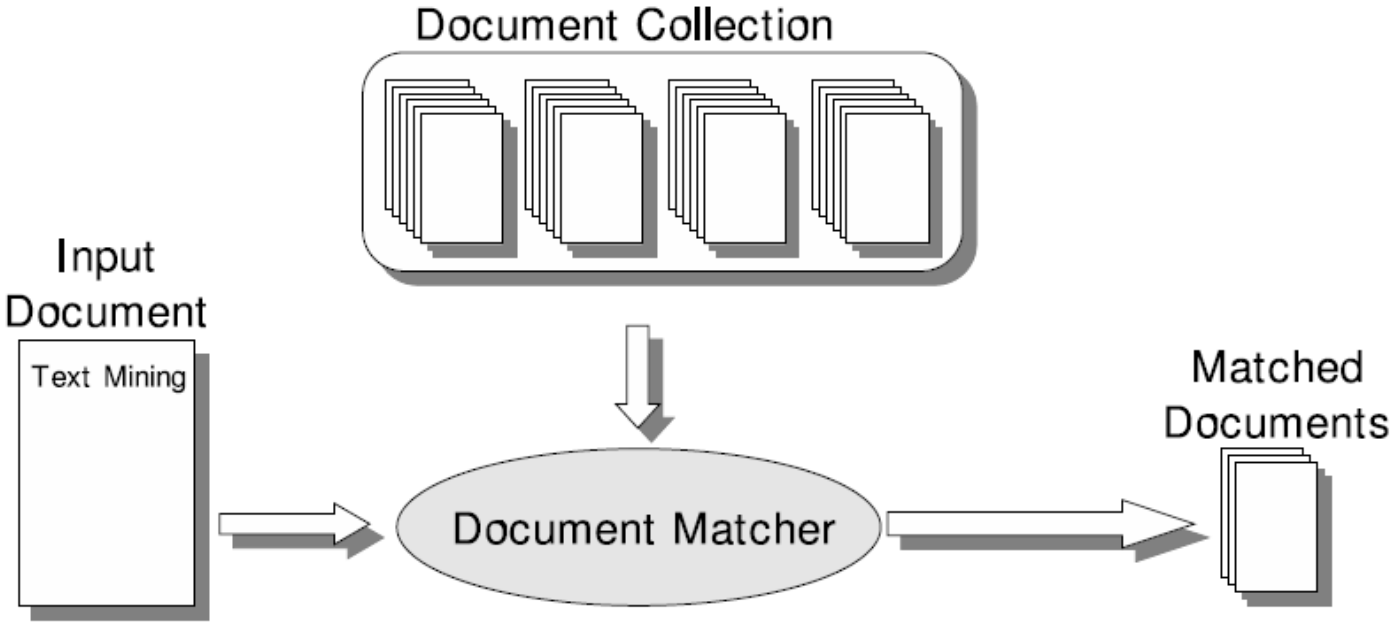
- Os exemplos de Text Mining também podem ser aplicados para documentos extraídos da web
- Classificar páginas web
  - saúde e não saúde;
  - notícias e blog;
  - Economia e esportes;
- Classificar e-mails recebidos
  - Redirecionar para o departamento correto
  - Detectar spam

Exemplo (Web Mining)

- Recuperação de informação
  - Dado um documento, quais documentos em um conjunto mais se assemelham?
  - Dado um conjunto de palavras, quais documentos da web são mais relevantes para estas palavras
    - Buscador

Exemplo (Web Mining)

- Recuperação de informação

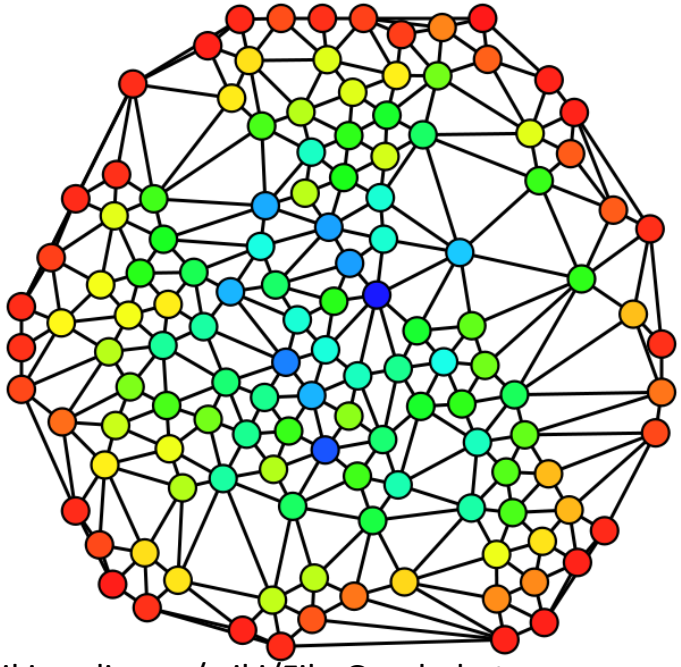


  - O documento de entrada (*input document*) pode ser as palavras de entrada de um buscador

