

Ciência de Dados e Big Data

Recuperação da Informação na Web e em Redes Sociais

PUC-Minas IEC | Pós-Graduação Lato Sensu

Zilton Cordeiro Jr.

Projeto Final

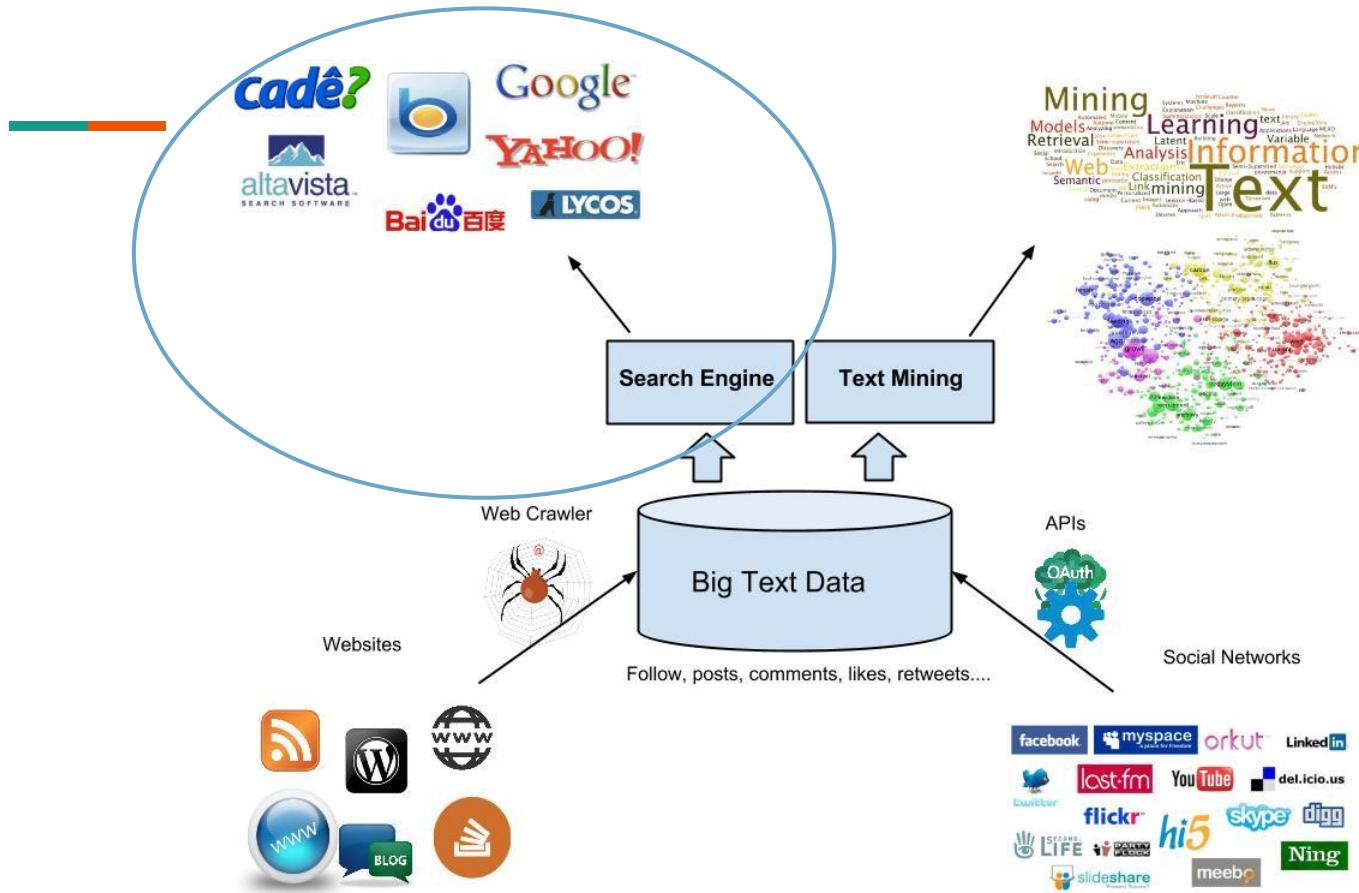


Pensaram em algo?

Projeto Final

- ❖ O **projeto final** consiste em realizar um estudo da Web para um assunto real e de livre escolha.
 - Exemplos: Automóveis, moda, música, imóveis...
- ❖ Será necessário
 - Coletar dados em texto de redes sociais e sites da Web
 - Analisar o conteúdo textual obtido
 - Analisar dados de relacionamentos entre usuários (i.e. nas redes)
 - **Relatório final**
- ❖ **Data de Entrega**
 - 15º dia após a última aula às 23:59hrs

Mineração da Web e Redes Sociais



Modelos em RI



Modelos em Recuperação da Informação

- Núcleo de qualquer sistema de recuperação de informação
- Utilizados para representar características semânticas dos elementos envolvidos nos sistemas
- Modelos clássicos: booleano, **vetorial** e probabilístico
- Modelos bastante utilizados: **vetorial**, language models e BM25 (que são probabilísticos)
- Diversos outros modelos na literatura

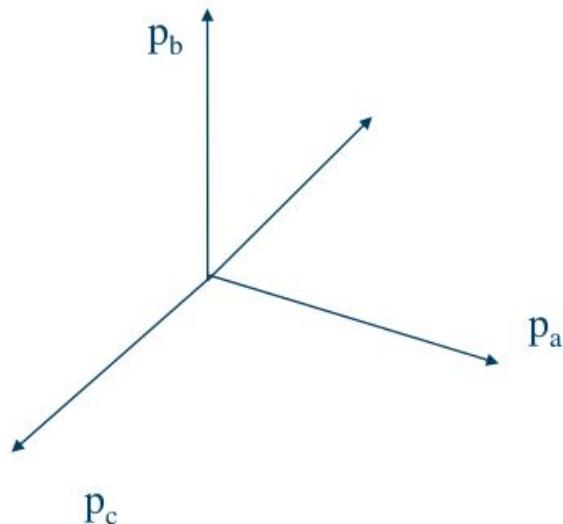
Modelo Vetorial

- Proposto em 1968 e continua sendo muito empregado **hoje em dia**

- Originalmente utilizado para resolver problemas de **busca**
- Sucesso reside na eficiência e nos bons resultados obtidos
- Todos os componentes do sistema são vistos como conjuntos de palavras
- Elementos a serem modelados são representados como vetores dentro de um **espaço vetorial**.
- Dimensão do espaço é dada pelo **número de palavras distintas**

Modelo Vetorial

- Número de palavras distintas da coleção de documentos determina dimensão do espaço onde os documentos e consultas serão representados



- Como determinar as coordenadas dos elementos ?

Modelo Vetorial

Um *corpus* contendo n documentos e i termos de indexação pode ser representado da seguinte forma:

	t_1	t_2	t_3	...	t_i
DOC₁	$w_{1,1}$	$w_{2,1}$	$w_{3,1}$...	$w_{i,1}$
DOC₂	$w_{1,2}$	$w_{2,2}$	$w_{3,2}$...	$w_{i,2}$
.
.
.
DOC_n	$w_{1,n}$	$w_{2,n}$	$w_{3,n}$...	$w_{i,n}$

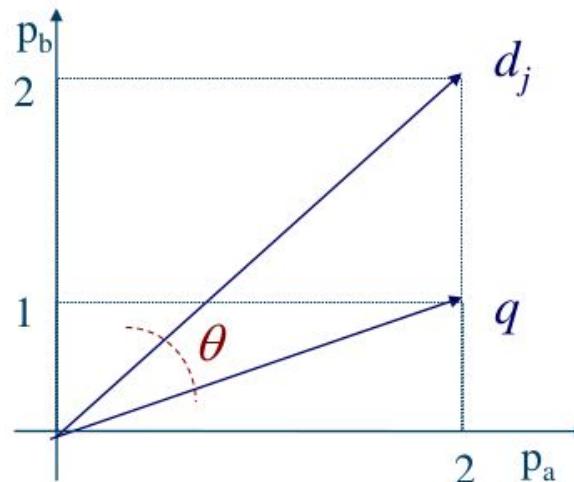
onde $w_{i,n}$ representa o peso do i -ésimo termo do n -ésimo documento.

Modelo Vetorial

- Componentes do sistema são vistos como vetores cujas coordenadas são determinadas pelas palavras que os descrevem

$$\vec{d}_j = (2, 2)$$

$$\vec{q} = (2, 1)$$



- A frequência de termos em comum é suficiente para similaridade?

Medidas de TF e IDF

Term Frequency - Inverse Document Frequency

- O IDF tenta expressar a “importância” de uma palavra dentro da coleção
- **N**: número total de documentos de uma coleção
- **n_t** : número de documentos onde a palavra t ocorreu

$$Idf(t) = \log\left(\frac{N}{n_t}\right)$$

- Quanto mais rara a palavra, maior seu idf !

Medidas de TF e IDF

Coordenada do doc
d no eixo t

$$Idf(t) = \log\left(\frac{N}{n_t}\right)$$

$$w(d,t) = tf(d,t) \times idf(t)$$

Frequência da palavra
t no documento d

Importância de t
na coleção

Medidas de TF e IDF

D1	A A A B
D2	A A C
D3	A A
D4	B B

$$idf(A) = \log\left(\frac{4}{3}\right) = 0,28$$

$$idf(B) = \log\left(\frac{4}{2}\right) = 0,69$$

TF-IDF dos Termos nos Documentos

- Pesos para termos do documento D1

D1	A A A B
----	---------

$$w(D1, A) = idf(A) \times tf(D1, A) = 0,28 \times 3 = 0,84$$

Exemplo de Processamento de Consultas

- Pesos para termos do documento D1

D1	A A A B
----	---------

$$w(D1, A) = idf(A) \times tf(D1, A) = 0,28 \times 3 = 0,84$$

$$w(D1, B) = idf(B) \times tf(D1, B) = 0,69 \times 1 = 0,69$$

Exemplo de Processamento de Consultas

- Pesos para termos do documento D1

D1	A A A B
----	---------

$$w(D1, A) = idf(A) \times tf(D1, A) = 0,28 \times 3 = 0,84$$

$$w(D1, B) = idf(B) \times tf(D1, B) = 0,69 \times 1 = 0,69$$

$$w(D1, C) = idf(C) \times tf(D1, C) = 1,38 \times 0 = 0$$

Exemplo de Processamento de Consultas

- Pesos para termos do documento D1

D1	A A A B
----	---------

$$w(D1, A) = idf(A) \times tf(D1, A) = 0,28 \times 3 = 0,84$$

$$w(D1, B) = idf(B) \times tf(D1, B) = 0,69 \times 1 = 0,69$$

$$w(D1, C) = idf(C) \times tf(D1, C) = 1,38 \times 0 = 0$$

$$\vec{D1} = (0,84; 0,69; 0)$$

Exemplo de Processamento de Consultas

- Para uma consulta “A B”


$$w(Q, A) = idf(A) \times tf(Q, A) = 0,28 \times 1 = 0,28$$

Exemplo de Processamento de Consultas

- Para uma consulta “A B”


