

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.

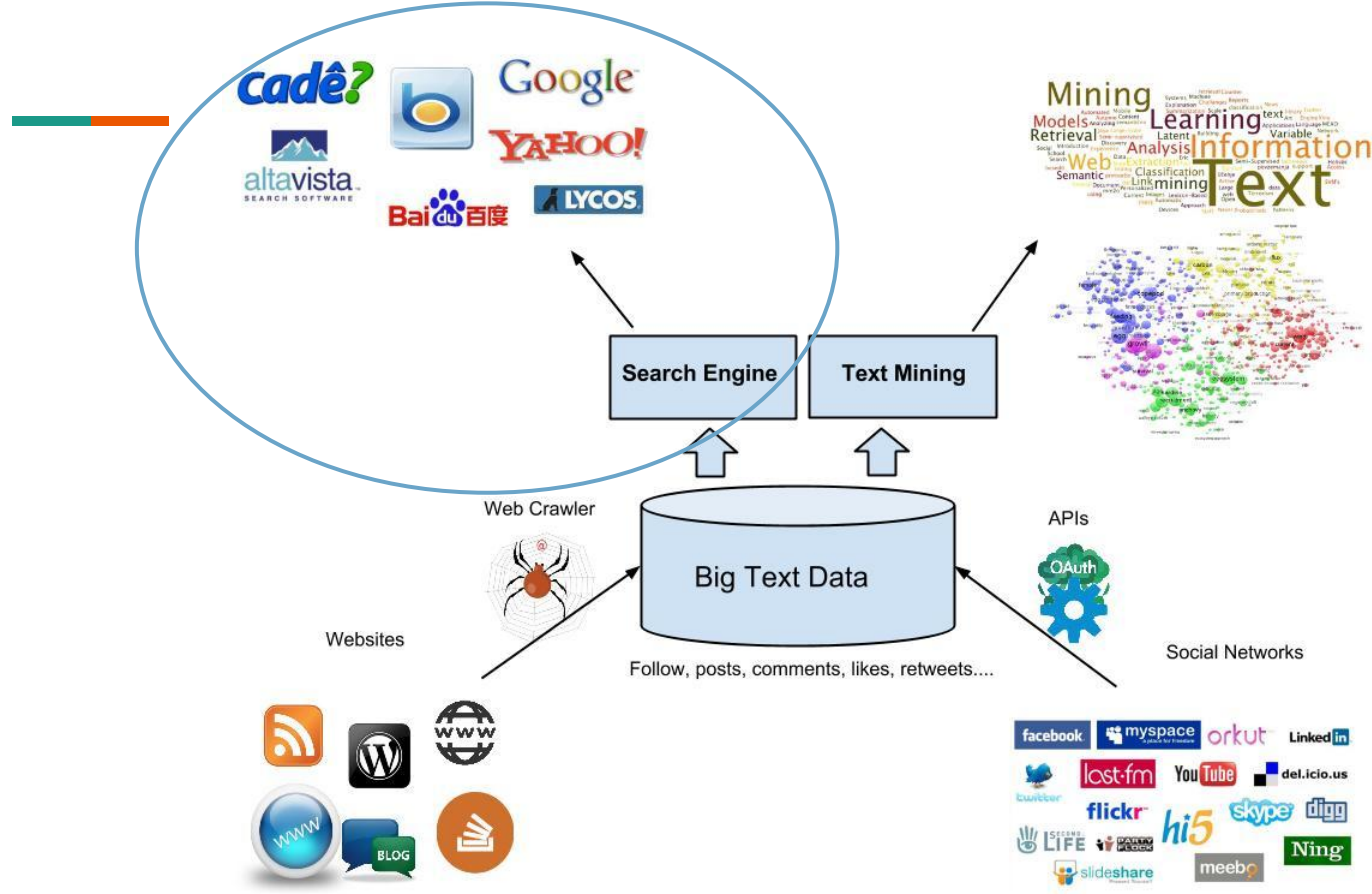
Projeto Final

- ❖ 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**
 - 15° dia após a última aula às 23:59hrs