Weiss SM, Indurkhya N, Zhang T. Fundamentals of Predictive Text Mining. 2010 edition. London ; New York: Springer; 2010. 226 p. - Cap. 1

Exemplo (Web Mining)

- Análise de redes sociais
  - Trabalha com Web Structure Mining (análise de links)
  - Quem são as pessoas com mais conexões em uma rede? Quem são as pessoas mais influentes?



https://commons.wikimedia.org/wiki/File:Graph_betweenness.svg

Exemplo (Web Mining)

- Análise de sentimento/opinião
  - A web é acessível a quase todas as pessoas. É comum elas escreverem opiniões sobre produtos, serviços e política
  - A análise de sentimento/opinião avalia se um texto escrito por um usuário é positivo ou negativo
  - O grande desafio é tratar linguagem natural e, muitas vezes, com grafia errada

Weiss SM, Indurkhya N, Zhang T. Fundamentals of Predictive Text Mining. 2010 edition. London ; New York: Springer; 2010. 226 p. - Cap. 1

Trabalho Prático – Aplicação de Web Mining

Avaliação/Projeto

- A aprovação na disciplina está condicionada ao desenvolvimento de um projeto
  - A nota final do curso será a nota do projeto
- Projeto: qualquer aplicação de web mining (tema livre)
- O projeto deverá envolver as seguintes etapas
  - Coleta dos dados
  - Pré-processamento
  - Modelagem (classificação, cluster, recuperação de informação, extração de informação)
  - Avaliação

Avaliação/Projeto

- O aluno poderá utilizar a ferramenta/tecnologia que desejar
  - Rapid Miner
  - Python
  - R
- Sugestões de bases de dados
  - Twitter
  - Notícias
  - Web sites
  - Kaggle (https://www.kaggle.com/datasets)

Softwares e Tecnologias

Preparação do Ambiente

- Vamos utilizar principalmente as seguintes ferramentas:
  - RapidMiner
  - Python 3, anaconda3 e Jupyter

RapidMiner

"RapidMiner Studio is a visual workflow designer that makes data scientists more productive, from the rapid prototyping of ideas to designing mission-critical predictive models."[1]

- O RapidMiner é um software para construção de processos de mineração e analíticos em geral
  - Análise de dados
  - Mineração de dados, textos e web
  - Preparação dos dados
  - Predição e classificação
  - Avaliação

[1] http://docs.rapidminer.com/studio/

Crawler utilizando RapidMiner

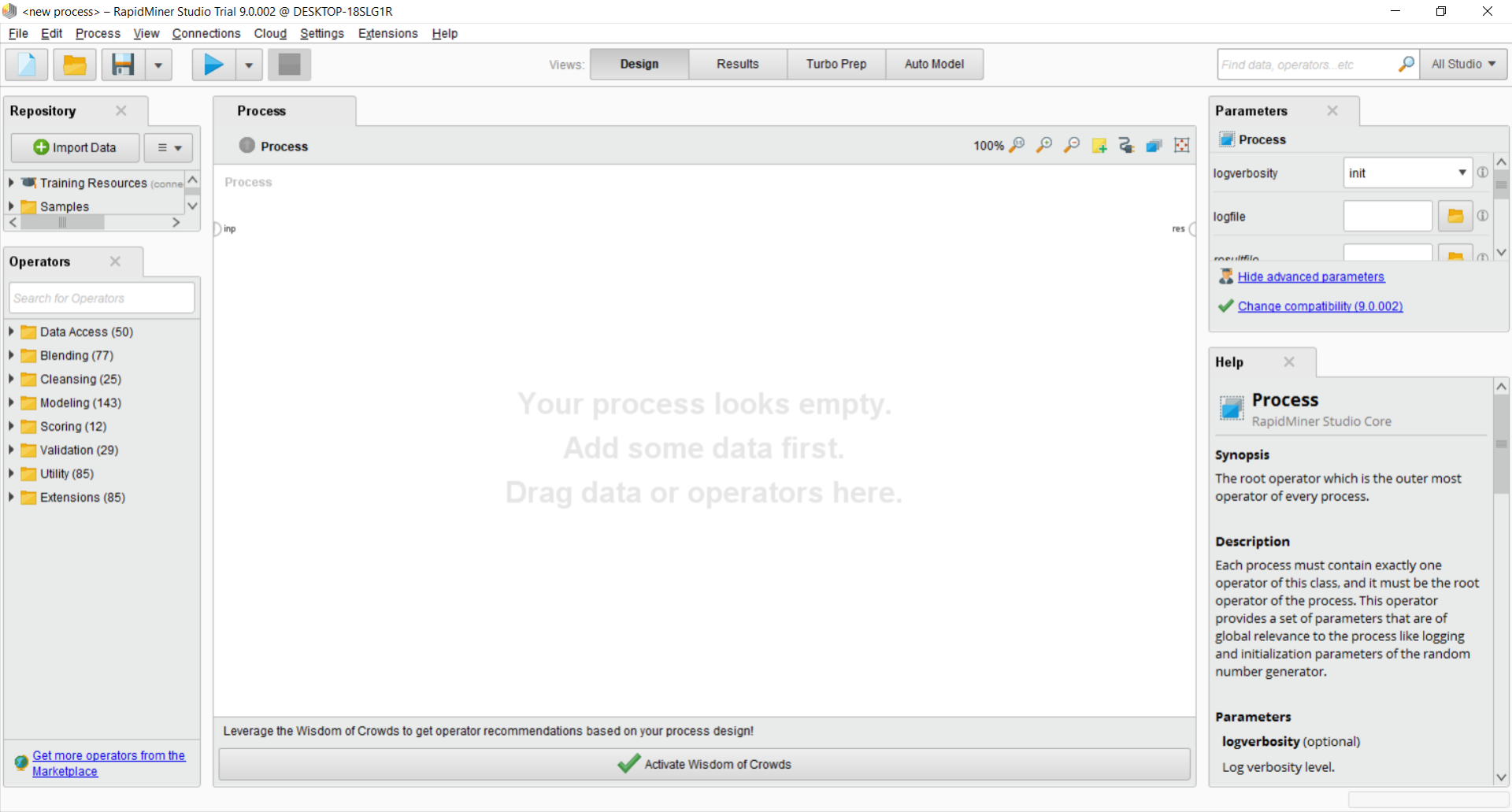
- Além dos processos de análise e predição o RapidMiner também contém métodos que auxiliam a construção do processo e da aplicação
  - Leitura e escrita de dados em diversos formatos
    - Bancos de dados (relacional e não relacional), CSV, Excel, XML, SAS, Access, SPSS, DBase, ...
  - Criar pastas e arquivos no computador
  - Loops e subprocessos
  - Transformação de dados
  - Conexão com Twitter e Crawler
- Website: https://rapidminer.com