$$w(Q, A) = idf(A) \times tf(Q, A) = 0,28 \times 1 = 0,28$$

$$w(Q, B) = idf(B) \times tf(Q, B) = 0,69 \times 1 = 0,69$$

Exemplo de Processamento de Consultas

- Para uma consulta “A B”


$$w(Q, A) = idf(A) \times tf(Q, A) = 0,28 \times 1 = 0,28$$

$$w(Q, B) = idf(B) \times tf(Q, B) = 0,69 \times 1 = 0,69$$

$$w(Q, C) = idf(C) \times tf(Q, C) = 1,38 \times 0 = 0$$

Exemplo de Processamento de Consultas

- Para uma consulta “A B”

$$w(Q, A) = idf(A) \times tf(Q, A) = 0,28 \times 1 = 0,28$$

$$w(Q, B) = idf(B) \times tf(Q, B) = 0,69 \times 1 = 0,69$$

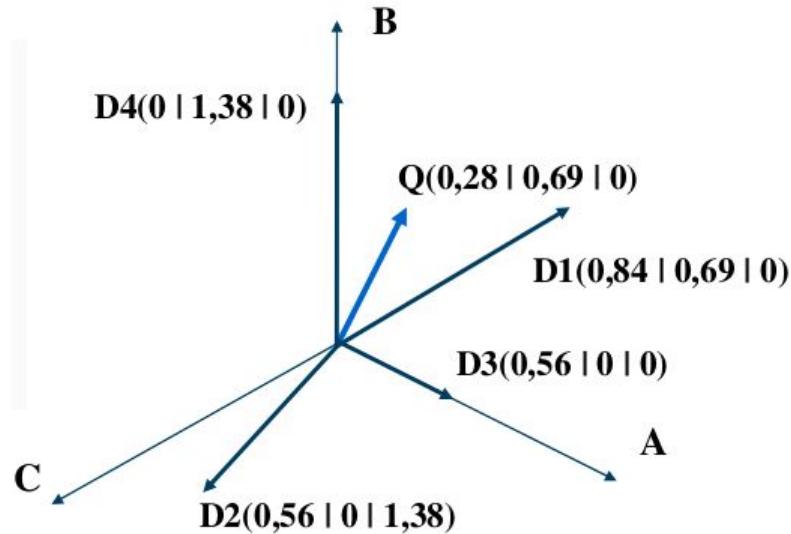
$$w(Q, C) = idf(C) \times tf(Q, C) = 1,38 \times 0 = 0$$

$$\vec{Q}1 = (0,28, 0,69, 0)$$

Implementação do Modelo Vetorial

$Q = "A\ B"$

D1	A A A B
D2	A A C
D3	A A
D4	B B



Similaridade

- As consultas são utilizadas para determinar importância das páginas
- Um modelo de RI é utilizado para computar a similaridade entre uma consulta Q e um determinado documento P (página web)

Similaridade

- Correlação entre dois vetores é utilizada para medir a proximidade entre os elementos reais modelados

The diagram illustrates the cosine similarity formula between a document d and a query q . It shows the formula as a fraction where the numerator is the dot product of the vectors, and the denominator is the product of their norms.

Acumuladores:

D1	D2	D3	D4
0,71	0,16	0,16	0,95

$$sim(d, q) = \cos\theta = \frac{\sum_{i=1}^t w(i, d) \times w(i, q)}{\sqrt{\sum (w(i, d))^2} \times \sqrt{\sum (w(i, q))^2}}$$

Norma do documento Norma da consulta

Acumuladores para cálculo da similaridade parcial

Acumuladores:

D1	D2	D3	D4
0,71	0,16	0,16	0,95

$$\text{sim}(d1, q) = \frac{\text{Acum}(d1)}{\|\vec{d1}\| \times \|\vec{q}\|} = \frac{0,71}{1,08 \times \|\vec{q}\|} = \frac{0,66}{\|\vec{q}\|}$$

$$\text{sim}(d2, q) = \frac{0,16}{1,49 \times \|\vec{q}\|} = \frac{0,17}{\|\vec{q}\|}$$

$$\text{sim}(d3, q) = \frac{0,16}{0,56 \times \|\vec{q}\|} = \frac{0,28}{\|\vec{q}\|}$$

$$\text{sim}(d4, q) = \frac{0,95}{1,38 \times \|\vec{q}\|} = \frac{0,69}{\|\vec{q}\|}$$

Exemplo de Processamento de Consultas

Q = "A B"

D1	A A A B	0,66
D2	A A C	0,17
D3	A A	0,28
D4	B B	0,69

Características do Modelo Vetorial

- Utiliza pesos tanto para os termos de indexação quanto para os termos da expressão de busca:
 - Permite o cálculo de um valor numérico que representa a relevância de cada documento em relação à busca;
- O resultado de uma busca é um conjunto de documentos ordenados pelo grau de similaridade da expressão de busca e cada documento do *corpus*;

Características do Modelo Vetorial

- O ordenamento permite restringir o resultado a um número máximo de documentos desejados;
- É possível, também, restringir a quantidade de documentos recuperados definindo um limite mínimo para o valor da similaridade;

Aplicações do Modelo Vetorial

- O modelo vetorial pode ser aplicado em qualquer tipo de problema de RI.

- Aplicação direta do modelo em sistemas de busca
- **Homogeneidade do Modelo Vetorial:**
 - Característica fundamental que permite uma grande variedade de operações relacionadas à recuperação de informação, incluindo:
 - Indexação,
 - Clustering (agrupamento),
 - *Relevance feedback*,
 - Classificação,
 - Reformulação da expressão de busca,
 - etc

Aplicações do Modelo Vetorial

❖ Filtragem com Modelo Vetorial

- Bases de dados contêm perfis no lugar de documentos
- Perfis são conjuntos de termos que descrevem os interesses dos usuários
- Documentos que chegam para o sistema são tratado como consultas

Dúvidas?

- As medidas que dão pesos para termos distintos, podem ser usadas como input para outro tipo de classificador?

- A similaridade de documentos pode ser utilizada para agrupar documentos similares (doc-doc) ao invés de consultas e documentos (termos - doc)?

Pré-processamento de documentos



Pré-processamento de documentos

- O pré-processamento de documentos é um importante procedimento empregado na construção de sistemas de RI
- Operações (ou transformações) textuais:
 - Análise léxica do texto
 - Eliminação de stopwords
 - Stemming das palavras
 - Seleção de termos ou palavras-chave

Pré-processamento de documentos

➤ Análise léxica

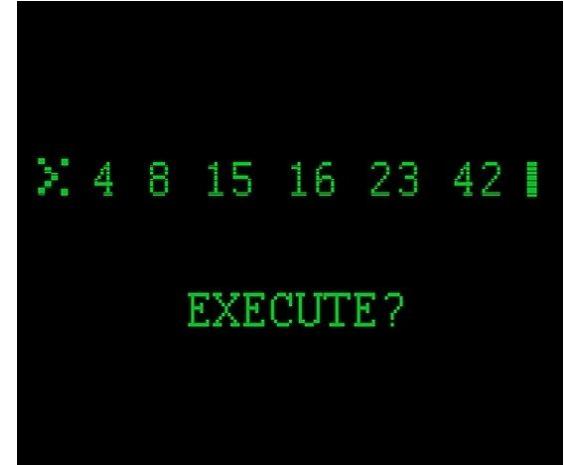
- Processo de conversão de uma sequência de caracteres em uma sequência de palavras (*Bag-of-words*)
- Usando somente espaços?
 - Dígitos
 - Hífen
 - Marcas de pontuação
 - Caixa das palavras (maiúsculas e minúsculas)



Pré-processamento de documentos

➤ Dígitos

- Números sozinhos são vagos
- 1987 pode representar um ano ou um número de pessoas em um registro
- Usualmente números não são considerados como termos de índice



```
3: 4 8 15 16 23 42 |  
EXECUTE?
```

Pré-processamento de documentos

➤ Hífen

- Difícil decisão para o analisador léxico
- Existem palavras que incluem hífens como parte integral
 - Guarda-chuva, B-52

Obs: Adote uma regra geral, mas tenha consciência das exceções.

Pré-processamento de documentos

➤ Marcas de pontuação

- Removidas por completo do texto
- Baixo risco de não interpretar palavras sem pontuação
 - “300 A.C.” será interpretado de maneira similar ao remover a pontuação

A collection of various punctuation marks including colons, exclamation points, commas, semicolons, double quotes, question marks, and forward slashes.

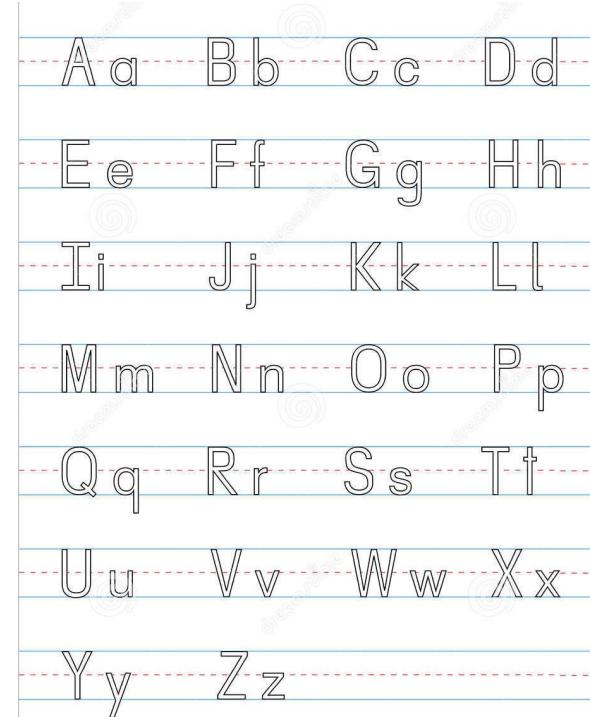
Pré-processamento de documentos

➤ Maiúsculas e Minúsculas

- Letras em maiúsculo ou minúsculo normalmente não têm impacto significante na identificação de termos de índice
- Normalmente todo o texto é convertido para maiúsculas ou minúsculas

Obs: Em alguns casos a semântica pode ficar comprometida

- Banco vs banco.



Pré-processamento de documentos

- **Eliminação de stopwords**

- Palavras muito frequentes entre os documentos de uma coleção não são boas como discriminantes
- Tais palavras são frequentemente chamadas de *stopwords* e geralmente são removidas dos termos de índice
- Exemplos: artigos, preposições, conjunções
 - o, a, portanto, logo, pois, como...