Busca Textual e Similaridade



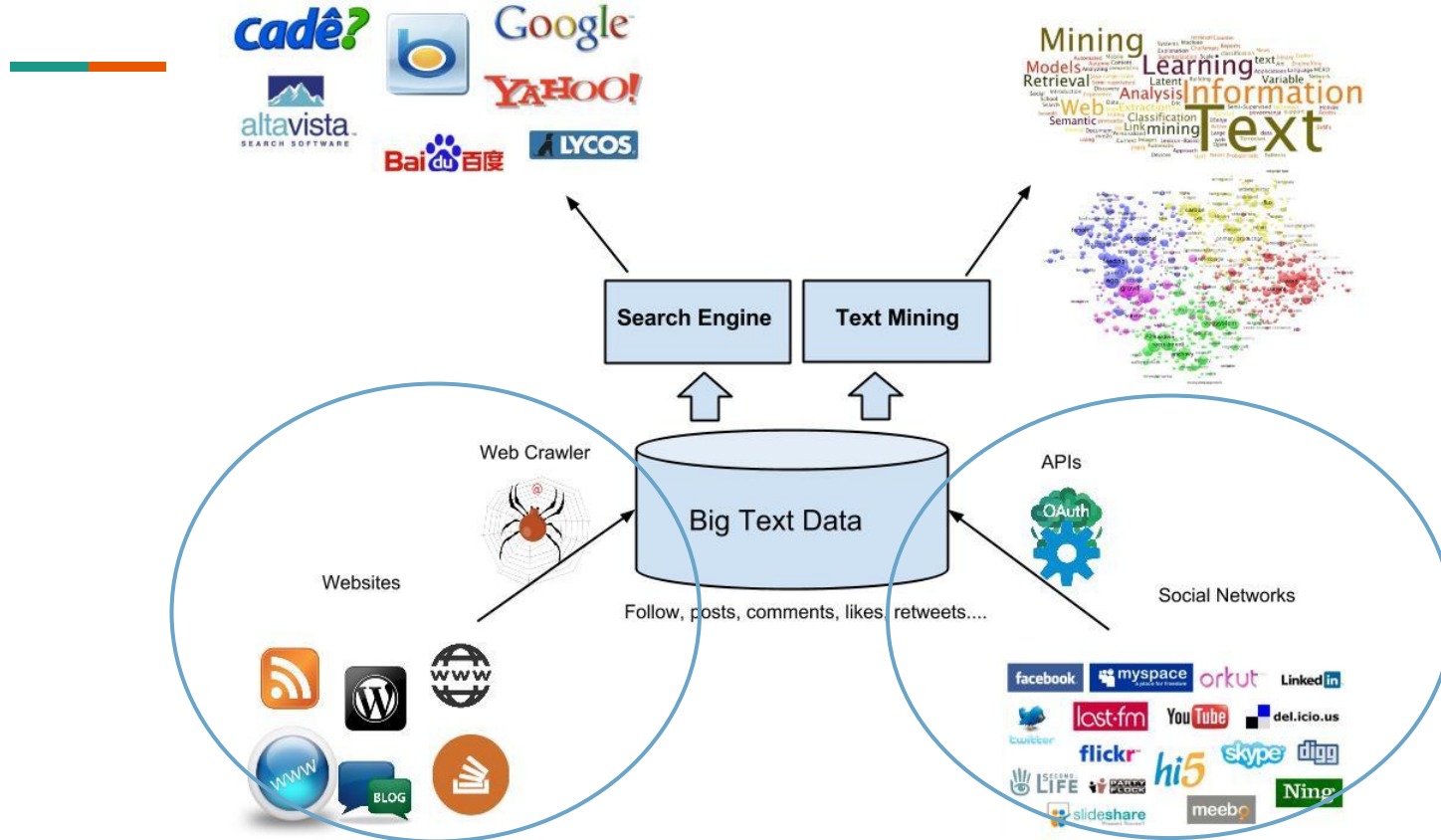
Mineração da Web e Redes Sociais



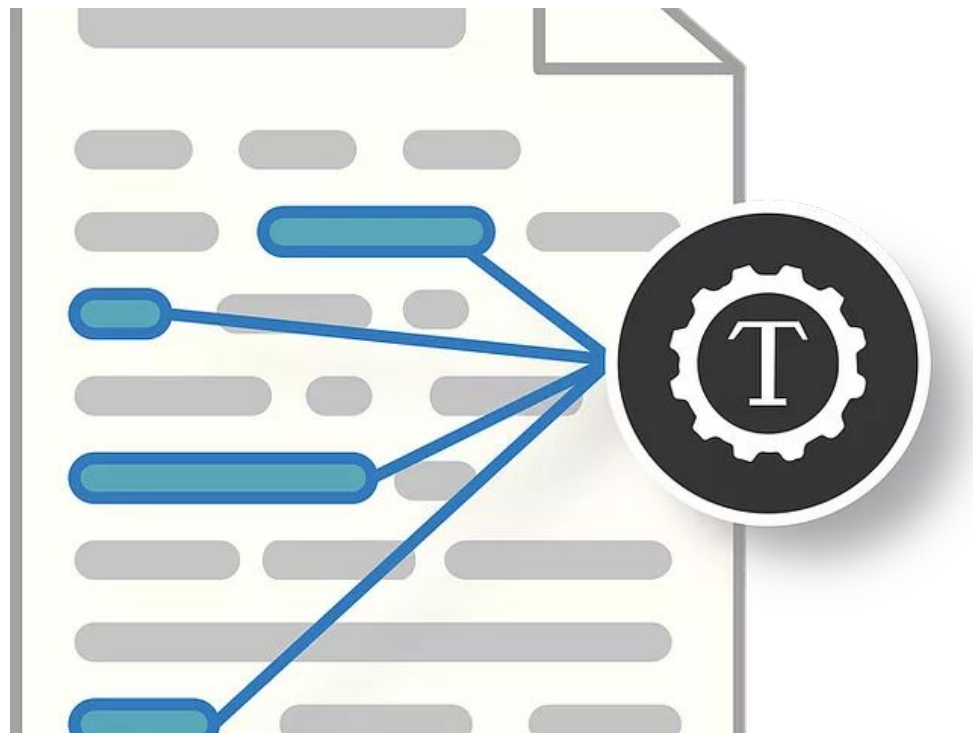
Coleta de Dados



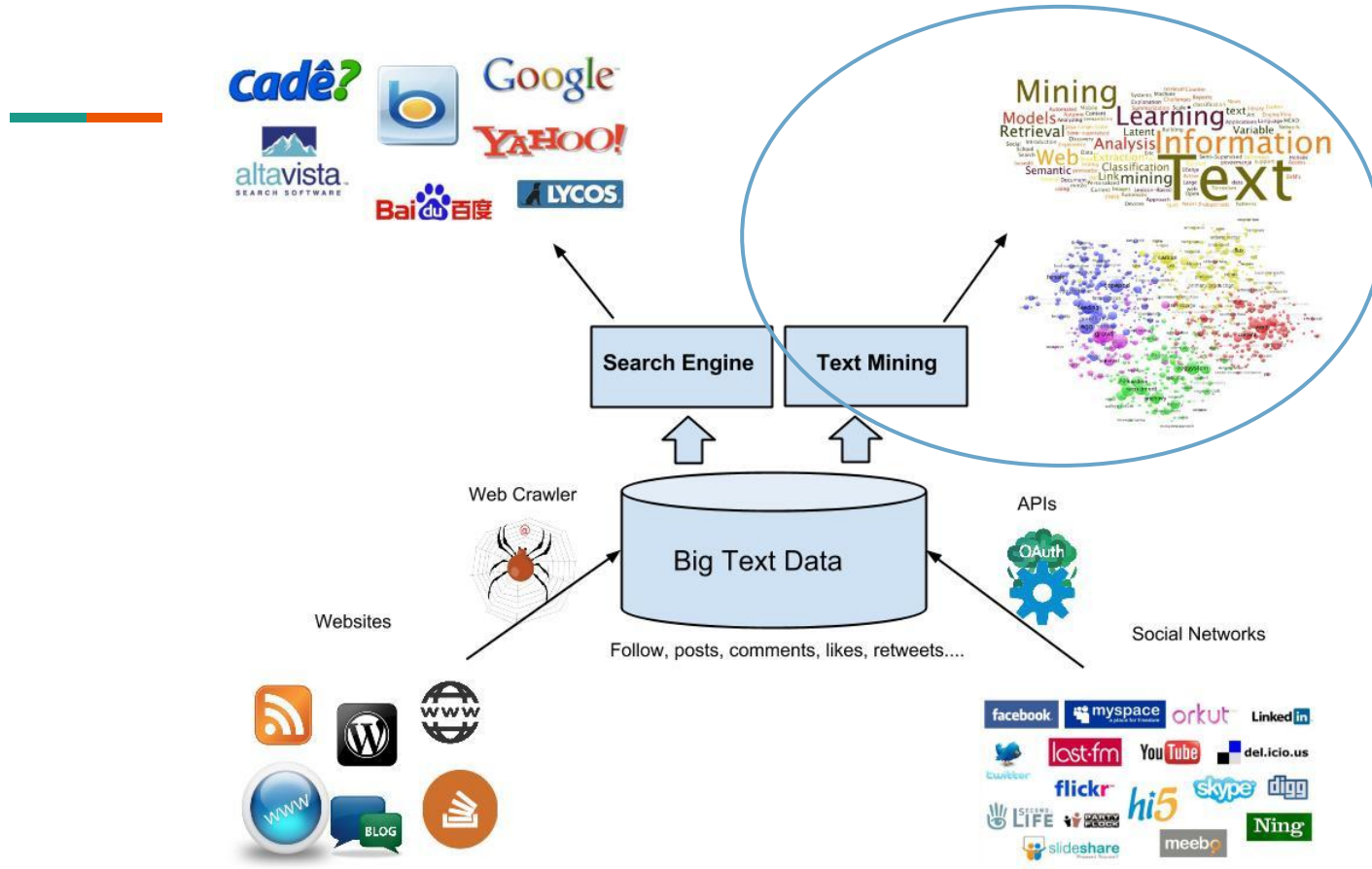
Mineração da Web e Redes Sociais



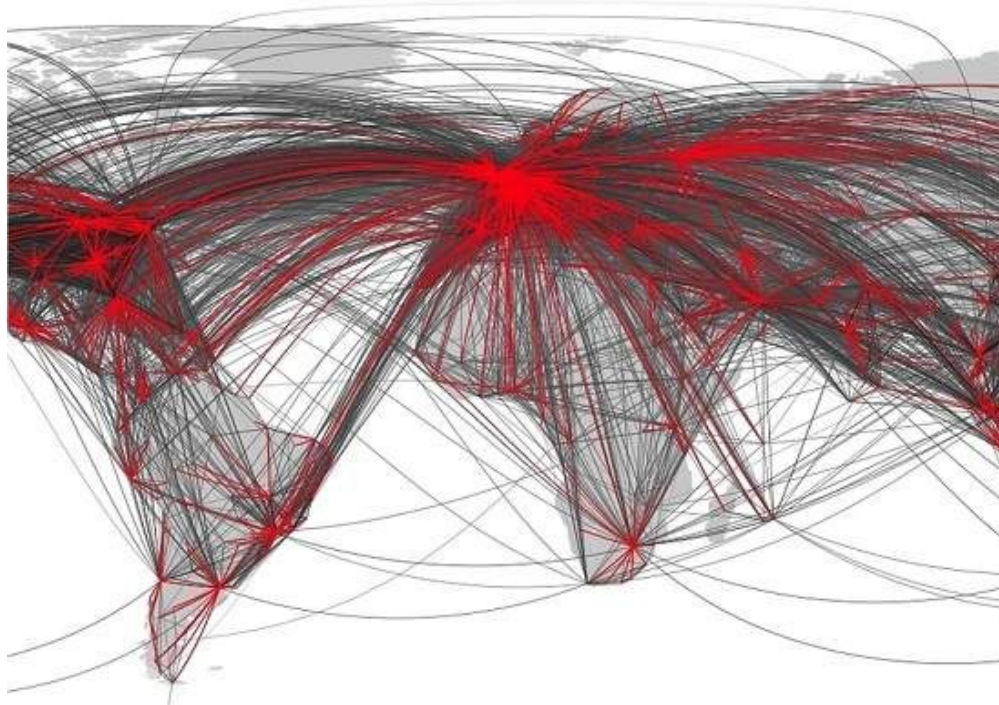
Mineração de Textos



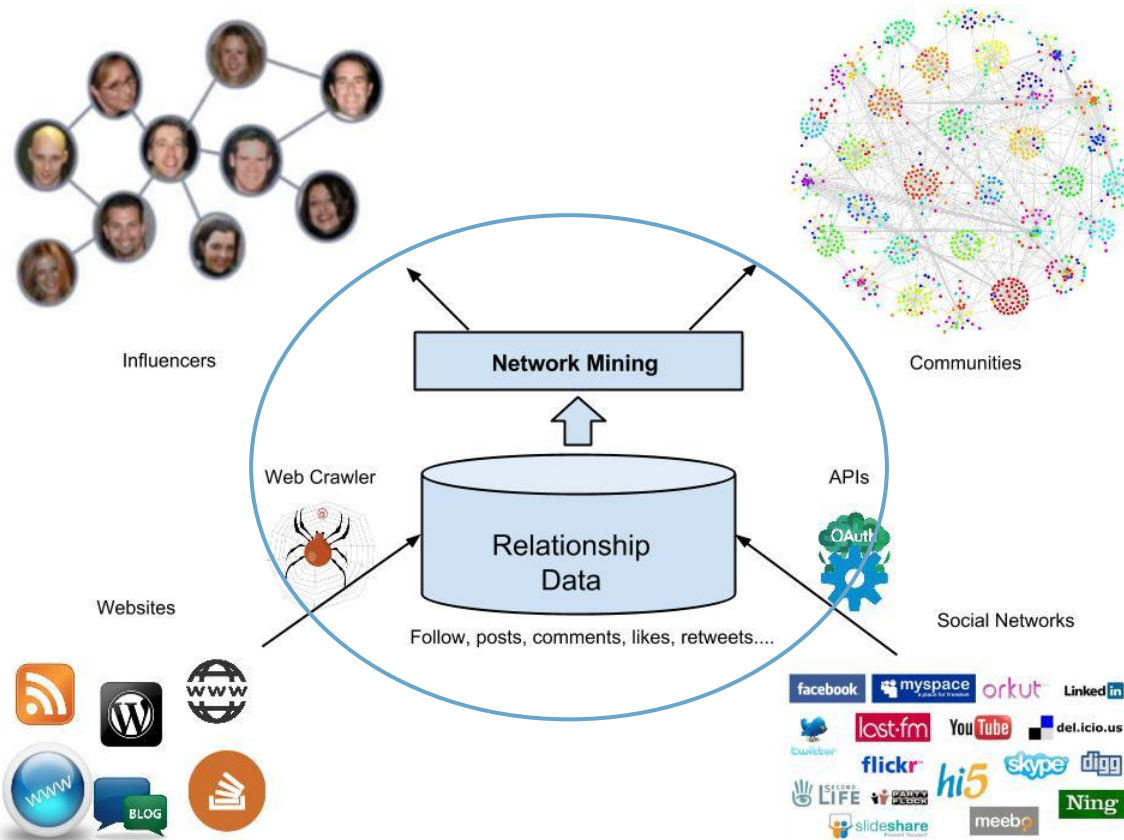
Mineração da Web e Redes Sociais



Grafos - Redes Complexas



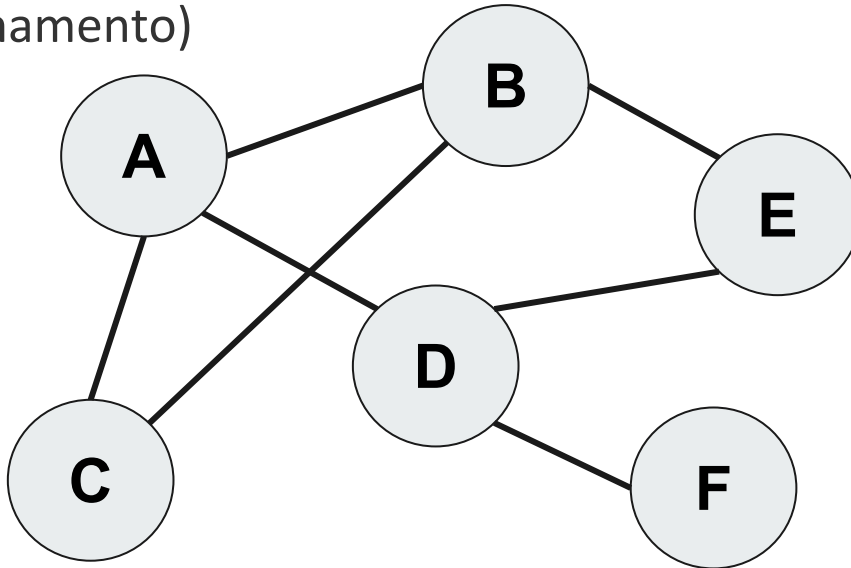
Network Mining



Grafos - Redes Complexas

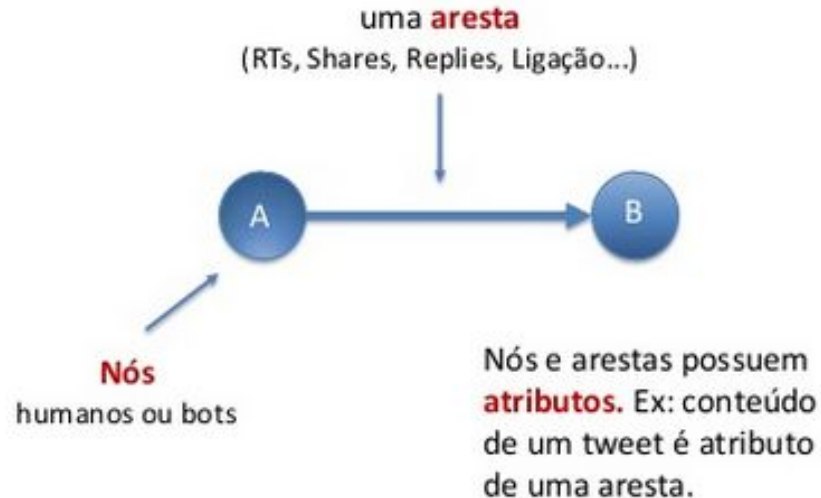
❖ O que são grafos?

- Um grafo é definido formalmente como $G = (V, E)$, onde V é o conjunto de **vértices** (entidades) conectados por E **arestas** (relacionamento)



Grafos - Redes Complexas

❖ Vértices e Arestas



Grafos - Redes Complexas

❖ Por que estudar grafos/redes?

- Importante ferramenta matemática com aplicação em diversas áreas do conhecimento
- Existem centenas de problemas computacionais que usam grafos com sucesso.
- Identificar a habilidade de comunicação entre duas entidades em uma rede
- Criar heurísticas ótimas/sub-rotinas para realizar busca de padrões em redes reais