Crawler utilizando RapidMiner

- A utilização do RapidMiner é muito fácil e intuitiva
- Basicamente consiste em criar processos analíticos e indicar o conjunto de tarefas sucessivas que o processo executará
  - O Processo é o ponto de partida do RapidMiner
- Os processos são compostos por operadores, que realizam as tarefas em sequência
- A “programação” é feita por
  - “arraste e solte” dos operadores para dentro do processo
  - ligação destes operadores para definir a sequência de tarefas
  - configuração dos parâmetros dos operadores

Faculdade IMPACTA TECNOLOGIA

RapidMiner



Faculdade IMPACTA TECNOLOGIA

RapidMiner

- O RapidMiner tem suporte para realizar tarefas de mineração de texto e web
- Deve-se instalar as extensões “Text Mining” e “Web Mining” para habilitar os operadores
- Acesse o Menu “Extensions” → “Marketplace (Updates and Extensions)”
- Digite “text” na caixa de pesquisa que ambas aparecerão
- Instale as duas extensões

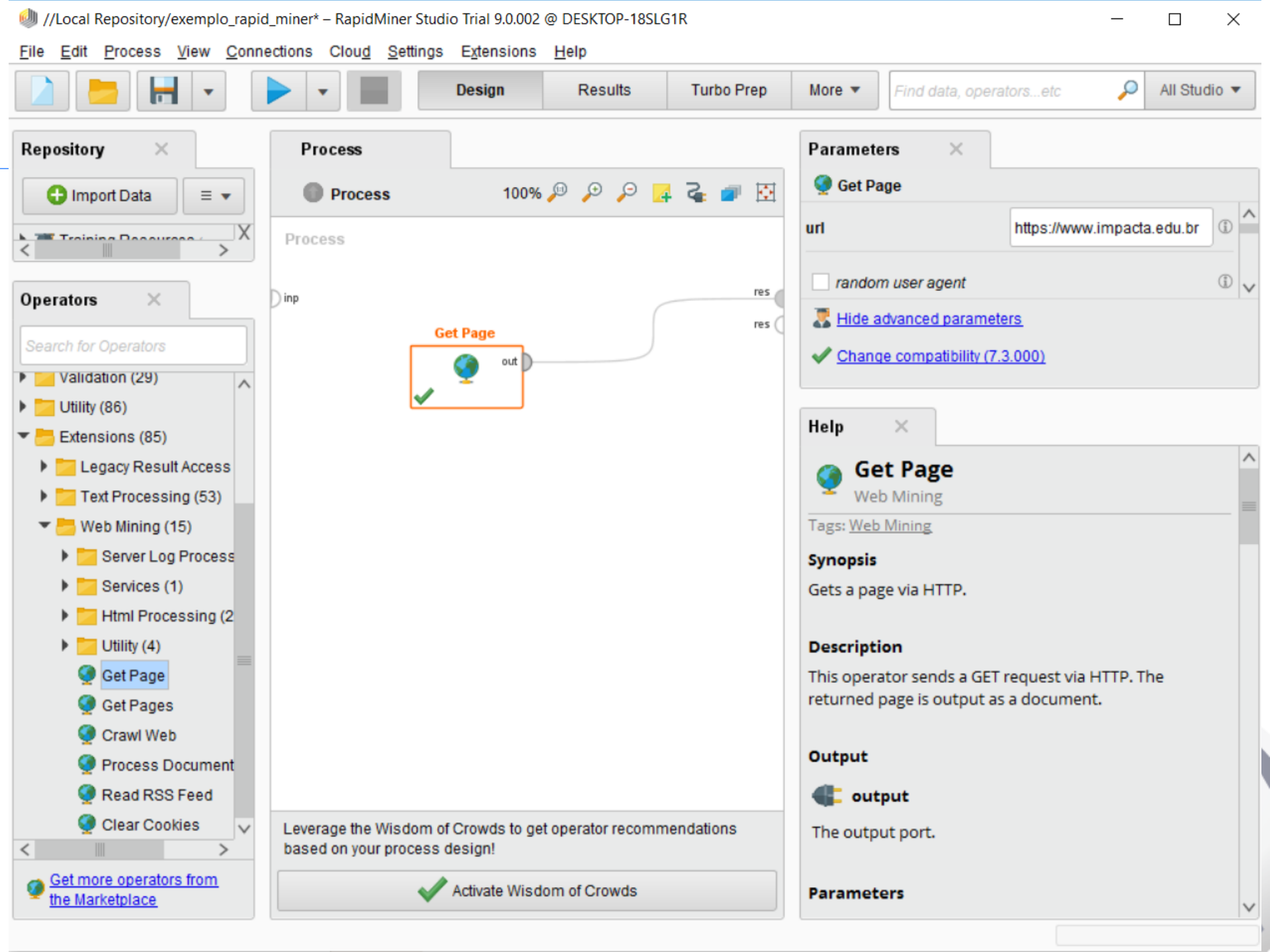
RapidMiner

- Faça o primeiro teste com o RapidMiner
- Procure o operador Get Page dentro do painel “Operators”, árvore Extensions -> Web Mining, do lado esquerdo da tela
  - Este operador recupera o conteúdo de uma página web
- Selecione o operador, arraste e solte par dentro da área branca no centro (Process)
- Com o operador selecionado, no painel de parâmetros (Parameters) ao lado direito coloque a URL que deseja pegar o conteúdo na caixa url

Faculdade IMPACTA TECNOLOGIA

RapidMiner

- Volte para o operador Get Page e ligue a saída do operador (out) na saída do processo (res)
- Aperte o Play (painel superior) e veja o resultado na visão Results (painel superior)


//Local Repository/exemplo_rapid_miner* – RapidMiner Studio Trial 9.0.002 @ DESKTOP-18SLG1R