Pré-processamento de documentos

- **Eliminação de stopwords**

- Eliminar *stopwords* reduz显著mente o tamanho do índice

Obs: Apesar dos benefícios, a eliminação de stopwords pode reduzir a revocação

- Uma busca por “ser ou não ser”

Pré-processamento de documentos

Stem*ming*

- Pode ser que um documento possua apenas uma variação da palavra procurada. Ex: Plurais, gerúndios e sufixos.
- Substituir as palavras pelos seus respectivos stems (radicais) pode superar parcialmente esse problema
- Stem é a porção de uma palavra que resta após a remoção de afixos (prefixos e sufixos)
 - casa, casinha, casinhas, casas = casa

Pré-processamento de documentos

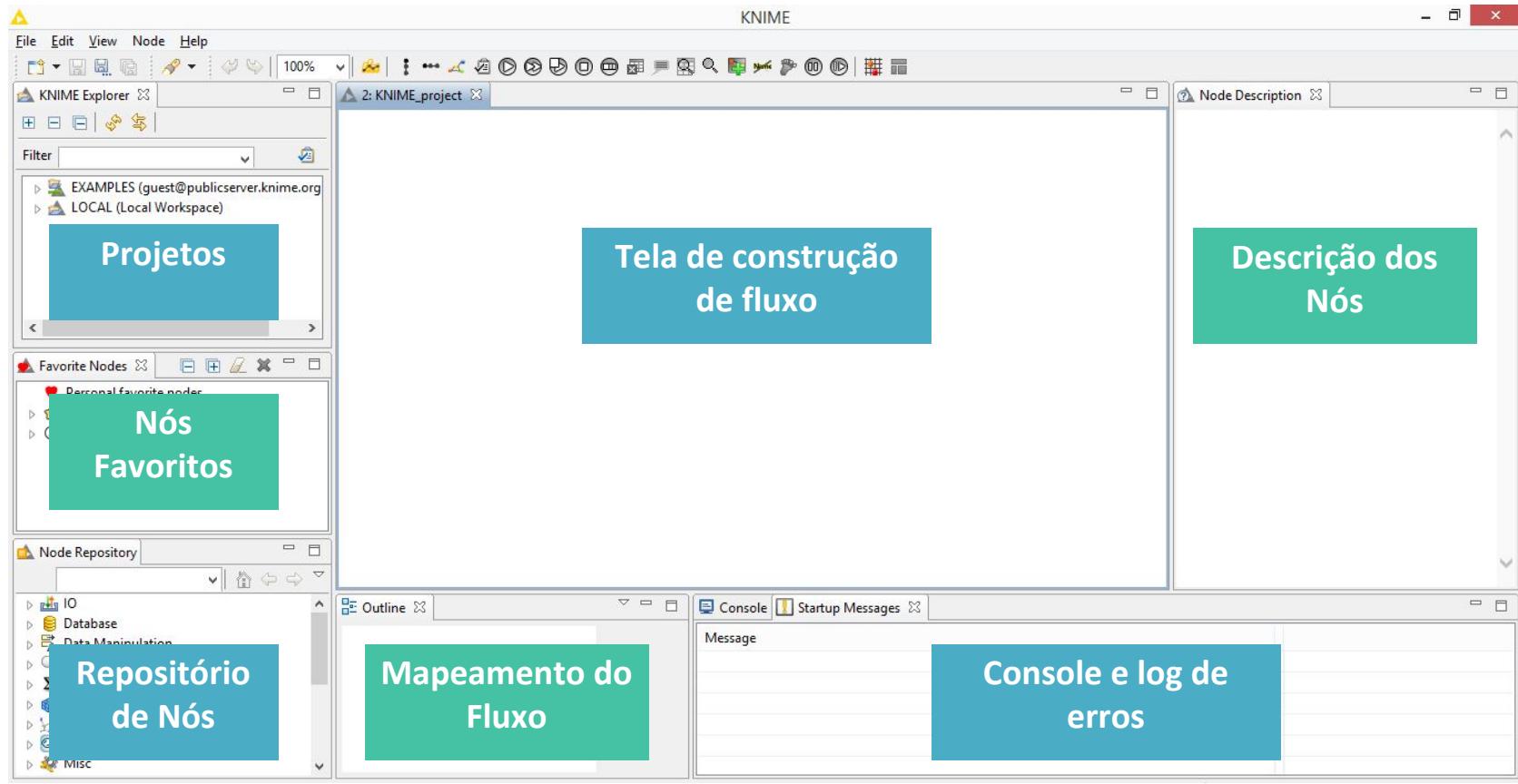
➤ Stemming

- Também reduz o tamanho da estrutura de indexação
- Existem controvérsias na literatura sobre os benefícios do stemming na performance da recuperação
- Obs: Em determinadas línguas o stemming pode ser difícil de se realizar, exigindo buscas em tabelas externas e algoritmos específicos