Grafos - Redes Complexas

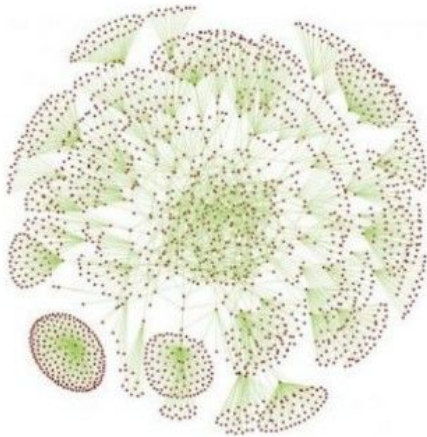
❖ O que podemos modelar por meio de grafos?

- Redes Tecnológicas
- Redes Sociais
- Redes de Informações
- Redes Biológicas

Grafos - Redes Complexas

❖ O que podemos modelar por meio de grafos?

➤ Redes Tecnológicas



Internet

Estrutura da Internet da universidade de San Diego/Califórnia



Sistema de Metrô

Sistema de metrô de Londres



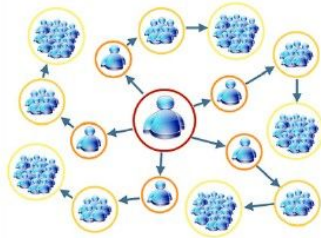
Mapa Hidrográfico

Bacia hidrográfica do rio Hérault (sul da França)

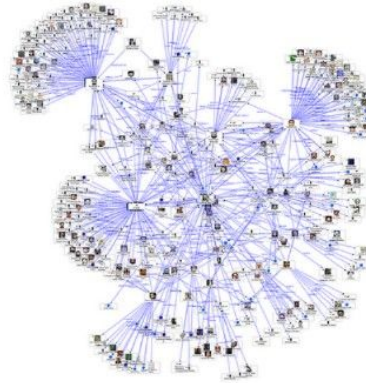
Grafos - Redes Complexas

❖ O que podemos modelar por meio de grafos?

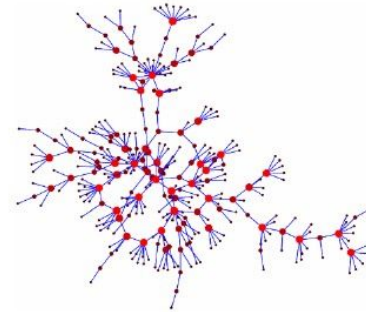
➤ Redes Sociais



Redes sociais (Facebook)



Crime Organizado

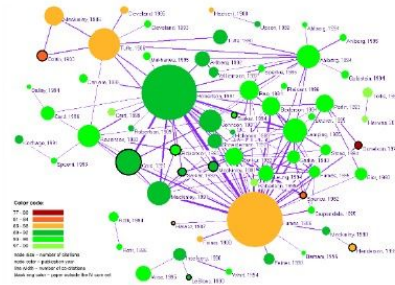


Contatos sexuais

Grafos - Redes Complexas

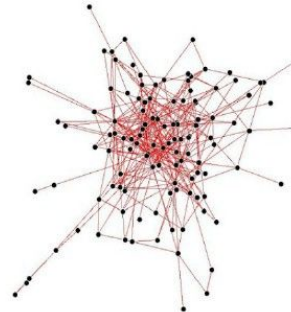
❖ O que podemos modelar por meio de grafos?

➤ Redes de Informações



Site corporativo

Citações

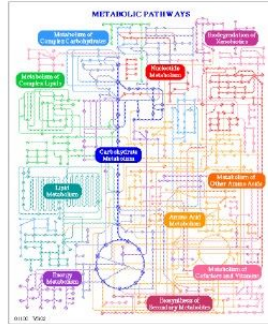


P2P - Gnutella

Grafos - Redes Complexas

◆ O que podemos modelar por meio de grafos?