File Edit Process View Connections Cloud Settings Extensions Help
Design Results Turbo Prep More
Find data, operators...etc
All Studio
Repository
Import Data
Operators
Search for Operators
Validation (29)
Utility (86)
Extensions (85)
Legacy Result Access
Text Processing (53)
Web Mining (15)
Server Log Process
Services (1)
Html Processing (2
Utility (4)
Get Page
Get Pages
Crawl Web
Process Document
Read RSS Feed
Clear Cookies
Get more operators from the Marketplace
Process
Process
100%
Process
inp
Get Page
out
res
res
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!
Activate Wisdom of Crowds
Parameters
Get Page
url
https://www.impacta.edu.br
random user agent
Hide advanced parameters
Change compatibility (7.3.000)
Help
Get Page
Web Mining
Tags: Web Mining
Synopsis
Gets a page via HTTP.
Description
This operator sends a GET request via HTTP. The returned page is output as a document.
Output
output
The output port.
Parameters

Result History | Document (Get Page)

Document

```
<!DOCTYPE html>
<html lang="pt-br" itemscope itemtype="https://schema.org/WebSite">
  <head>
    <meta charset="UTF-8">
    <meta name="viewport" content="width=device-width,initial-scale=1,maximum-scale=1.0,user-scalable=0">
                                <meta name="google-site-verification" content="UqmFbE7eMs-8xlWxaSyb6E3yt8LMNeEyDGc

<!-- Google Tag Manager -->
<script type="text/javascript" async defer>
  (function(w,d,s,l,i){w[l]=w[l]||[];w[l].push({'gtm.start':
  new Date().getTime(),event:'gtm.js'});var f=d.getElementsByTagName(s)[0],
  j=d.createElement(s),dl=l!='dataLayer'?'&l='+l:'';j.async=true;j.src=
  'https://www.googletagmanager.com/gtm.js?id='+i+dl;f.parentNode.insertBefore(j,f);
  })(window,document,'script','dataLayer','GTM-W4ND9BQ');
</script>
<!-- End Google Tag Manager -->
```