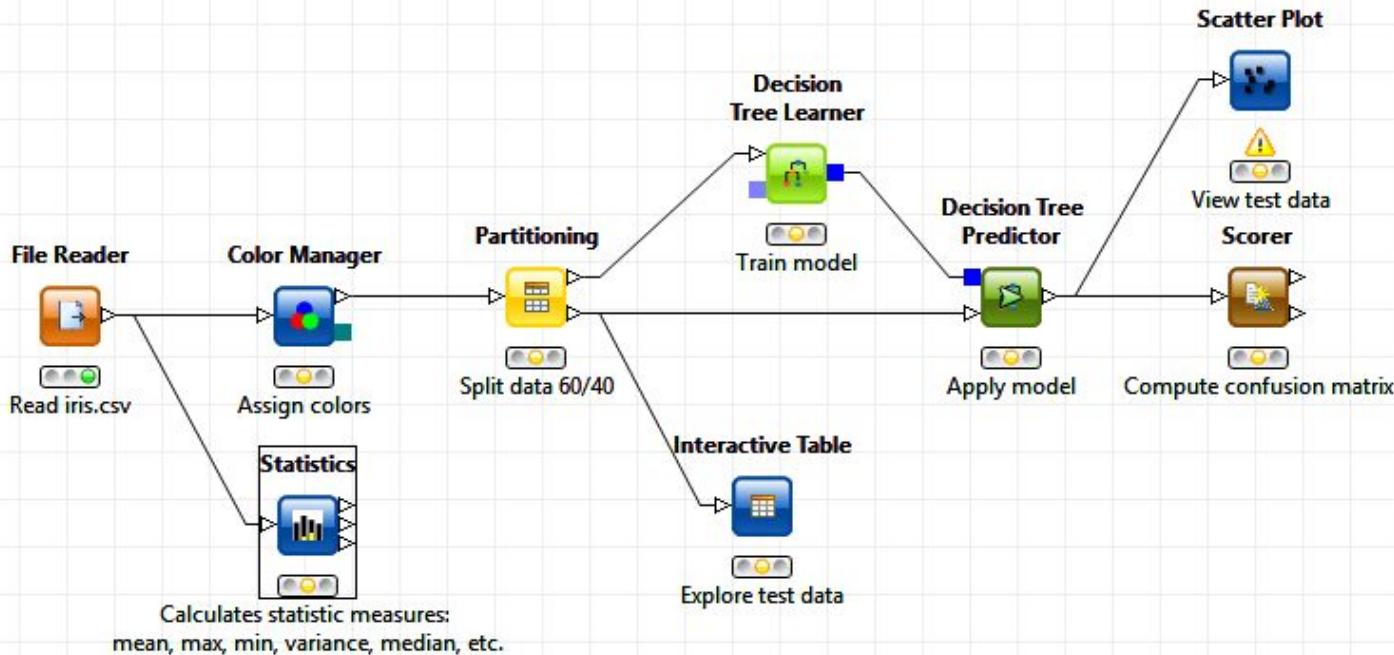
Introdução ao KNIME



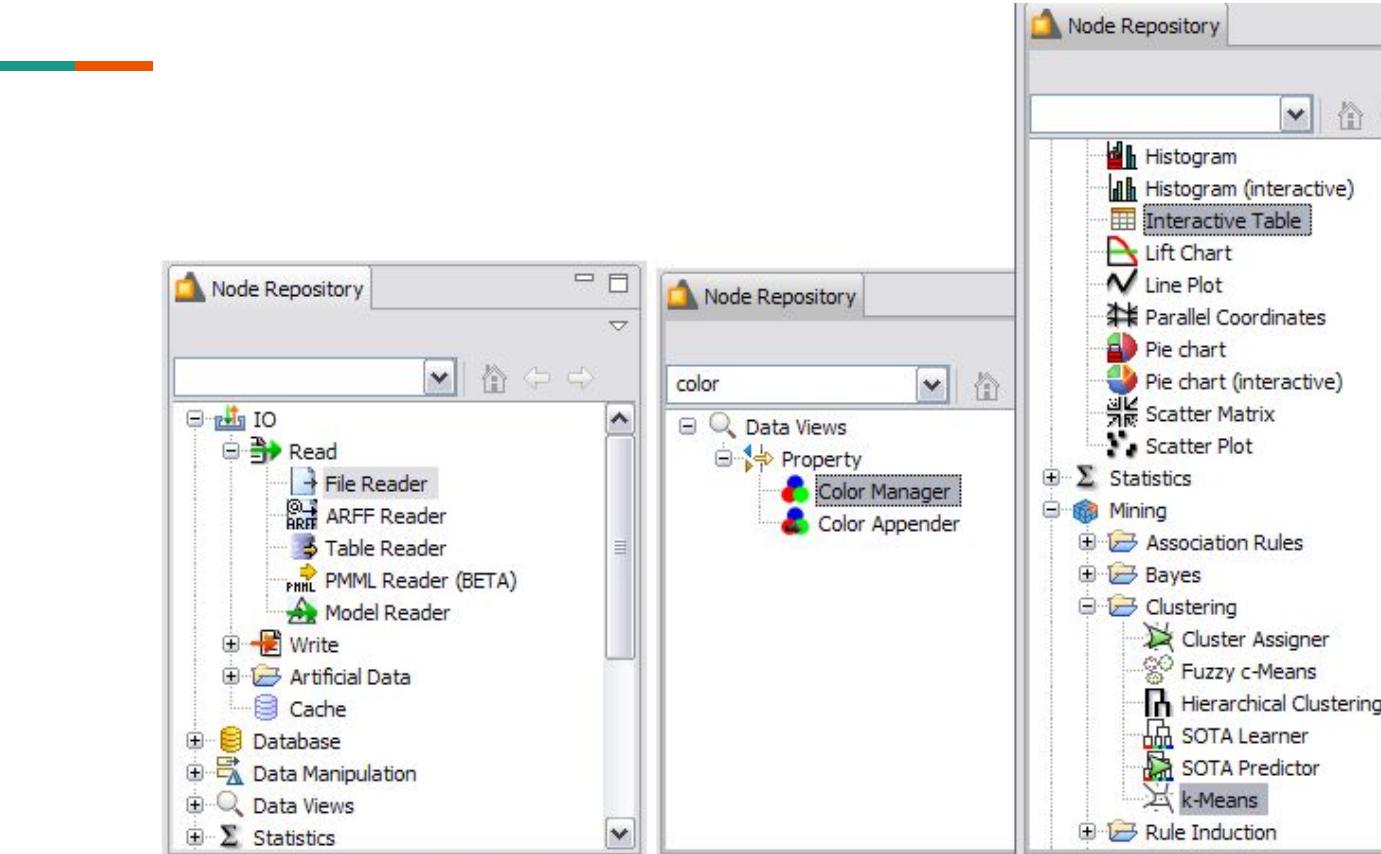
Tela Principal



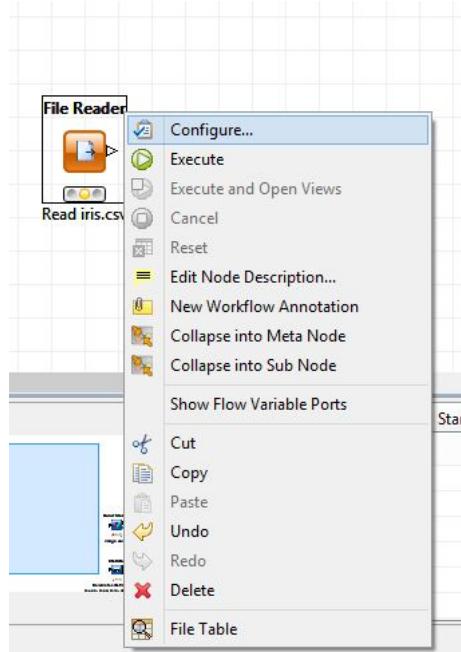
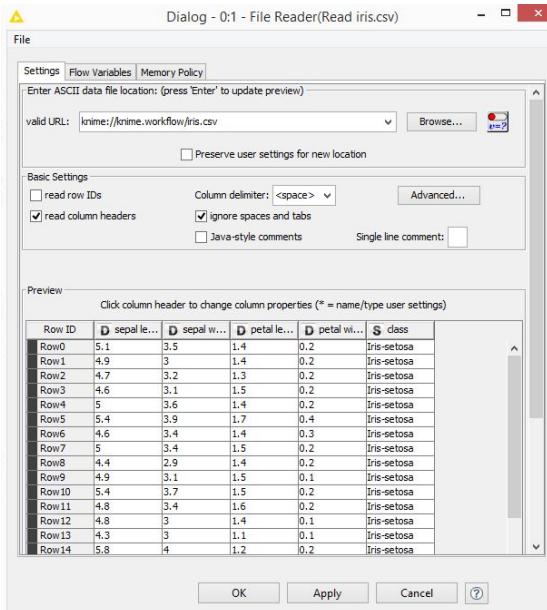
Exemplo de Fluxo



Adição de um componente (Node)



Configuração de um Nó



File Reader

This node can be used to read data from an ASCII file or URL location. It can be configured to read various formats. When you open the node's configuration dialog and provide a filename, it tries to guess the reader's settings by analyzing the content of the file. Check the results of these settings in the preview table. If the data shown is not correct or an error is reported, you can adjust the settings manually (see below).

The file analysis runs in the background and can be cut short by clicking the "Quick scan", which shows if the analysis takes longer. In this case the file is not analyzed completely, but only the first fifty lines are taken into account. It could happen then, that the preview appears looking fine, but the execution of the File Reader fails, when it reads the lines it didn't analyze. Thus it is recommended you check the settings, when you cut an analysis short.

Dialog Options

ASCII file location

Enter a valid file name or URL. When you press ENTER, the file is analyzed and the settings pre-set. You can also choose a previously read file from the drop-down list, or select a file from the "Browse..." dialog.

Preserve user settings

If checked, the checkmarks and column names/types you explicitly entered are preserved even if you select a new file. By default, the analyzer starts with fresh default settings for each new file location.

Execução de um Nó

The screenshot shows the KNIME Analytics Platform interface. The main window displays a workflow titled "Example Workflow". A context menu is open over a "File Reader" node, with the "Execute" option highlighted. To the right, the "Node Description" panel is open for the "File Reader" node, providing documentation and configuration options. The interface includes various toolbars, a file browser, and other panels like "Outline" and "Startup Messages".

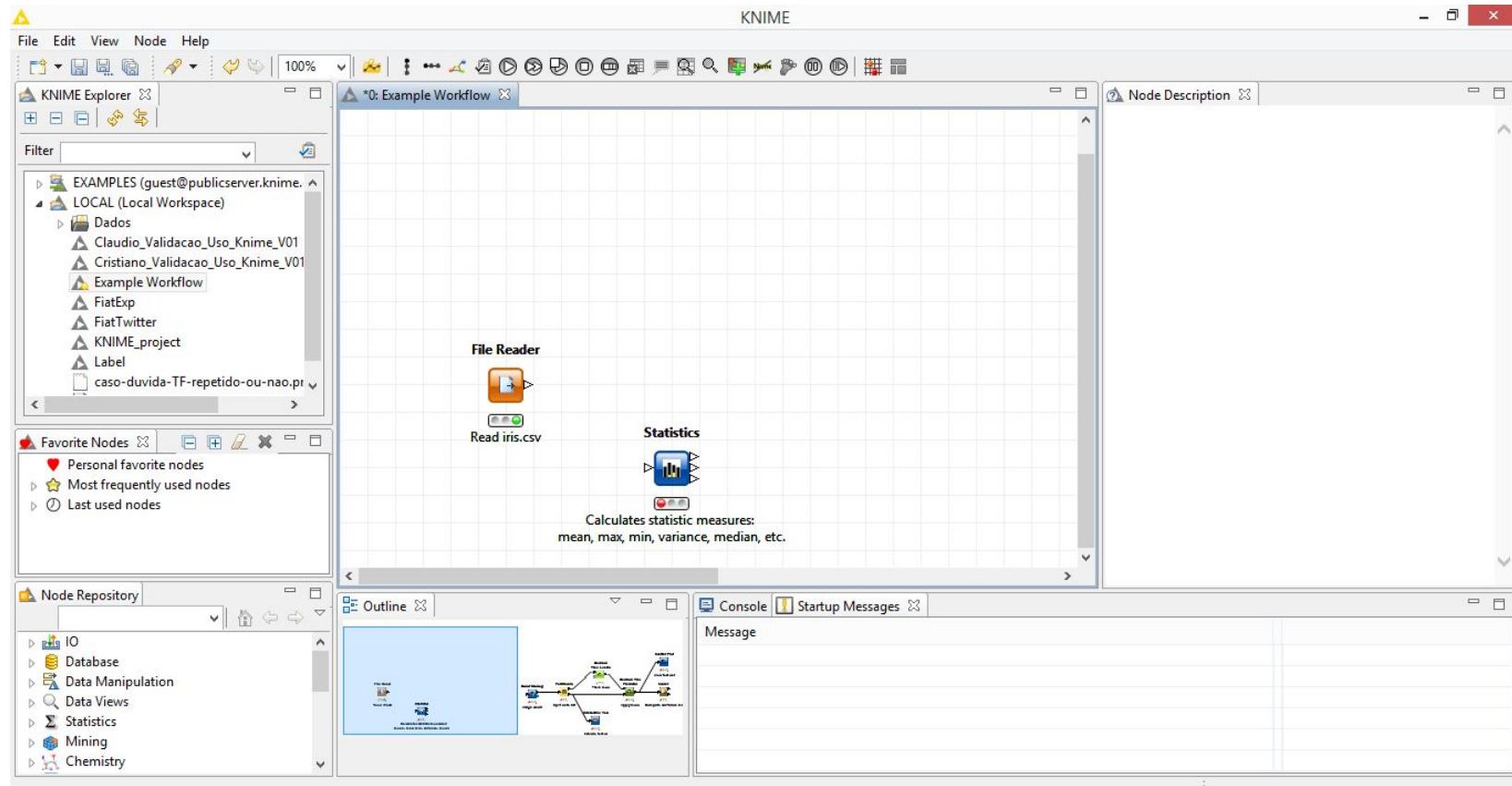
File Reader

This node can be used to read data from an ASCII file or URL location. It can be configured to read various formats. When you open the node's configuration dialog and provide a filename, it tries to guess the reader's settings by analyzing the content of the file. Check the results of these settings in the preview table. If the data shown is not correct or an error is reported, you can adjust the settings manually (see below).

The file analysis runs in the background and can be cut short by clicking the "Quick scan", which shows if the analysis takes longer. In this case the file is not analyzed completely, but only the first fifty lines are taken into account. It could happen then, that the preview appears looking fine, but the execution of the File Reader fails, when it reads the lines it didn't analyze. Thus it is recommended you check the settings, when you cut an analysis short.

Dialog Options

Conexão entre Nós



Conexão entre Nós

The screenshot displays the KNIME Analytics Platform interface, which includes the following components:

- KNIME Explorer**: Shows the project structure with nodes like "EXAMPLES", "LOCAL (Local Workspace)", and "Dados".
- Example Workflow**: The main workspace where a workflow is being built. It consists of a "File Reader" node (with the configuration "Read iris.csv") connected to a "Statistics" node. A tooltip for the "Statistics" node states: "Calculates statistic measures: mean, max, min, variance, median, etc."
- Node Description**: A panel on the right providing detailed information about the selected "File Reader" node. It includes a title "File Reader" and a description: "This node can be used to read data from an ASCII file or URL location. It can be configured to read various formats. When you open the node's configuration dialog and provide a filename, it tries to guess the reader's settings by analyzing the content of the file. Check the results of these settings in the preview table. If the data shown is not correct or an error is reported, you can adjust the settings manually (see below).". It also notes: "The file analysis runs in the background and can be cut short by clicking the "Quick scan", which shows if the analysis takes longer. In this case the file is not analyzed completely, but only the first fifty lines are taken into account. It could happen then, that the preview appears looking fine, but the execution of the File Reader fails, when it reads the lines it didn't analyze. Thus it is recommended you check the settings, when you cut an analysis short."
- Dialog Options**: A section within the Node Description panel titled "Dialog Options" with a dotted underline.
- Outline**: A panel showing the hierarchical structure of the workflow.
- Console**: A panel displaying "Startup Messages" and a "Message" log area.
- Node Repository**: A panel listing categories of nodes: IO, Database, Data Manipulation, Data Views, Statistics, Mining, and Chemistry.

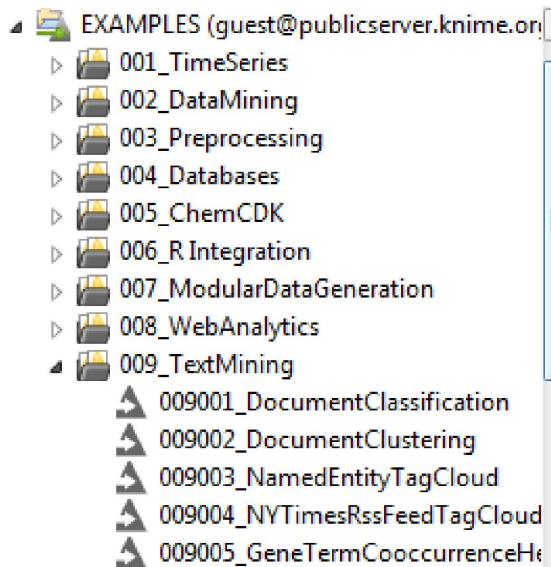
Import/Export de Projetos

The image shows two windows from the KNIME software interface:

- Import Window:** A dialog titled "Import" with the sub-section "KNIME workflow projects". It contains two radio button options: "Select root directory:" (selected) and "Select archive file:". Below these are "Browse..." buttons. A "Projects:" list area is present with buttons for "Select All", "Deselect All", and "Refresh". At the bottom is a checked checkbox "Copy projects into workspace" and a footer with buttons for "< Back", "Next >", "Finish", and "Cancel".
- Export Window:** A dialog titled "Export" with the sub-section "KNIME Workflow project". It has a "Select project to export:" field containing "KNIME_project" with a "Select..." button. An "Export file name (zip):" field contains "C:\export.zip" with a "Select..." button. Under "Options", there is a checked checkbox "Exclude data from export.". At the bottom are "Finish" and "Cancel" buttons.