➤ Redes Biológicas



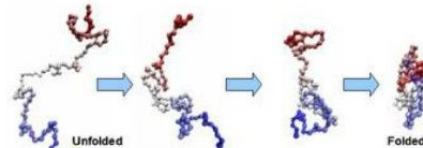
Mapa metabólico



Interações proteicas (levedura)



Etanol



Enovelamento proteico

Métricas em Redes Complexas

❖ Medidas de Centralidade

- Grau
- Closeness (Proximidade)
- Betweenness (Intermediação)

Métricas em Redes Complexas

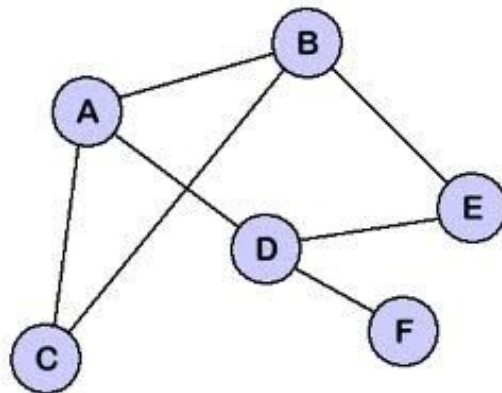
❖ Grau do Nó

- É uma medida relativa aos vértices de um grafo
- O grau de um vértice é dado pelo número de arestas que lhe são incidentes
- Exemplo:

Grau 3 = A, B, D

Grau 2 = C, E

Grau 1 = F

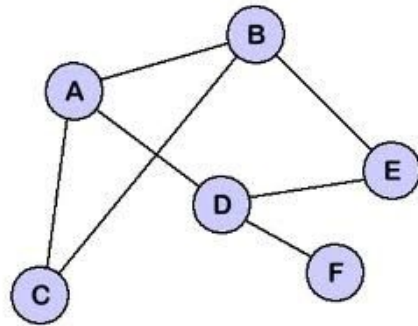


Métricas em Redes Complexas

❖ Closeness - Proximidade

- É definida pelo comprimento de **caminhos mais curtos**
- **Caminhos mais curtos** representam a menor distância entre pares de vértices
- Em um grafo **sem arestas ponderadas** o caminho é definido pelo números de arestas de um ponto a outro
- Exemplo:

A - F = (A-C-B-E-D-F) (5)
(A-B-E-D-F) (4)
(A-D-F) (2)



Métricas em Redes Complexas

❖ Closeness - Proximidade

- Define o quanto cada vértice está próximo dos demais
- Quanto mais central é o vértice menor é a distância total para todos os outros vértices
- Exemplo:

F = 11 D= 7

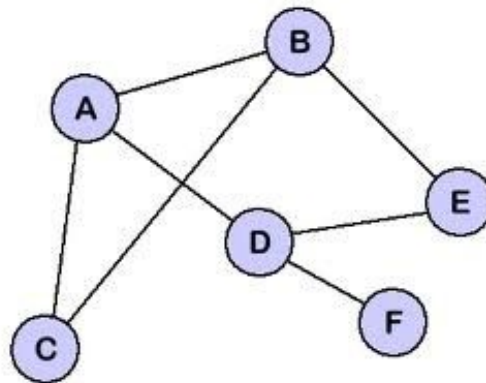
1-D 1-F

2-E 1-E

2-A 1-A

3-B 2-B

3-C 2-C



Métricas em Redes Complexas

❖ Betweenness – Intermediação (Vértice)

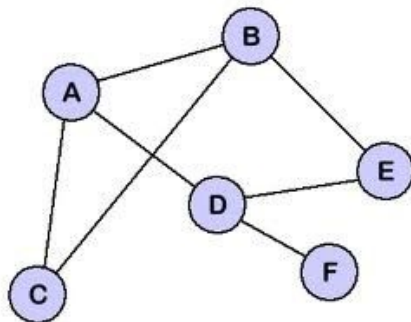
- Define o número de vezes que um **vértice** age como ponte ao longo do caminho mais curto entre dois outros vértices
 1. Para cada par de vértices calcular os caminhos mais curtos entre eles
 2. Para cada par de vértices determinar a fração de caminhos mais curtos que passam através do vértice em questão
 3. Somar esta fração de todos os pares de vértices.

➤ Exemplo:

D = A – E

A – F

C – F



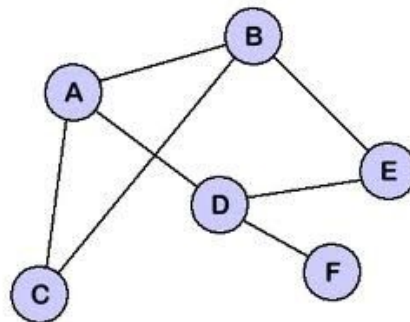
Métricas em Redes Complexas

❖ Betweenness – Intermediação (Aresta)

➤ Quantifica o número de vezes que uma **aresta** age como ponte ao longo do caminho mais curto entre dois outros vértices

➤ Exemplo:

$$\begin{aligned} A - D &= A - E \\ &A - F \end{aligned}$$



Métricas em Redes Complexas

❖ Medidas de Importância

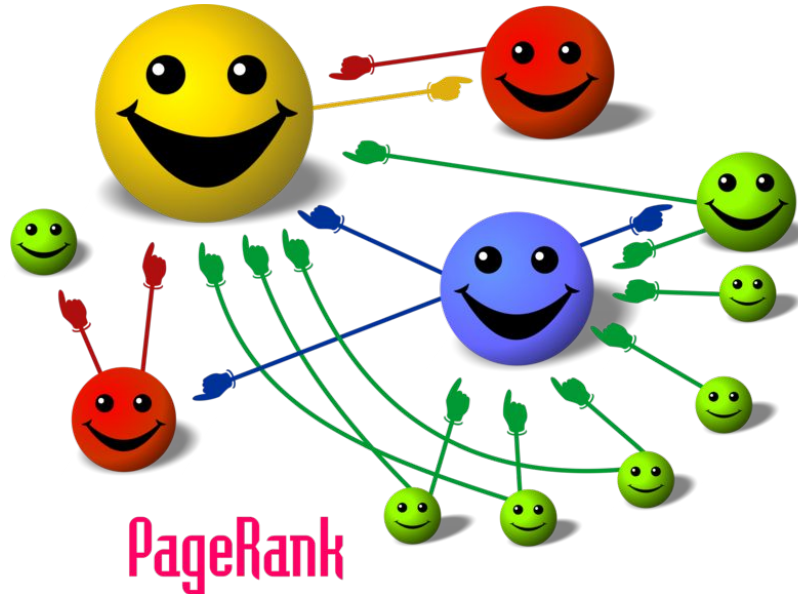
➤ PageRank

➤ HITS

Métricas em Redes Complexas

❖ PageRank

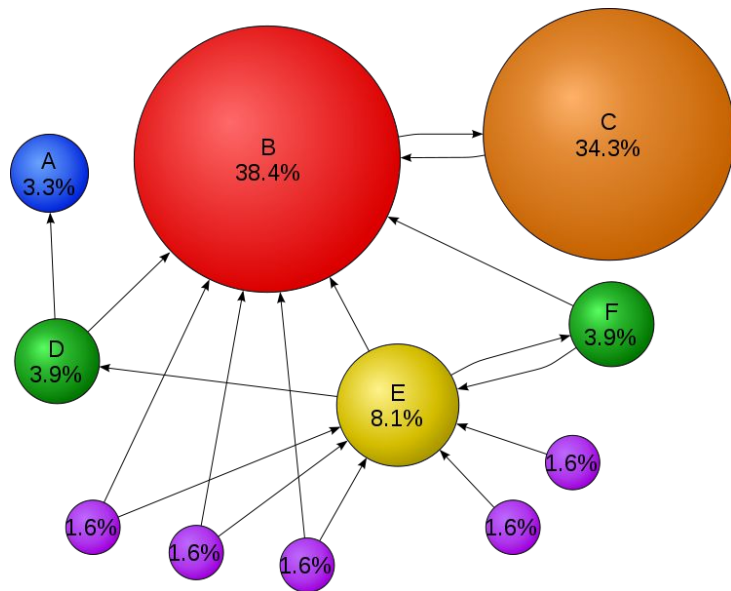
- PageRank procura expressar a probabilidade de um caminhante aleatório no grafo chegar a um vértice P



Métricas em Redes Complexas

❖ PageRank

- Considera o quanto um vértice é referenciado (Observe B)
- Se quem aponta para o vértice também é importante (Observe C)



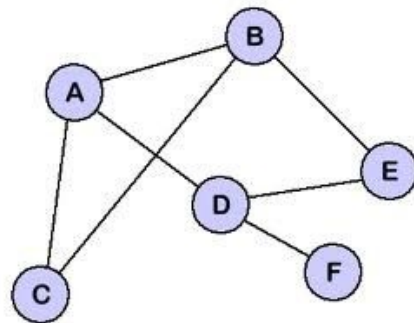
Métricas em Redes Complexas

❖ PageRank

- A importância de um vértice P é dada pela seguinte equação:

$$PR(A) = \frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)}.$$

- $PR(i)$ - PR do vértice i que aponta para A
Probabilidade inicial de todas $1/N$
- $L(i)$ - quantidade de links de saída em i