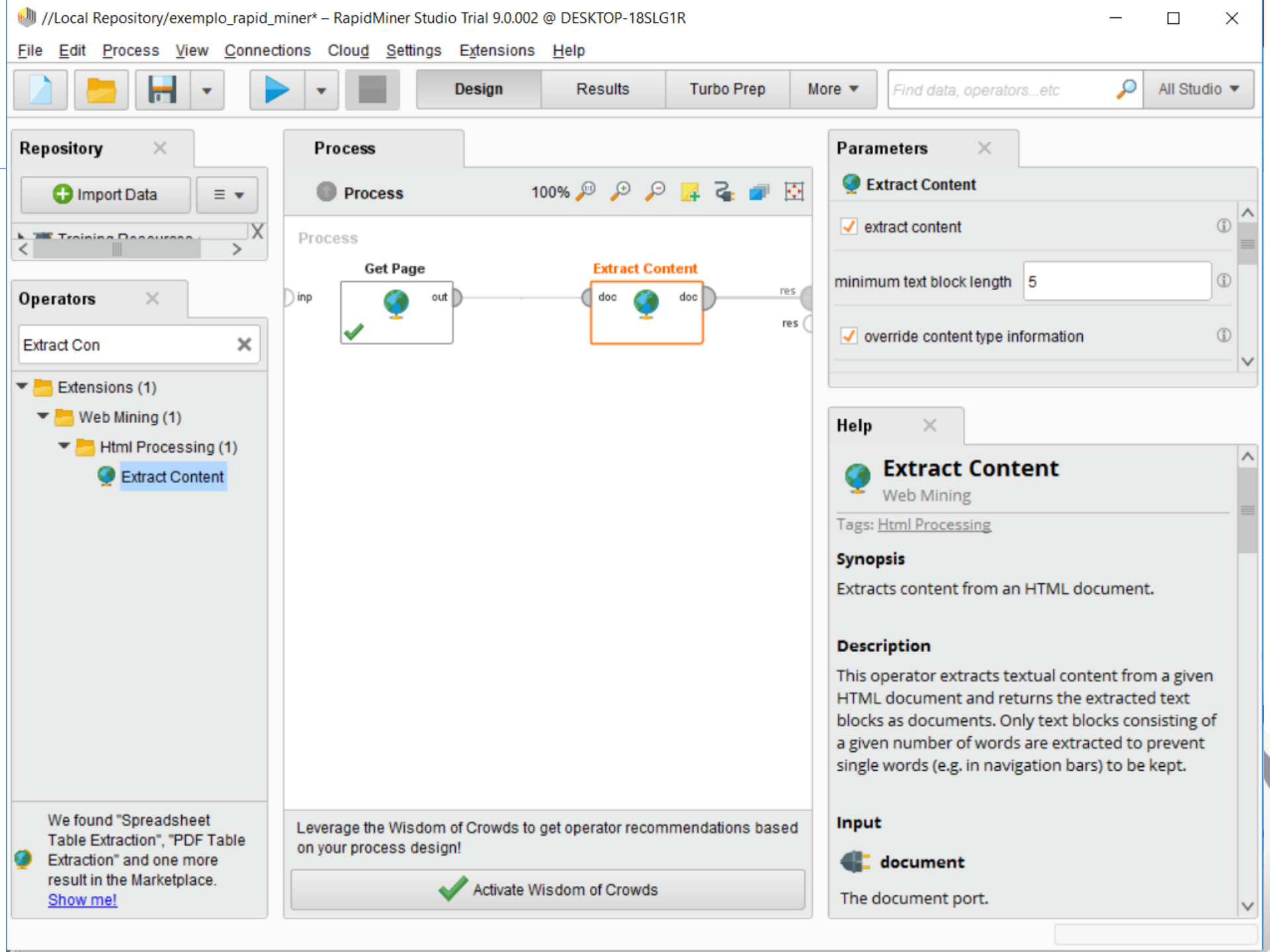

```
<!DOCTYPE html>
<html lang="pt-br" itemscope itemtype="https://schema.org/WebSite">
  <head>
    <meta charset="UTF-8">
    <meta name="viewport" content="width=device-width,initial-scale=1,maximum-scale=1.0,user-scalable=0">
                                <meta name="google-site-verification" content="UqmFbE7eMs-8xlWxaSyb6E3yt8LMNeEyDGc

<!-- Google Tag Manager -->
<script type="text/javascript" async defer>
  (function(w,d,s,l,i){w[l]=w[l]||[];w[l].push({'gtm.start':
  new Date().getTime(),event:'gtm.js'});var f=d.getElementsByTagName(s)[0],
  j=d.createElement(s),dl=l!='dataLayer'?'&l='+l:'';j.async=true;j.src=
  'https://www.googletagmanager.com/gtm.js?id='+i+dl;f.parentNode.insertBefore(j,f);
  })(window,document,'script','dataLayer','GTM-W4ND9BQ');
</script>
<!-- End Google Tag Manager -->
```


URL	https://w...
Respon...	200
Respon...	OK
Content-...	text/html;...
Content-...	44314
Date	Sep 11, ...
Last-Mo...	?
Expires	Nov 19, ...