Workflows de Exemplo

O servidor público do KNIME oferece workflows de exemplo para diversos tipos de tarefas.





KNIME - Similaridade de Docs

KNIME - Similaridade de Docs

❖ Similaridade de documentos

- **Fluxo: 2-SimilaritySearch-pratica-KNIME3**
- Prática em **similaridade por frequência de termos**
- Exercício utilizando **TF-IDF**
- **Faça o download do Fluxo do Aluno (ZIP)**

KNIME - Similaridade de Docs

- ❖ Acesso aos materiais da disciplina...

<https://goo.gl/ps88N4>



Estrutura de Documentos

❖ Atributos

- Texto
- Autor
- Fonte
- Categoria
- Tipo

 Document
"Aschenputtel"
"Dornröschen"
"Frau Holle"
"Hans im Glück"

Lembre-se dos **documentos** em R!

Leitura de Documentos

Utilizaremos uma base previamente coletada e transformada no tipo **Document** do KNIME

Na próxima aula veremos como coletar e criar documentos novos!

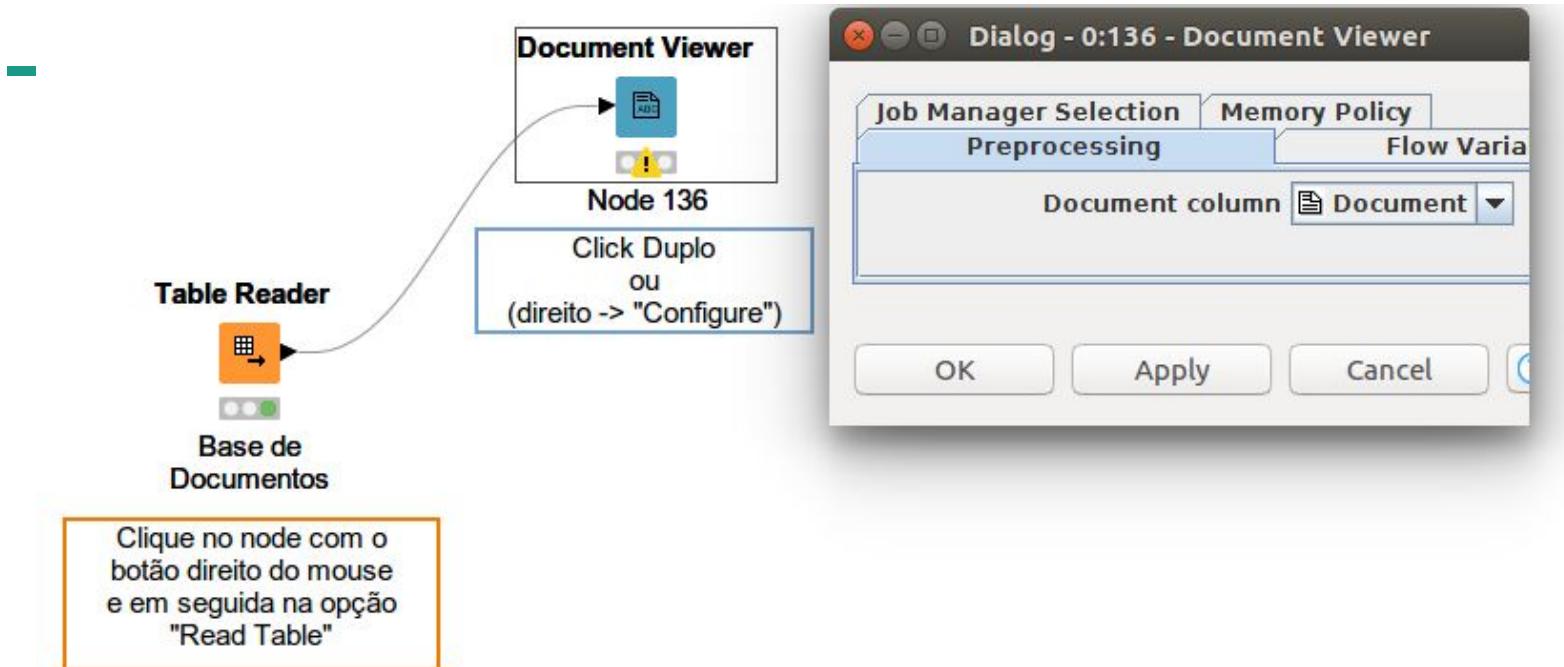


A screenshot of the KNIME 'Table Reader' window titled 'Read table - 0:124 - Table Reader (Base de...)'.

The window displays a table with 21 rows, labeled 'Table "default" - Rows: 215'. The columns are 'Row ID' and 'Document'.

Row ID	Document
Row0	"drakam"
Row1	"monstromedieval"
Row2	"thais_aoliveira"
Row3	"EricRaupp"
Row4	"glaubermacario"
Row5	"leiaorgawa"
Row6	"Murilinho4269"
Row7	"GomesMazurca"
Row8	"GeGe_Crazy_One"
Row9	"danielgeneralli"
Row10	"Pedrexox23"
Row11	"ananaulabertoni"

Visualização dos Documentos



Sempre que necessário, podemos visualizar a lista de documentos com o **Document Viewer**.

Visualização dos Documentos

The diagram illustrates the process of visualizing documents. It starts with a **Table Reader** node (orange icon) connected to a **Document Viewer** node (blue icon). The **Document Viewer** node is labeled **Node 136**. A callout box indicates to "Click Duplo ou (direito -> "Configure")". Another callout box provides instructions: "Para visualizar os documentos: "Execute and Open Views" ou "Execute" -> "View:Document View"".

Table Reader

Base de Documentos

Clique no node com o botão direito do mouse e em seguida na opção "Read Table"

Document Viewer

Node 136

Para visualizar os documentos:
"Execute and Open Views" ou
"Execute" -> "View:Document View"

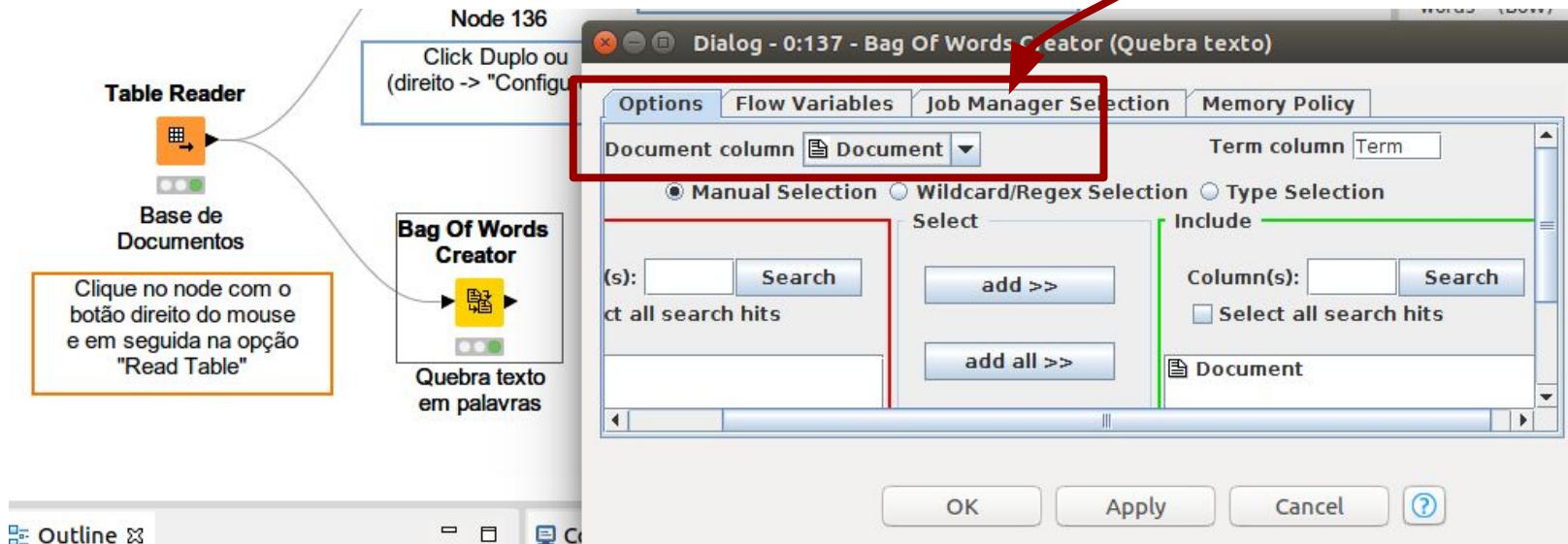
Document View - 0:136 - Document Viewer

#	Document Title	Authors	Source	Category
1	00110011o	- ジギリ	00110011o	crz, sp
2	AlinedeOz	Aline de Oz	AlinedeOz	Brasil
3	AnJuliaOliveira	Ana Julia♥	AnJuliaOliveira	
4	BennyCohen10	Benny Cohen	BennyCohen10	Belo Horizonte
5	BiaRosenburg	Beatriz Rosenburg	BiaRosenburg	Rio de Janeiro, Brazil
6	Brasangola_	- Brasangola_	Brasangola_	Indaiatuba-SP
7	CarlosNilson	Carlos Nilson Jr	CarlosNilson	
8	CarlosNilson	Carlos Nilson Jr	CarlosNilson	
9	Carolina_MariaF	Carolina Fernandes	Carolina_MariaF	São Paulo, Brazil

Basta clicar duas vezes no documento desejado para abri-lo

Bag of Words

Atenção para a coluna que será processada!



Com o *Bag of Words* podemos quebrar o texto “bruto” dos documentos em palavras

Bag of Words

The screenshot shows the KNIME interface. On the left, there is a 'Bag Of Words Creator' node with a yellow icon and three colored dots below it. A callout bubble points from this node to the text 'Quebra texto em palavras'. To the right is a table titled 'Documents output table - 0:137 - Bag Of Words'. The table has columns 'Row ID', 'Document', and 'Term'. The data rows show various terms extracted from documents, such as 'drakam', 'https[]', and URLs like '://t.co/L1zea9YeWH[]'. The table has 3565 rows and 2 columns.