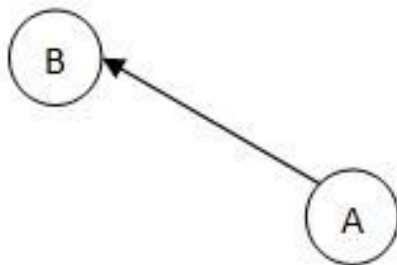


Métricas em Redes Complexas

❖ PageRank - Desafios

➤ Vértices sem ligações

- A não recebe links de ninguém e passa a ter $PR=0$
- B recebe 0 de A

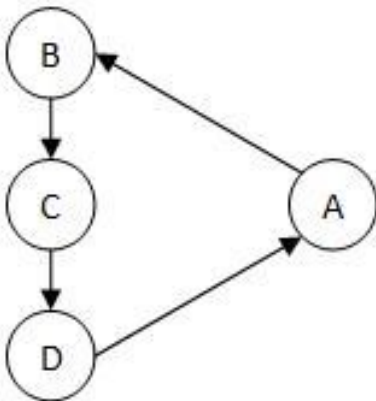


Métricas em Redes Complexas

❖ PageRank - Desafios

➤ Ciclos

- Cálculo do PageRank fica "preso" no ciclo infinito
- Em cada iteração o valor de PageRank é transmitido de um vértice para outro do ciclo e não há convergência



Métricas em Redes Complexas

❖ PageRank - Dump Factor

$$PR(A) = \frac{1-d}{N} + d \left(\frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)} + \dots \right).$$

- d - dump factor (geralmente entre 0.1 e 0.9)
Probabilidade de um caminhante aleatório continuar seguindo os links.
Do contrário acontecerá um “teletransporte” para qualquer outro vértice
- N - Total de páginas

Métricas em Redes Complexas

❖ HITS

- Utiliza valores de **hub** e **autoridade** para definir a reputação de uma página P.

Hub de uma página “P” – é dado em função dos valores de autoridade das páginas **para onde ela aponta**.

Autoridade de uma página “P” – é dada em função dos valores de hub das páginas **que apontam para P**.

- Um bom hub é uma página que aponta para boas autoridades e uma boa autoridade é uma página apontada por bons hubs.

Métricas em Redes Complexas

❖ Medidas em Redes Sociais

Você saberia dizer uma aplicação para essas métricas em sistemas de RI, como as máquinas de busca?

E nas mídias sociais?



KNIME - Análise de Redes

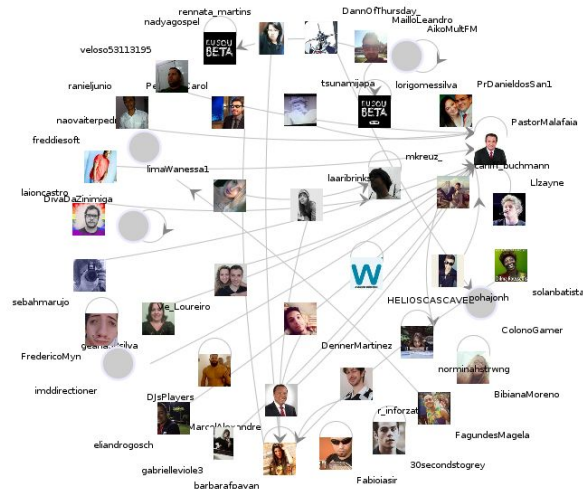
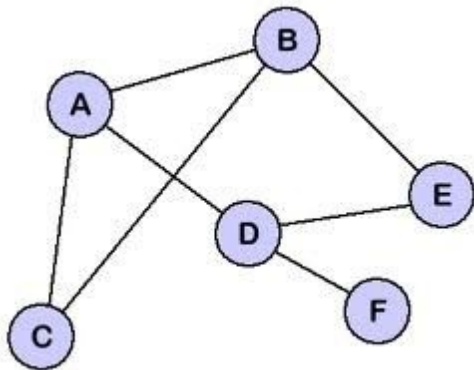


- Fluxo: **Network Analysis - Boticario - pratica**
- Análise de influenciadores no twitter
- Encontrar o componentes fortemente conectados

Coleta em Redes Sociais

❖ Selecionando os principais tweets

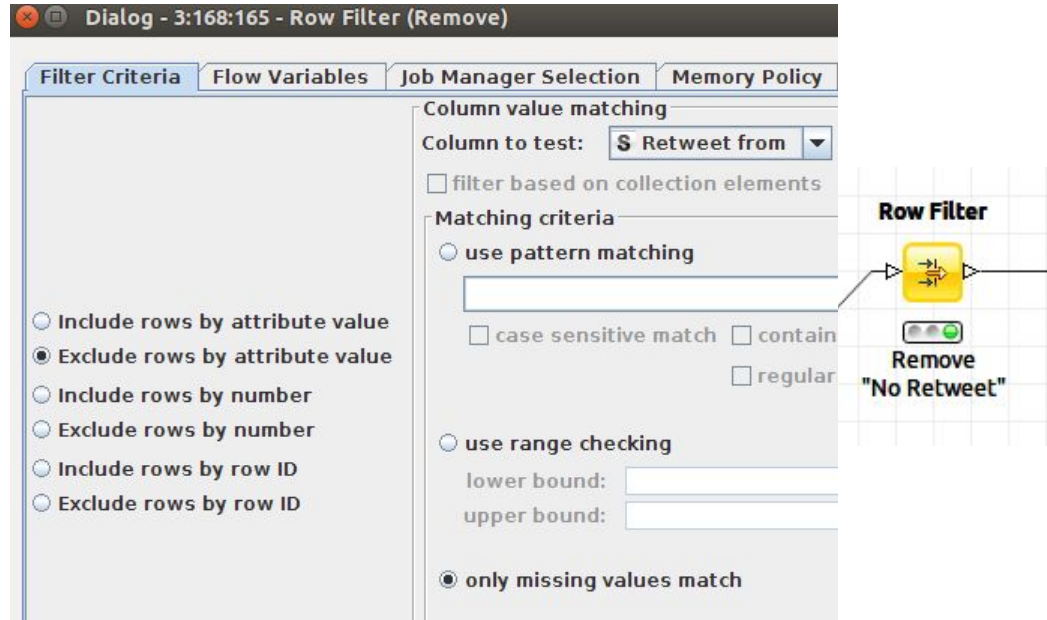
Uma vez que você tenha os **tweets coletados** o primeiro passo é extrair os **principais** tweets da base, ou seja aqueles que **foram compartilhados** (retweets).



Coleta em Redes Sociais

❖ Limpar tweets sem retweets

Com isso é possível diminuir o número de conexões a serem analisadas, focando apenas nos posts que existem conexões na rede



Coleta em Redes Sociais

❖ Contar interações conexões entre usuários

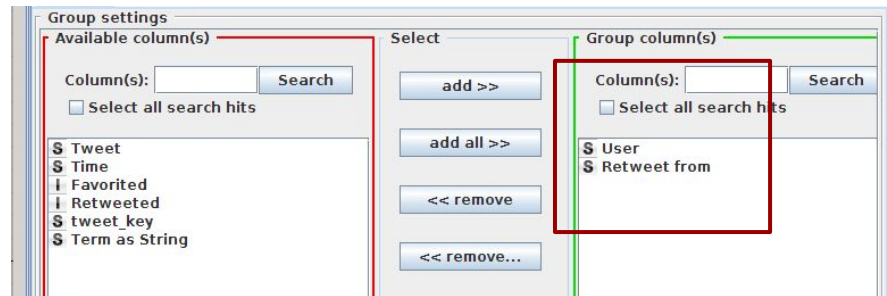
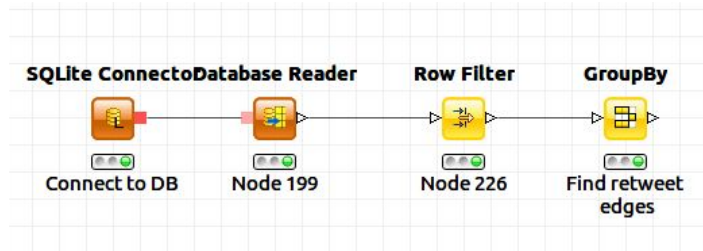
Em seguida, vamos encontrar as conexões entre os usuários através dos seus compartilhamentos. Assim você tem dados para formar um grafo conectado de usuários e retweets.