RapidMiner

- Agora procure o operador Extract Content (Extensions -> Web Mining ->Html Processing)
  - Dica: Você pode procurar pelo nome do operador na caixa de texto do painel Operators
  - Extract Content remove o conteúdo HTML de uma página web, mantendo apenas o texto
- Ligue a saída do Get Page (out) na entrada de Extract Content (doc)
- A saída de Extract Content (doc) deve ser ligada na saída do Processo (res)
- Veja o resultado


//Local Repository/exemplo_rapid_miner* – RapidMiner Studio Trial 9.0.002 @ DESKTOP-18SLG1R
File Edit Process View Connections Cloud Settings Extensions Help
Design Results Turbo Prep More
Find data, operators...etc
All Studio
Repository
Import Data
Operators
Extract Con
Extensions (1)
Web Mining (1)
Html Processing (1)
Extract Content
We found "Spreadsheet Table Extraction", "PDF Table Extraction" and one more result in the Marketplace.
Show me!
Process
Process
100%
Process
Get Page
inp
out
Extract Content
doc
doc
res
res
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!
Activate Wisdom of Crowds
Parameters
Extract Content
extract content
minimum text block length
5
override content type information
Help
Extract Content
Web Mining
Tags: Html Processing
Synopsis
Extracts content from an HTML document.
Description
This operator extracts textual content from a given HTML document and returns the extracted text blocks as documents. Only text blocks consisting of a given number of words are extracted to prevent single words (e.g. in navigation bars) to be kept.
Input
document
The document port.

Result History | Document (Extract Content)

Document

Faculdade Impacta - Faculdade Impacta SABE COMO A TECNOLOGIA MUDA O ? ÚLTIMAS VAGAS Considerada uma das melhores faculdades com foco em tecnologia do País, a Faculdade Impacta oferece cursos de Graduação, Pós-Graduação e MBA e cursos técnicos. Conheça quais são as áreas de conhecimento que a Impacta aborda e saiba um pouco mais sobre cada curso: Graduação Cursos na área de Tecnologia, Gestão e Design Ver cursos Pós e MBA Diferentes áreas de conhecimento para você crescer no mercado Ver cursos Técnico Acelere sua inserção no mercado de trabalho com cursos de qualidade Ver cursos A Faculdade Impacta é altamente reconhecida por sua atuação no mercado da educação, o que comprova o compromisso da Instituição com o ensino de qualidade e a dedicação do Grupo Impacta em transformar a sociedade através da tecnologia. Desta maneira, ao longo dos anos a Faculdade Impacta tem recebido prêmios e reconhecimentos que são resultados do trabalho contínuo de professores e funcionários e da determinação de seus alunos. As salas de aula da Faculdade Impacta são equipadas com lousas eletrônicas, que capturam todos os pontos das aulas presenciais, como voz, escrita dinâmica, vídeos e navegação do professor. Todo o conteúdo é gravado e disponibilizado em uma sala de ensino virtual para que o aluno assista quando e onde quiser, de qualquer plataforma digital. Profissões Confira 5 vantagens de cursar uma pós-graduação Saiba mais Tecnologia Pós em Cloud Computing fecha parceria com a Cisco Saiba mais Carreira 5 dicas práticas para conseguir um bom estágio durante a faculdade Saiba mais Profissões Como convencer meu chefe a investir na minha educação? Saiba mais Profissões Quais as vantagens de fazer uma pós em Cloud Computing? Saiba mais Profissões Confira 5 vantagens de cursar uma pós-graduação Saiba mais Tecnologia Pós em Cloud Computing fecha parceria com a Cisco Saiba mais Carreira 5 dicas práticas para conseguir um bom estágio durante a faculdade Saiba mais Profissões Como convencer meu chefe a investir na minha educação? Saiba mais Profissões Quais as vantagens de fazer uma pós em Cloud Computing? Saiba mais Confira os