Row ID	Document	Term
Row0	"drakam"	drakam[]
Row1	"drakam"	https[]
Row2	"drakam"	://t.co/L1zea9YeWH[]
Row3	"drakam"	Correio[]
Row4	"drakam"	Brazilense[]
Row5	"drakam"	descobre[]
Row6	"drakam"	a[]
Row7	"drakam"	verdadeira[]
Row8	"drakam"	identidade[]

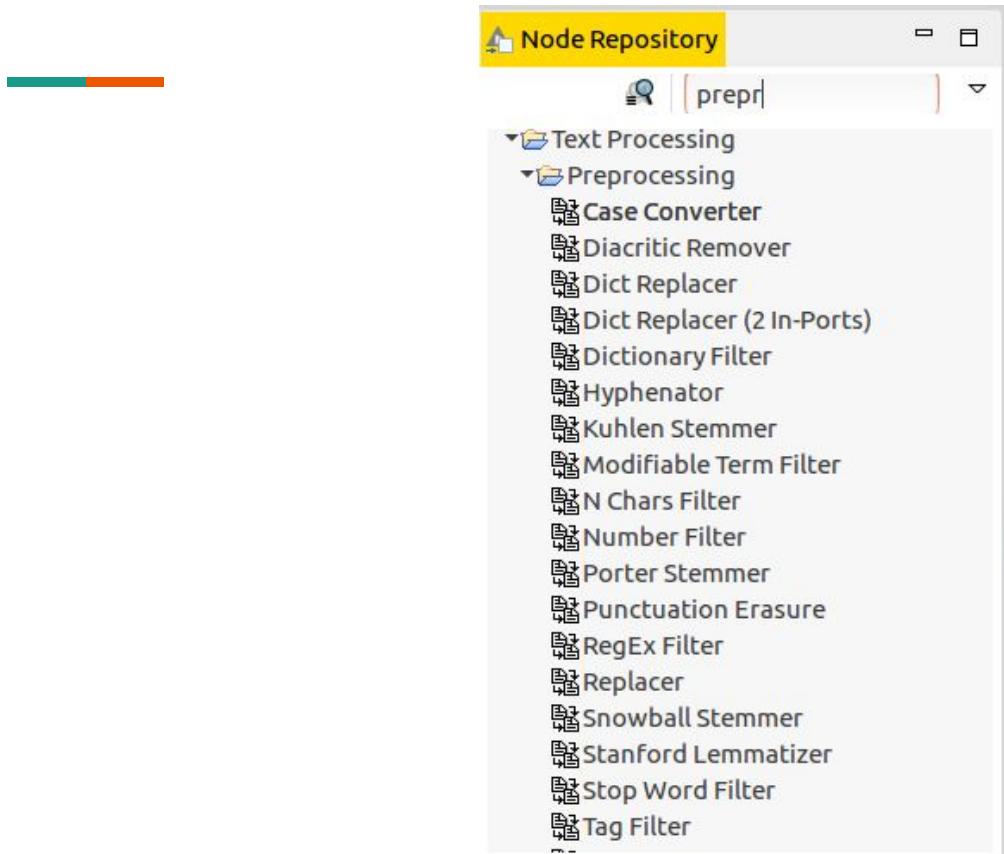
Retorna uma tabela com **termos em cada documento**.
Observe na lista de termos que o texto não está normalizado
(pontuações, artigos, mix de minúsculas/maiúsculas ...)

Survivor Guide

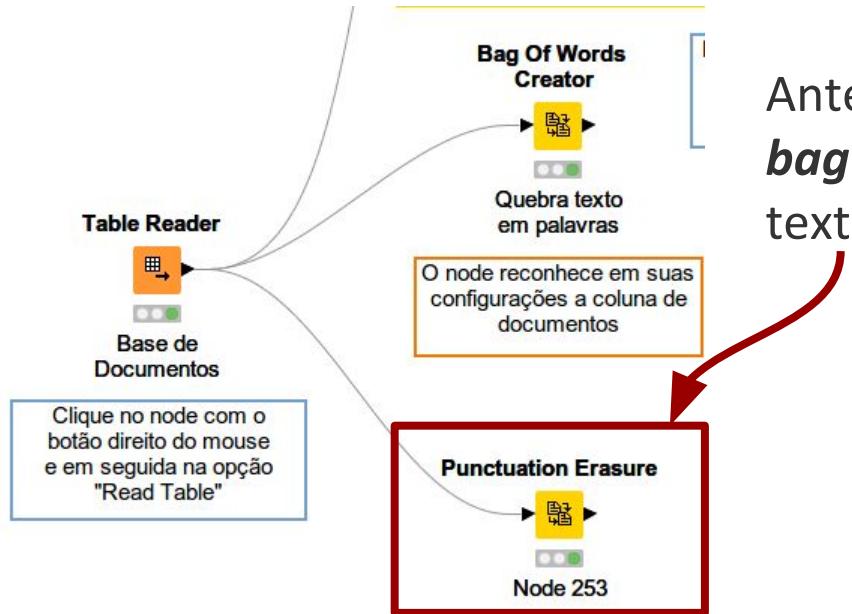
Você já deve ser capaz de **localizar** nodes por nome, abrir menu de **configurações**, **executar** e **visualizar** resultados.

Caso contrário **SOS Professor!**

Opções de Pré-processamento



Remoção de Pontuação



Antes de visualizar os termos com o *bag of words*, vamos **preprocessar** o texto

Com o *Punctuation Erasure* podemos **apagar** caracteres de pontuação encontrados nos **documentos**

Remoção de Pontuação

The diagram illustrates the KNIME workflow for removing punctuation from documents. It starts with a 'Bag Of Words Creator' node, followed by a 'Punctuation Erasure' node (Node 253). A callout box indicates that the 'Punctuation Erasure' node recognizes the document column in its configuration. To the right, a screenshot of the 'Punctuation Erasure' dialog shows the 'Preprocessing' tab selected. The 'Document column' dropdown is set to 'Document'. A red box highlights the 'Append column:' field, which contains 'Preprocessed Document'. A red arrow points from this field to a text box stating: 'Por padrão o KNIME adiciona uma coluna com o documento alterado e mantém uma cópia do original'. A callout box also provides instructions: 'Para visualizar a tabela de termos: "Execute" -> "Document output table"'.

Bag Of Words Creator

Quebra texto em palavras

O node reconhece em suas configurações a coluna de documentos

Punctuation Erasure

Node 253

Para visualizar a tabela de termos:
"Execute" -> "Document output table"

Por padrão o KNIME adiciona uma coluna com o documento alterado e mantém uma cópia do original

Dialog - 0:253 - Punctuation Erasure

Job Manager Selection Memory Policy

Preprocessing Flow Variables

Document column: Document

Replace column:

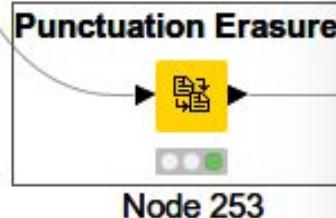
Append column: Preprocessed Document

Ignore unmodifiable flag

Com o *Punctuation Erasure* podemos apagar caracteres de pontuação encontrados nos documentos

Remoção de Pontuação

O node reconhece em suas configurações a coluna de documentos



Preprocessed documents. - 0:253 - Punctuation Eraser

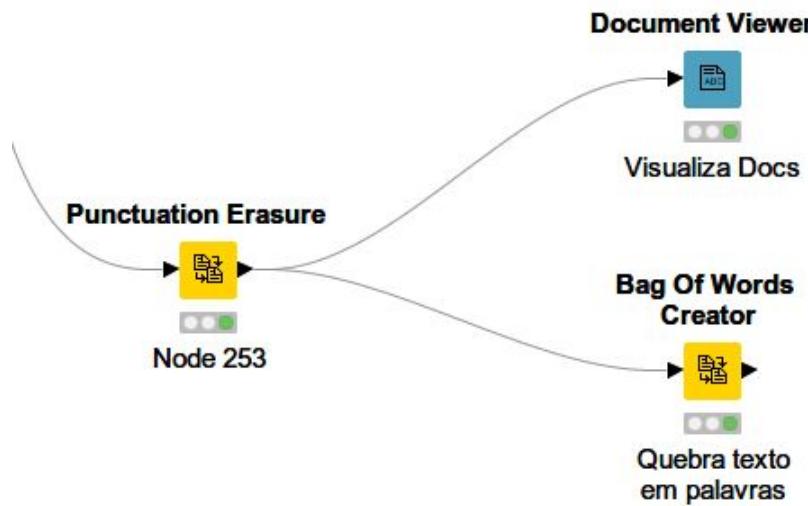
File Hilite Navigation View

Table "default" - Rows: 554 Spec - Columns

Row ID	Document	Preproc...
Row0	"fala_sampai...	"falasampai...
Row1	"MineXis"	"MineXis"
Row2	"KassioFrazao"	"KassioFraz...
Row3	"melodyolive...	"melodyoliv...
Row4	"guilhermebl...	"guilherme...
Row5	"rafaelmach...	"rafaelmac...
Row6	"roodrigoma...	"roodrigom...

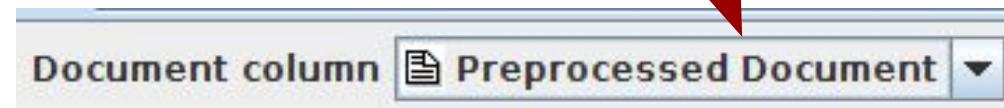
A saída do node (*Botão direito do mouse -> Preprocessed...*) é a lista de documentos **originais (Document)** e a nova lista com os documentos **alterados (Preprocessed)**

Remoção de Pontuação



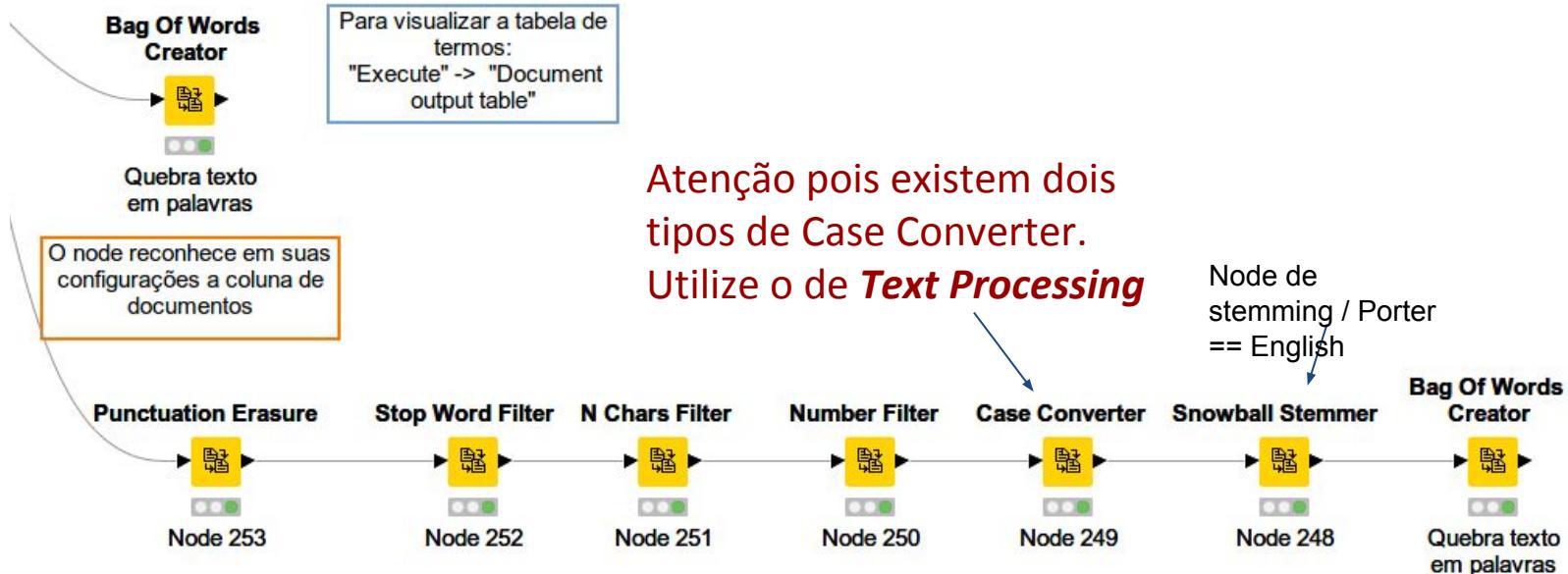
Atenção: Sempre fique atento a coluna em que pretende visualizar ou processar.