Table "database" - Rows: 29						Spec - Columns: 6	Properties	Flow Variables
Row ID	S User	S Tweet	S Time	I Fav...	I R...	S Retweet f...		
Row2	rennata_marti...	RT @barbarafpav...	2015-06-07 ...	0	6	barbarafpav...		
Row3	DannOfThursd...	RT @ColonoGam...	2015-06-06 ...	0	9	ColonoGamer		
Row4	r_inforzato	RT @barbarafpav					Table "default" - Rows: 27	
Row5	solanbatista	RT @barbarafpav						
Row6	DennerMartinez	RT @barbarafpav					Spec - Columns: 3	
Row7	PereiraBCarol	RT @barbarafpav					Properties	Flow
Row8	laaribrinks	RT @barbarafpav					Row ID	S User
							S Retweet from	I Count(Time)
							Row0	DannOfThursday_
							Row1	DennerMartinez
							Row2	FagundesMagela
							Row3	Llwayne
							Row4	MailloLeandro
							Row5	MailloLeandro
								nadyagospel

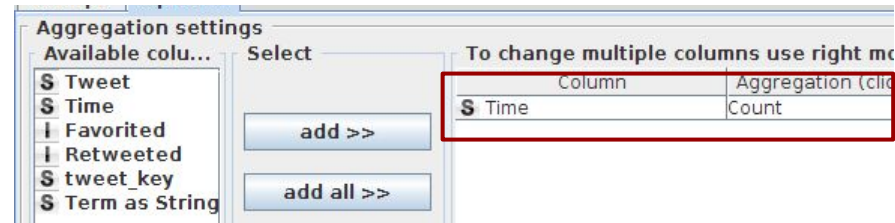
Coleta em Redes Sociais

❖ Agrupar usuários por retweets

(Agrupe Users e Retweet from)



Através da aba **Manual Aggreagation**, conte (Count) quando (Time) um usuário *retuitou* outro usuário



Coleta em Redes Sociais

- ❖ Resultado da contagem de conexões entre usuários

Table "default" - Rows: 27		Spec - Columns: 3	Properties	Flow
Row ID	S User	S Retweet from	I Count(Time)	
Row0	DannOfThursday_	ColonoGamer	1	
Row1	DennerMartinez	barbarafpavan	1	
Row2	FagundesMagela	naovaiterpedro	1	
Row3	Lzayne	PastorMalafaia	1	
Row4	MailloLeandro	lorigomessilva	1	
Row5	MailloLeandro	nadyagospel	2	

Observe que:

DannOfThursday_ retuitou ColonoGamer 1 vez

MailloLeandro retuitou nadyagospel 2 vezes

Coleta em Redes Sociais

❖ Criando Grafo ou Rede de Conexões

Options Advanced Options Flow Variables Job Manager Selection Memory Policy

Node settings

Node id column: (opt.) \$ User

Node label column: (opt.) \$ User

Second node id column: (opt.) \$ Retweet from

Second node label column: (opt.) \$ Retweet from

Edge settings

Edge id column: (opt.) ? <RowID>

Edge label column: (opt.) \$ User

☒ Create directed edges

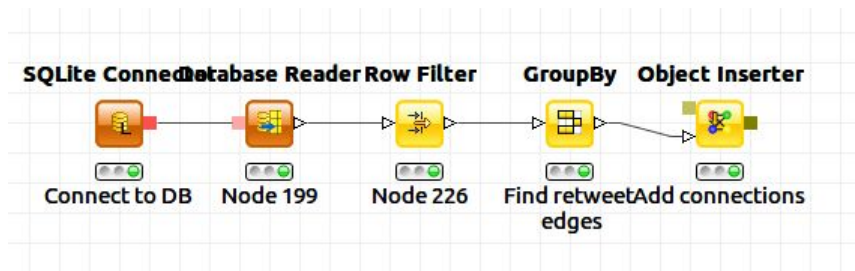
Weight settings

☐ None ☐ Default ☒ Column

Default weight: 1,0

Weight column: Count(Time)

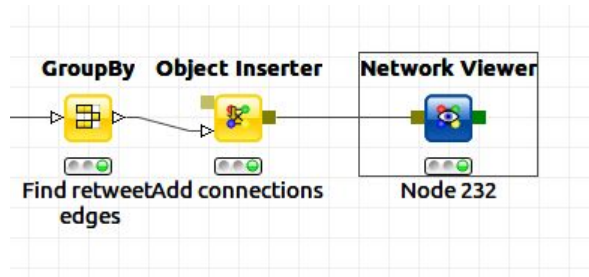
☐ All nodes have same weight



Com o **Object Inserter** definimos quem serão os vértices (Nodes) a serem conectados, o que identifica as arestas (Edge) se haverá um peso para cada conexão (Weight Column)

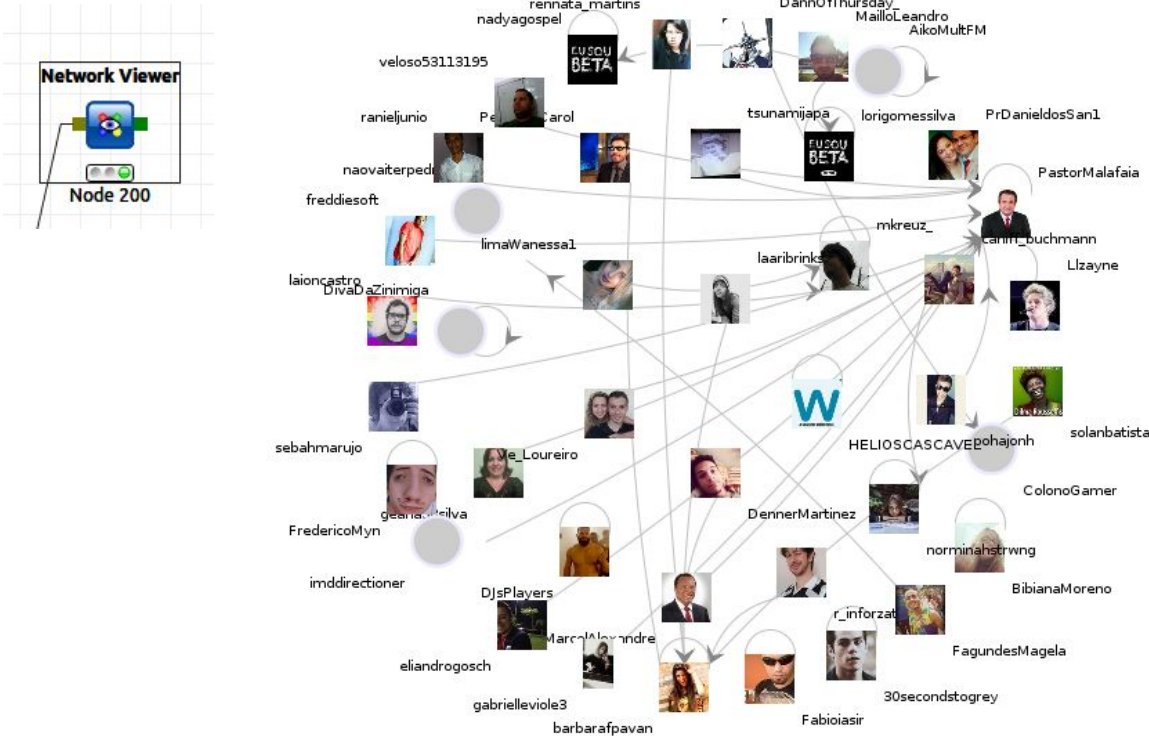
Coleta em Redes Sociais

❖ Rede de retweets (influenciadores)



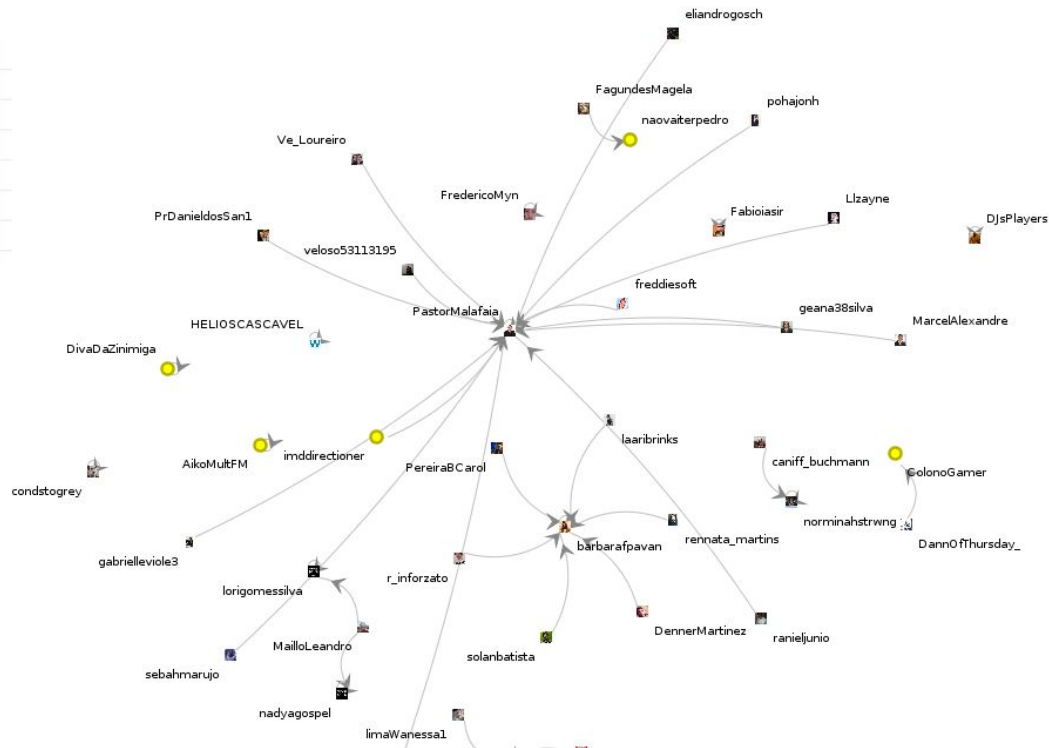
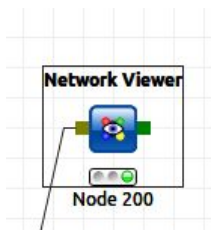
Coleta em Redes Sociais