Faculdade Impacta - Faculdade Impacta SABE COMO A TECNOLOGIA MUDA O ? ÚLTIMAS VAGAS Considerada uma das melhores faculdades com foco em tecnologia do País, a Faculdade Impacta oferece cursos de Graduação, Pós-Graduação e MBA e cursos técnicos. Conheça quais são as áreas de conhecimento que a Impacta aborda e saiba um pouco mais sobre cada curso: Graduação Cursos na área de Tecnologia, Gestão e Design Ver cursos Pós e MBA Diferentes áreas de conhecimento para você crescer no mercado Ver cursos Técnico Acelere sua inserção no mercado de trabalho com cursos de qualidade Ver cursos A Faculdade Impacta é altamente reconhecida por sua atuação no mercado da educação, o que comprova o compromisso da Instituição com o ensino de qualidade e a dedicação do Grupo Impacta em transformar a sociedade através da tecnologia. Desta maneira, ao longo dos anos a Faculdade Impacta tem recebido prêmios e reconhecimentos que são resultados do trabalho contínuo de professores e funcionários e da determinação de seus alunos. As salas de aula da Faculdade Impacta são equipadas com lousas eletrônicas, que capturam todos os pontos das aulas presenciais, como voz, escrita dinâmica, vídeos e navegação do professor. Todo o conteúdo é gravado e disponibilizado em uma sala de ensino virtual para que o aluno assista quando e onde quiser, de qualquer plataforma digital. Profissões Confira 5 vantagens de cursar uma pós-graduação Saiba mais Tecnologia Pós em Cloud Computing fecha parceria com a Cisco Saiba mais Carreira 5 dicas práticas para conseguir um bom estágio durante a faculdade Saiba mais Profissões Como convencer meu chefe a investir na minha educação? Saiba mais Profissões Quais as vantagens de fazer uma pós em Cloud Computing? Saiba mais Profissões Confira 5 vantagens de cursar uma pós-graduação Saiba mais Tecnologia Pós em Cloud Computing fecha parceria com a Cisco Saiba mais Carreira 5 dicas práticas para conseguir um bom estágio durante a faculdade Saiba mais Profissões Como convencer meu chefe a investir na minha educação? Saiba mais Profissões Quais as vantagens de fazer uma pós em Cloud Computing? Saiba mais Confira os

URL	https://w...
Respon...	200
Respon...	OK
Content-...	text/html;...
Content-...	44314
Date	Sep 11, ...
Last-Mo...	?
Expires	Nov 19, ...
Title	Faculda...
Language	?
Descripti...	A Faculd...
Keywords	?
Robots	index, fol...

Faculdade IMPACTA TECNOLOGIA

Python

- Vamos utilizar o Python a partir do Anaconda3 (https://www.anaconda.com/download/#windows), escrevendo os programas no Jupyter Notebook
- Principais bibliotecas utilizadas
  - numpy
  - pandas
  - bs4
  - urllib
  - sklearn
  - nltk

Python

- Abra o Jupyter Notebook
- Crie um novo notebook
- Coloque o seguinte código, para capturar uma página web e extrair o conteúdo textual

```
from urllib.request import urlopen
from bs4 import BeautifulSoup
html = urlopen("https://pt.wikipedia.org/wiki/Cerveja").read()
bsObj = BeautifulSoup(html, "html.parser")
# remover script e style
for script in bsObj(["script", "style"]): script.extract()
texto = bsObj.text
#remover quebras de linha excessivas
linhas = (linha.strip() for linha in texto.splitlines())
linhas = (frase.strip() for linha in linhas for frase in linha.split("  "))
texto = ' '.join(linha for linha in linhas if linha)
texto
```