Utilizar a coluna errada pode te confundir nos resultados



OPCIONAL: Para conferir alterações nos documentos, você visualizar os documentos com o *Document Viewer*, ou os termos com o *Bag of Words*

Pré-processamento do Texto



Utilizem nodes de pré-processamento para tratar o conjunto de palavras extraído da coleção de textos.
Procurem **investigar as opções** de configurações.

KNIME - Frequência



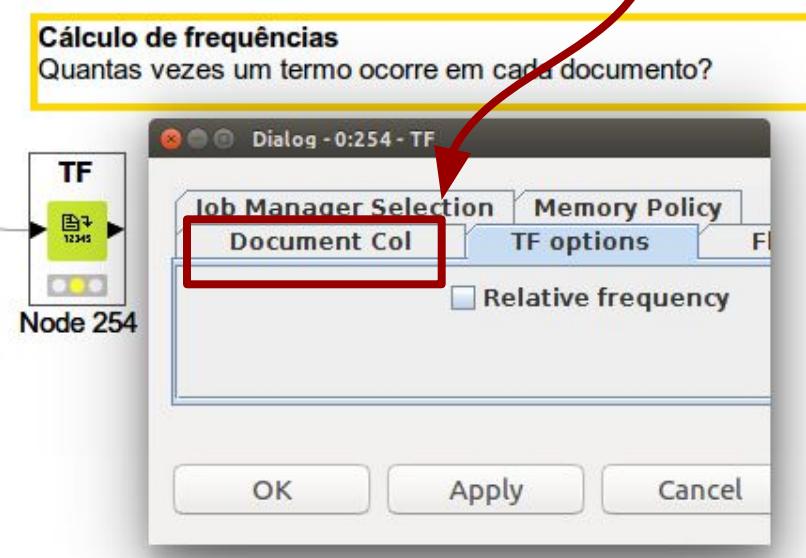
Cálculo de Frequênciā

➤ TF

- TF absoluto = Nº de ocorrências do termo t
- TF relativo = Nº de ocorrências do termo t / Nº de termos
- IDF = $\log(1 + \text{Nº de docs} / \text{Nº docs com termo t})$
- ICF = $\log(1 + \text{Nº de categorias} / \text{Nº de categorias com termo t})$
- **IDF * TF**

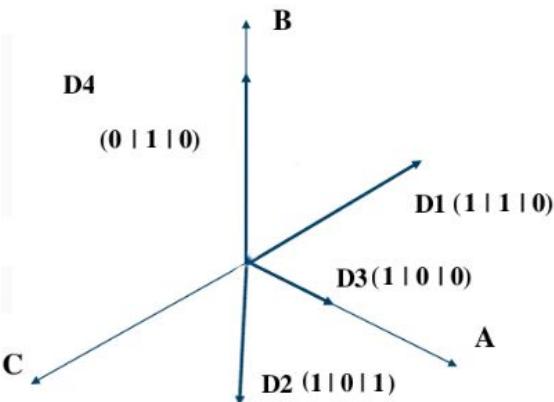
Frequência

Não esqueça de configurar
a origem



Opção de frequência **absoluta** ou frequência **relativa** ao nº total termos do documento.

Vetor de Bits (Freq=0 ou 1)



Por enquanto não vamos utilizar a frequência(TF). Vamos considerar 0 ou 1 como peso dos termos.

Documents output table - 0:9 - Document vector

Spec - Columns: 1077 Properties Flow Variables

Table "default" - Rows: 20

Row ID	Document	D endoc...	D cancer	D autot...	D inflam
1	"Autotaxin is an in...	1	1	1	1
2	"CD14 knockdown ...	0	1	0	1
3	"Cocaine mediate...	0	0	0	1
4	"Early Growth Res...	0	0	0	1
5	"IL6/JAK1/STAT3 sig...	0	1	0	0
6	"Interleukin-6 cont...	0	0	0	0
7	"Knockdown of AD...	0	0	0	0
8	"Lumican overexpr...	0	0	0	0
9	"MTOR regulates t...	0	0	0	0
10	"MicroRNA214 is A...	0	0	0	0
11	"Microenvironment...	0	0	0	0

Dialog - 0:34 - Document vector

Snowball Stemmer

Job Manager Selection Memory Policy

Options Flow Variables

Document column Orig Document

Ignore tags

Bitvector

Vector value Abs

As collection cell

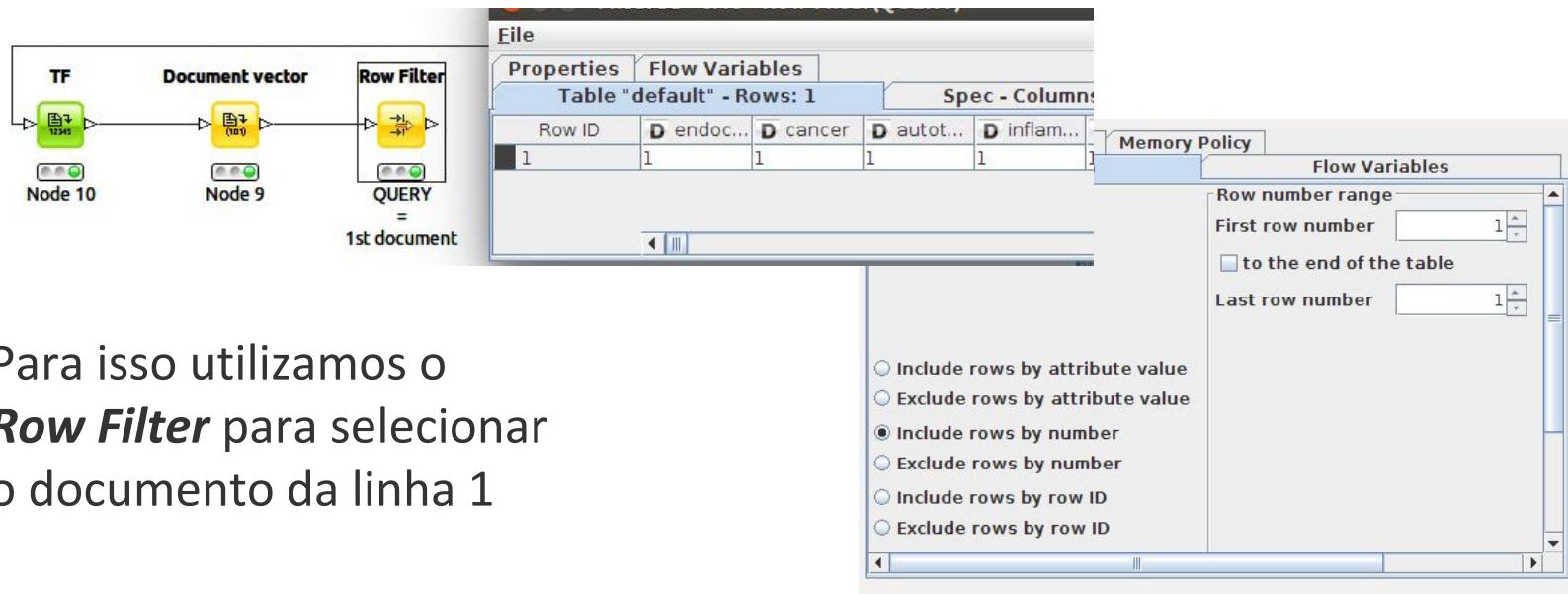
OK Apply Cancel ?

TF Document vector

Node 10 Node 9

Simulação de Consulta

- ❖ Vamos utilizar como consulta um dos documentos e procurar por outros similares a este documento na coleção

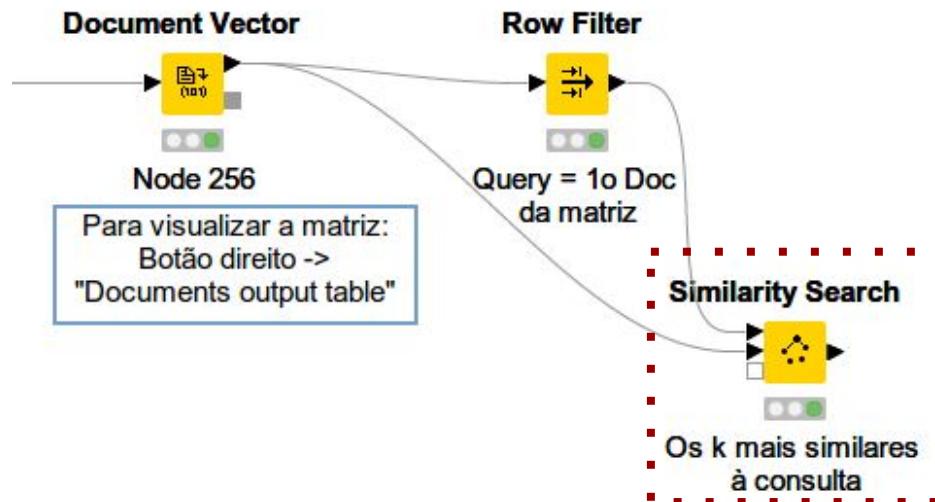


Para isso utilizamos o
Row Filter para selecionar
o documento da linha 1

Similaridade

- ❖ Utilizamos então os vetores de todos os documentos da base

Com o node ***Similarity Search*** podemos utilizar a medida do cosseno para determinar as distâncias entre os documentos.