❖ Vamos enriquecer a rede com outros atributos



Coleta em Redes Sociais

❖ E até outros formatos de visualização

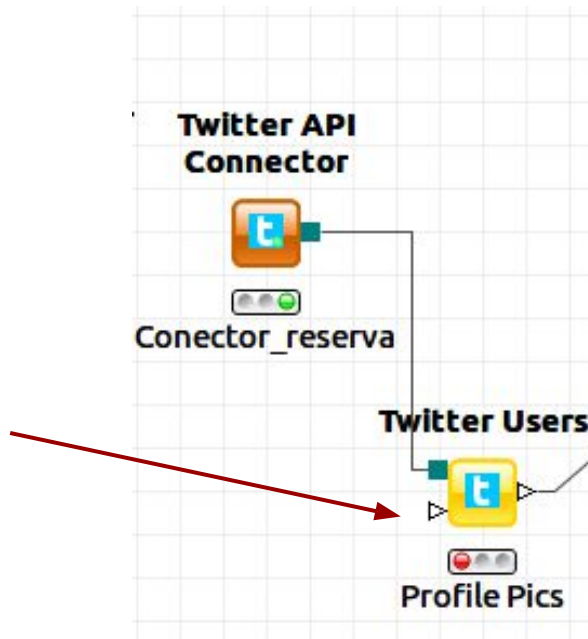


Coleta em Redes Sociais

❖ Buscando informações de usuários no Twitter

Com o node **Twitter Users** buscamos informações de cada usuário.

Precisamos da lista de usuários a serem buscados.



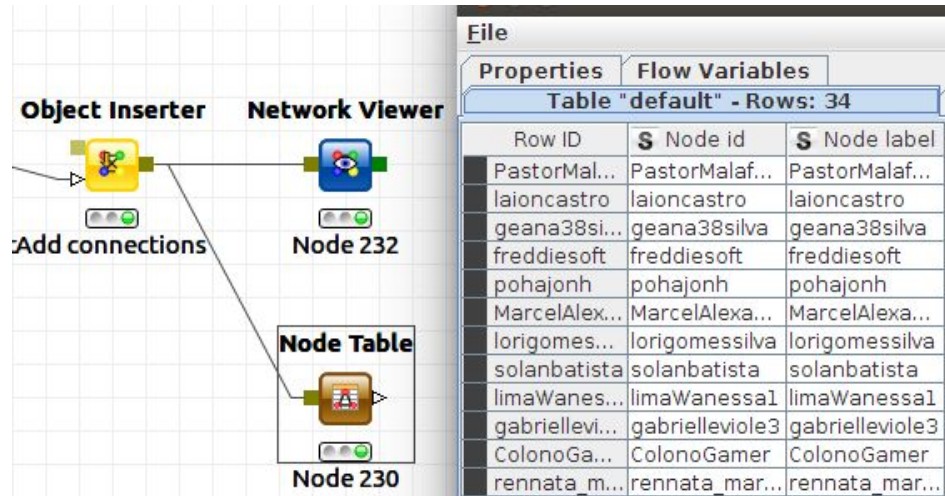
Análise de Redes

❖ Informações de Vértices

O **Node** (no sentido de Vértice) **Table** gera uma tabela com os vértices da rede

Podemos buscar informações já filtradas apenas para essa rede específica.

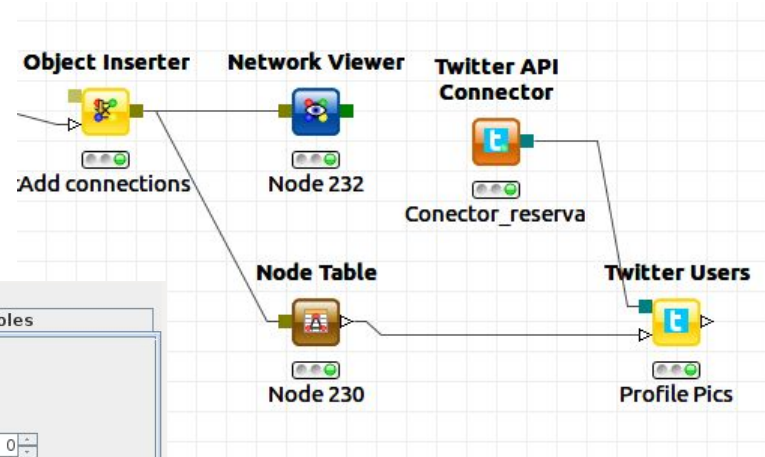
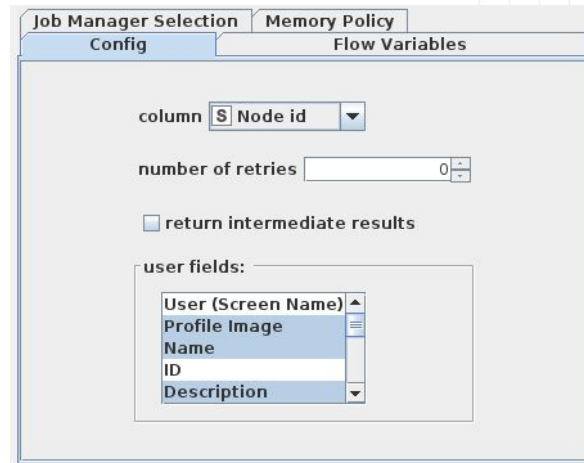
Com o Node Table extraímos os ids dos vértices da rede, ou seja nomes de usuários.



Análise de Redes










❖ Busca de dados de usuários

Agora podemos fornecer o ID dos usuários no node **Twitter Users** e buscar informações no Twitter.



Análise de Redes

❖ Agora temos dados para enriquecer a rede e suas respectivas conexões

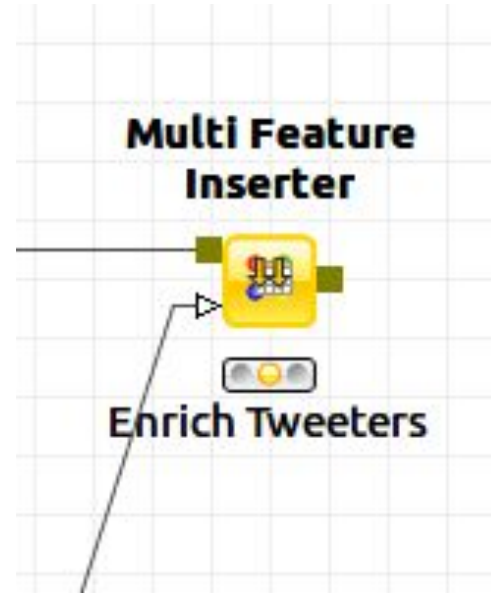
Table "default" - Rows: 22 Spec - Columns: 14 Properties Flow Variables														
Row ID	S Node id	S Node label	PRG	User - Profile image	S User - Name	S User - ...	S User - ...	S User - Creation t...	S User - ...	S User - ...	S User - ...	I User - ...	I User - ...	I Us
MaillioLean...	MaillioLeandro	MaillioLeandro			leandro guimaraes		?	2015-05-27 20:57:42	pt		?	1129	282	1
Lizayne	Lizayne	Lizayne			Ooopssss Yne	DIRECTIO...	?	2012-11-03 22:58:21	pt	Brazil	?	72396	2847	1820
r_inforzato	r_inforzato	r_inforzato			Ricardo Inforzato	Empreen...	https://t....	2008-09-08 02:36:02	pt	São Paul...	?	127695	3792	1206
MarcelAlex...	MarcelAlexa...	MarcelAlexa...			Marcel Alexandre	Minha po...	https://t....	2009-09-25 14:38:31	pt	Manaus -...	?	37190	25008	4953
eliandrogo...	eliandrogosch	eliandrogosch			Eliandro Gosch		http://t.c...	2010-06-03 20:10:28	en		?	685	46	223
solanbatista	solanbatista	solanbatista			solange			2009-07-25 10:57:47	pt		?	66283	1465	3321
PrDaniel...	PrDanieldos...	PrDanieldos...			Daniel S. Lobo	Pasteur ... Fondateu...	?	2012-09-14 19:21:52	pt	Genebra,...	?	5632	136	5626
veloso531...	veloso5311...	veloso5311...			BrasilTerradoNun...		?	2014-06-14 23:05:55	pt		?	18765	698	1355
sebahmar...	sebahmarujo	sebahmarujo			שבחמתי	O cara q...	?	2010-09-09 18:58:57	pt	Santo An...	?	4611	129	2727

Coleta em Redes Sociais

❖ Inserindo atributos na rede

Com o node **Muiti Feature Inserter** inserimos os dados na rede.

Quais entradas do node?

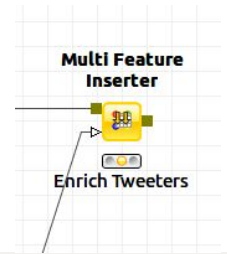


Coleta em Redes Sociais

❖ Inserindo atributos na rede

Nesse exemplo estamos inserindo a foto do usuário e o número de followers.