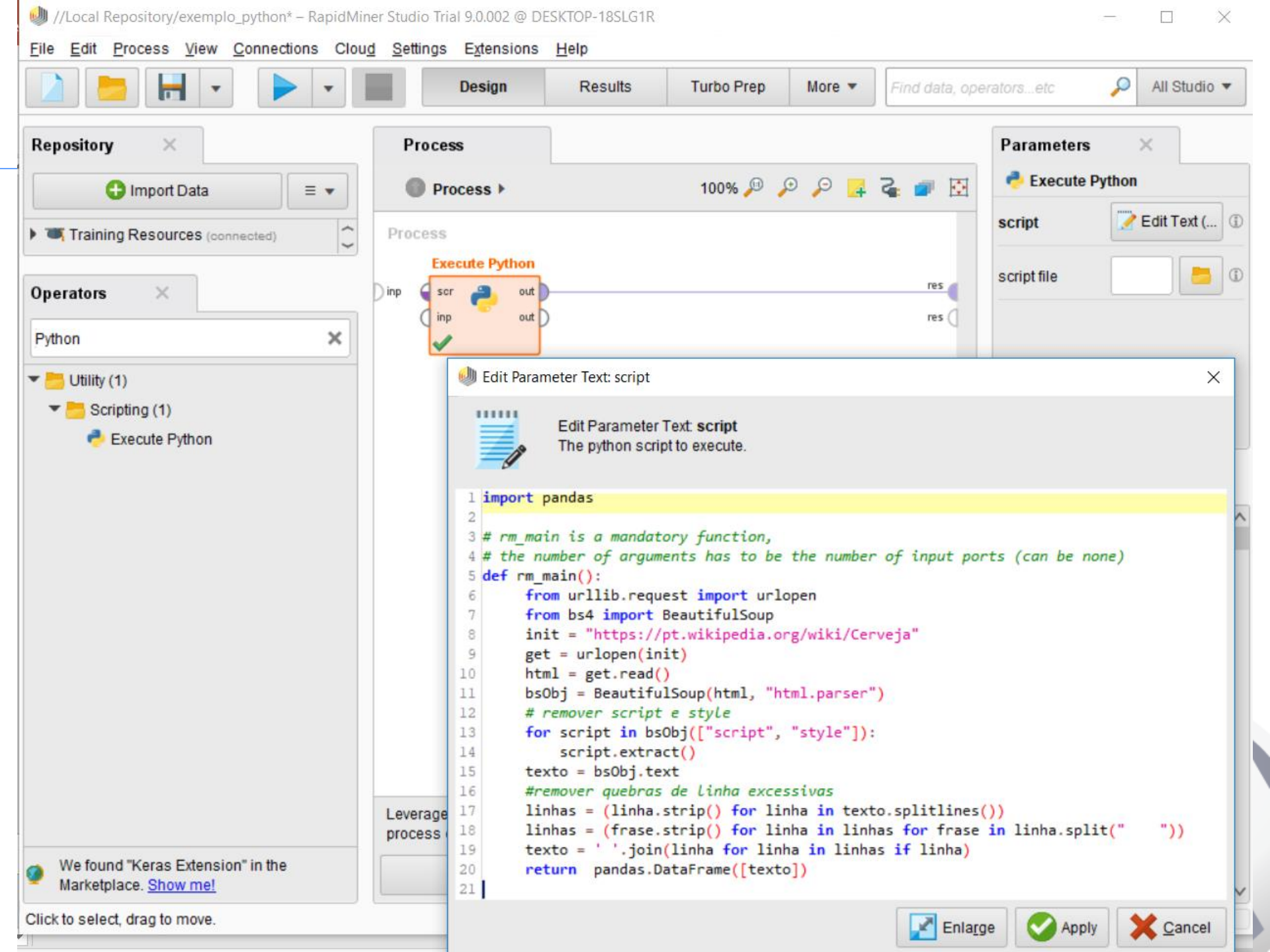
RapidMiner + Python

- É possível executar scripts python diretamente do RapidMiner
- Instale a extensão Python
- Procure o operador “Execute Python” (Utility -> Scripting) e coloque no Processo
- Configure o caminho do Python
  - Menu Settings → Preferences -> Aba Python Scripting
  - Selecione o caminho do Python da instalação do Anaconda
- Clique duas vezes no operador “Execute Python” que colocou no processo e coloque o seguinte código dentro da função rm_main (deve ter essa função)

RapidMiner + Python

```
from urllib.request import urlopen
        from bs4 import BeautifulSoup
        html = urlopen("https://pt.wikipedia.org/wiki/Cerveja").read()
        bsObj = BeautifulSoup(html, "html.parser")
        # remover script e style
        for script in bsObj(["script", "style"]): script.extract()
        texto = bsObj.text
        #remover quebras de linha excessivas
        linhas = (linha.strip() for linha in texto.splitlines())
        linhas = (frase.strip() for linha in linhas for frase in linha.split("   "))
        texto = ' '.join(linha for linha in linhas if linha)
        return  pandas.DataFrame([texto])
```

- É o mesmo código executado no Jupyter, exceto pelo retorno
- Ligue a saída deste operador na saída do processo e veja o resultado


//Local Repository/exemplo_python* – RapidMiner Studio Trial 9.0.002 @ DESKTOP-18SLG1R
File Edit Process View Connections Cloud Settings Extensions Help
Design Results Turbo Prep More
Find data, operators...etc
All Studio
Repository
Import Data
Training Resources (connected)
Operators
Python
Utility (1)
Scripting (1)
Execute Python
We found "Keras Extension" in the Marketplace. Show me!
Click to select, drag to move.
Process
Process
100%
Process
Execute Python
inp
scr
inp
out
out
res
res
Parameters
Execute Python
script
Edit Text (...
script file
Edit Parameter Text: script
Edit Parameter Text: script
The python script to execute.
1 import pandas
2
3 # rm_main is a mandatory function,
4 # the number of arguments has to be the number of input ports (can be none)
5 def rm_main():
6 from urllib.request import urlopen
7 from bs4 import BeautifulSoup
8 init = "https://pt.wikipedia.org/wiki/Cerveja"
9 get = urlopen(init)
10 html = get.read()
11 bsObj = BeautifulSoup(html, "html.parser")
12 # remover script e style
13 for script in bsObj(["script", "style"]):
14 script.extract()
15 texto = bsObj.text
16 #remover quebras de linha excessivas
17 linhas = (linha.strip() for linha in texto.splitlines())
18 linhas = (frase.strip() for linha in linhas for frase in linha.split(" "))
19 texto = ' '.join(linha for linha in linhas if linha)
20 return pandas.DataFrame([texto])
21
Enlarge
Apply
Cancel

Redes sociais - Twitter

- Acesse https://developer.twitter.com e inscreva-se para poder utilizar a API do Twitter de coleta
- Utilizaremos o twitter nas últimas aulas do curso

Faculdade
IMPACTA
TECNOLOGIA