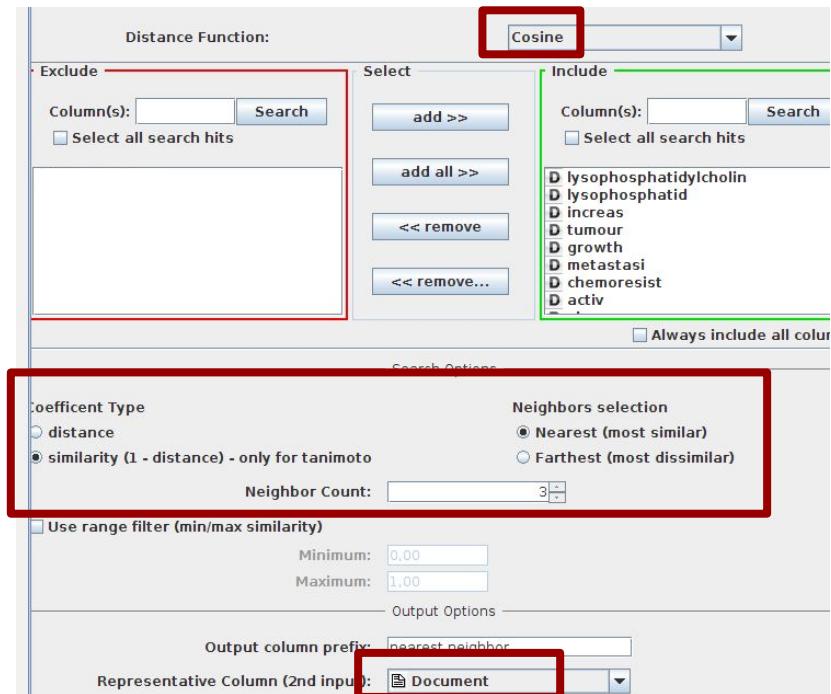


Lembrando que por enquanto utilizamos apenas 0 ou 1 como peso dos termos. Ainda não utilizamos nem TF nem IDF.

Similaridade

❖ Configuração do node *Similarity Search*

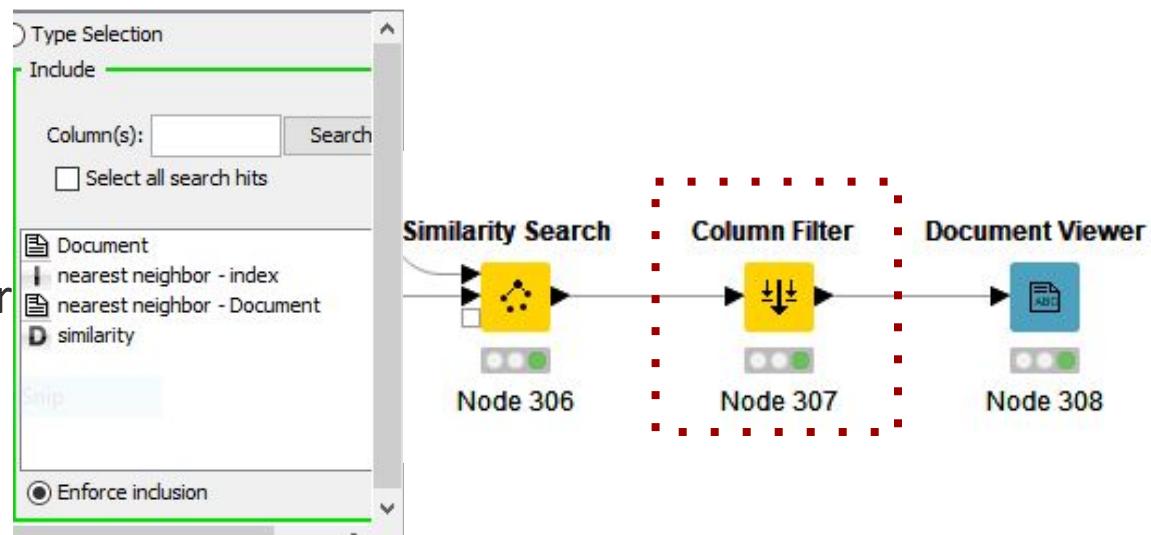
- Configuramos a medida de distância
- Incluímos os termos (colunas da matriz de vetores).
- Escolhemos a medida: similaridade ou distância ? deseja os vizinhos mais próximos ou distantes?
- Qual coluna representa a coleção de documentos (col “Document”)



Resultados

❖ Filtragem de resultados

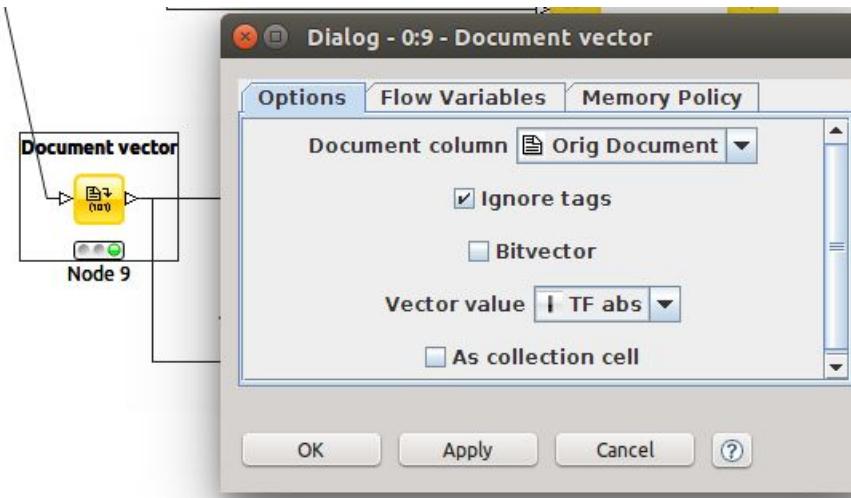
Para reduzir o número de colunas e melhorar a visualização da tabela de similaridades, basta utilizar o *Column Filter*



Em seguida é possível visualizar os documentos buscados no topo do ranking

Similaridade

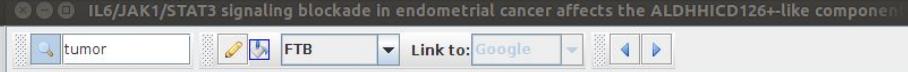
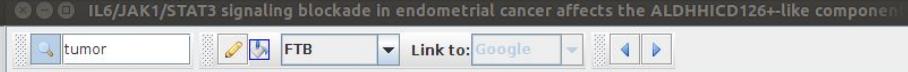
- ❖ Agora, utilize a medida do cosseno com base em TF



- Houve alteração no ranking e no peso da similaridade?

Similaridade

❖ Vantagem do stemming (**tumour** ~ **tumor**)

	
<p>Autotaxin is an inflammatory mediator and therapeutic target in thyroid cancer.</p> <p>JOURNAL_TITLE Endocrine-related cancer</p> <p>TITLE Autotaxin is an inflammatory mediator and therapeutic target in thyroid cancer.</p> <p>ABSTRACT Autotaxin is a secreted enzyme that converts extracellular lysophosphatidylcholine to lysophosphatidate. In cancers, lysophosphatidate increases tumour growth, metastasis and chemoresistance by activating six G-protein coupled receptors. We examined >200 human thyroid biopsies. Autotaxin expression in metastatic deposits and primary carcinomas was 4-10-fold higher than in benign neoplasms or normal thyroid tissue. Autotaxin immunohistochemical staining was also increased in benign neoplasms with leukocytic infiltrations. Malignant tumours were distinguished from benign tumours by high tumour autotaxin, lysophosphatidate levels and inflammatory mediators including IL1β, L6, IL8, GMCSF, TNFa, CCL2, CXCL10 and PDGF-AA. We determined the mechanistic explanation for these results and revealed a vicious regulatory cycle in which lysophosphatidate increased the secretion of 16 inflammatory modulators in papillary thyroid cancer cultures. Conversely, treating cancer cells with 10 inflammatory cytokines and chemokines or PDGF-AA and PDGF-BB increased autotaxin secretion. We confirmed that this autotaxin/inflammatory cycle occurs in two SCID mouse models of papillary thyroid cancer by blocking lysophosphatidate signalling using the autotaxin inhibitor, DNO-8430506. This decreased the levels of 16 inflammatory mediators in the tumours and this was accompanied by a 50-60% decrease in tumour volume. This resulted from a decreased mitotic index for the cancer cells and decreased levels of VEGF and angiogenesis in the tumours. Our results demonstrate that the autotaxin/inflammatory cycle is a focal point for driving malignant thyroid tumour progression and possibly treatment resistance. Inhibiting autotaxin activity provides an effective and novel strategy for decreasing the inflammatory phenotype in thyroid carcinomas, which should complement other treatment modalities.</p>	

Exercício - Similaridade

Multiplicando colunas com o Nó Math Formula

The screenshot shows the configuration of a 'Math Expression' node. In the 'Column List' section, 'column1' and 'column1 (#1)' are selected. In the 'Function' list, 'x * y' is highlighted. In the 'Expression' field, the formula '\$column1\$ * \$column1 (#1)\$' is entered. The 'Description' panel indicates 'Multiplication between'.

Sendo uma tabela de entrada:

Row ID	column1	column...
Row0	2	3
Row1	4	2
Row2	6	1

É possível expressões matemáticas como:
\$coluna_a\$ * \$coluna_b\$

Table "default" - Rows: 3			Spec
Row ID	column1	column...	produto
Row0	2	3	6
Row1	4	2	8
Row2	6	1	6

Obs: O “\$” é apenas um delimitador de nomes de colunas

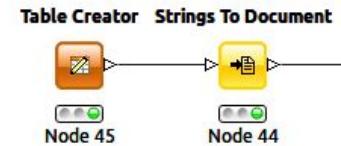
Exercício - Similaridade

- ❖ **Tarefa** - Utilize a medida do cosseno no fluxo de similaridade com base em **TF-IDF**

Procure avaliar mudanças nos resultados

Por exemplo:

- Houve alteração no ranking e no peso da similaridade?
- Você considera esse resultado mais, ou menos relevante?
- **Desafio** - Considerando apenas o TF, qual o documento mais relevante para a consulta ***“Torcedor Misterioso do Brasil”***
 - Dica: Criar uma string e transformar em um documento.Criando uma linha na tabela com o texto inteiro e usando o node ***Strings to Document***



Similaridade

❖ Aplicações

- Ranking de similaridade
- Classificação ou categorização de documentos
- Clustering de documentos

Projeto Final

❖ Utilidade do exercício para o projeto final

- Pré-processamento do texto
- Mostrar **documentos mais similares** em relação a outro documento
- Resgatar documentos **mais relevantes** a determinada consulta de termos específicos
- Agrupar documentos (clustering)
- Classificar documentos (se houver um conjunto conhecido de documentos com uma classe atribuída a cada um)

A disciplina - RI

❖ Plano de Ensino

- **Unidade 01:** Conceitos de inteligência competitiva e coletiva, crowdsourcing e redes sociais. Recuperação da informação e Máquinas de busca. Desafios da Mineração na web e nas redes sociais. Exemplos de projetos da disciplina.

- **Unidade 02:** Algoritmos e soluções para problemas de busca e extração de informação da WWW. Ferramenta e prática de processamento textual e recuperação de informação.

Esse assunto era o que você imaginava?