Porém é possível selecionar diversos atributos da lista



The screenshot shows the 'Options' tab of a software interface. It has four sub-tabs: 'Options', 'Flow Variables', 'Job Manager Selection', and 'Memory Policy'. The 'Options' tab is active.

ID settings

Edge id column: [?] <none> Node id column: [S] Node id

Feature settings

☒ Manual Selection ☐ Wildcard/Regex Selection ☐ Type Selection

Exclude (highlighted with a red box)

Filter

- [S] Node id
- [S] Node label
- [PRG] User - Profile image
- [S] User - Name
- [S] User - Description
- [S] User - URL
- [S] User - Creation time
- [S] User - Language
- [S] User - Location
- [S] User - Time Zone

☐ Enforce exclusion

Include (highlighted with a green box)

Filter

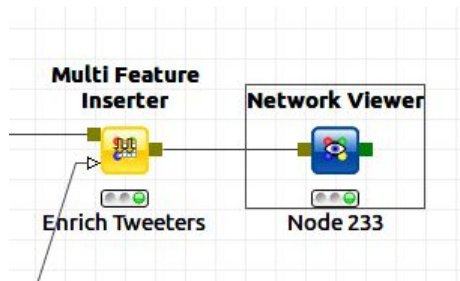
- [?] Node id - Profile image
- [?] Node id - Followers

☒ Enforce inclusion

Coleta em Redes Sociais

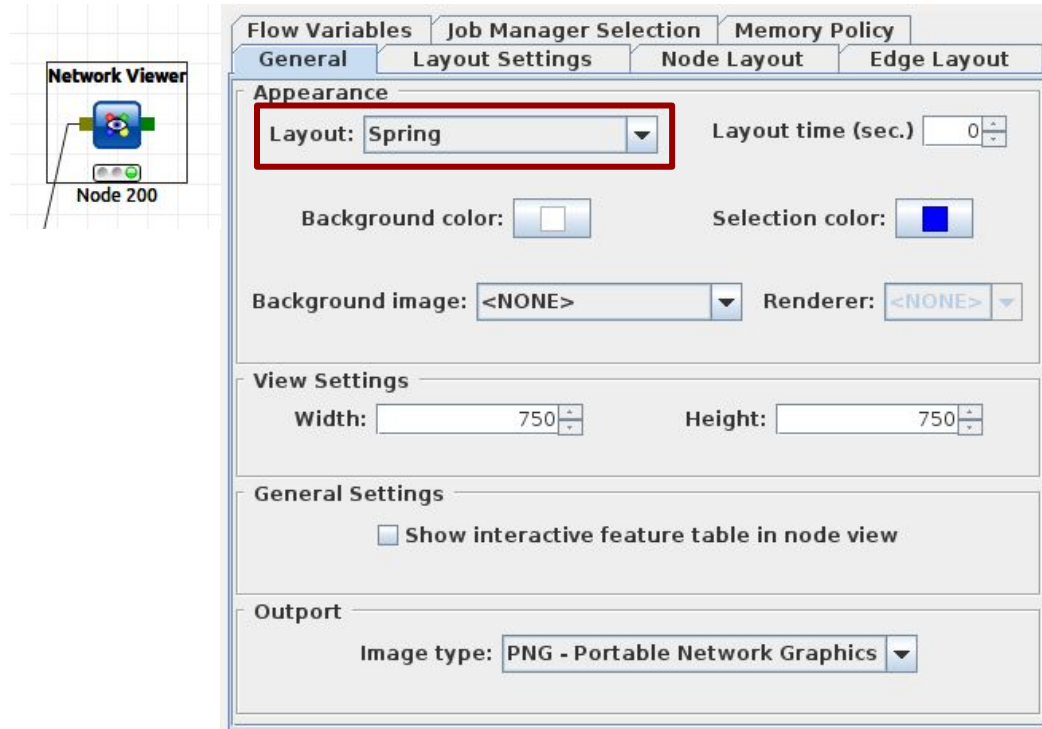
- ❖ Inserindo a foto do perfil do usuário na rede

Em **Node Layout** configuramos o campo **Icon** para utilizar a coluna “User - Profile image” com fotos que coletamos de cada usuário.



Coleta em Redes Sociais

- ❖ Customize o layout do grafo através das abas General



❖ Positivos/Negativos



- Grafo dos usuários que postaram tweets **Positivos**
- Grafo dos usuários que postaram tweets **Negativos**



Calculando outras Métricas

Network Analysis - SchoolsWiki

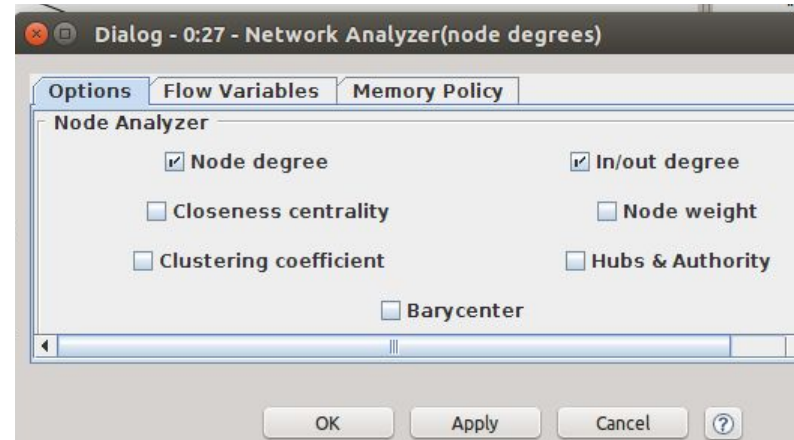
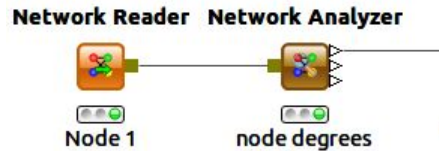


- Fluxo: **NetworkAnalysisSchoolsWiki-pratica**
- Análise básica em redes
- Encontrar o componente mais conectado
- Top vértices com o maior grau total/de entrada/de saída

Análise de Redes

❖ Métricas

Leitura habitual
da rede com o
Network Reader



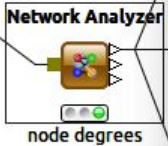
O **Network Analyzer** permite calcular várias métricas.

- Vamos calcular o grau do vértice, ou seja, o número total de ligações de cada vértice possui
- Além disso o grau de entrada e saída separadamente

Análise de Redes

❖ Métricas

Resultado do
Network Analyzer



Network Analyzer
node degrees

Node statistics table - 0:27 - Network Analyzer(node degrees)

File

Properties Flow Variables

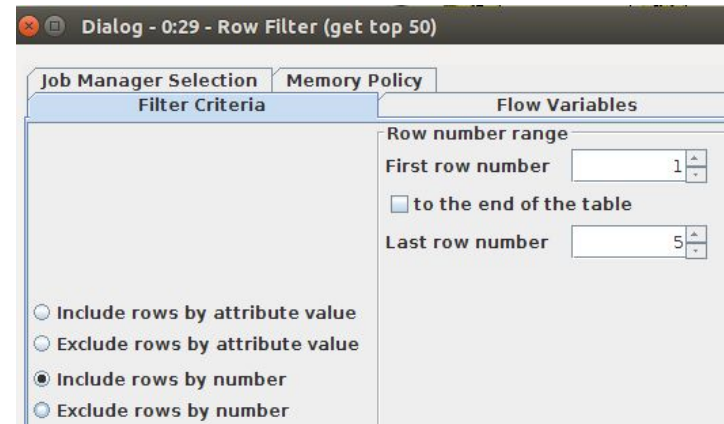
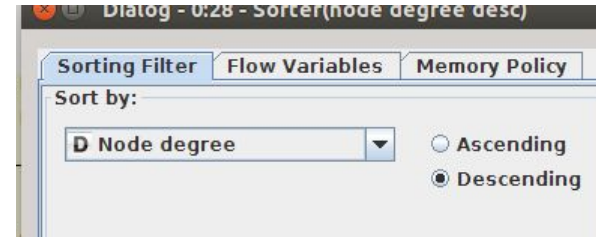
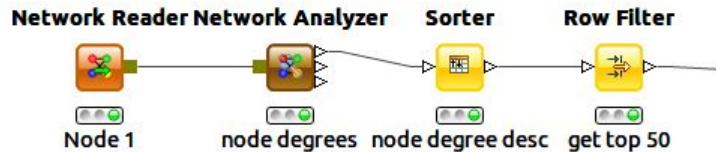
Table "spec_name" - Rows: 5538 Spec - Columns: 7

Row ID	S Object id	D Node...	D No...	D In deg...	D In d...	D
Eilmer of ...	Eilmer of Malm...	9	0.004	3	0.001	6
List of Pri...	List of Prime M...	150	0.069	2	0.001	14
Battle of M...	Battle of Mosc...	46	0.021	13	0.006	33
Byzantine ...	Byzantine Emp...	269	0.124	180	0.083	89
1937	1937	277	0.127	137	0.063	14
Abbadid	Abbadid	12	0.006	1	0	11
1913	1913	236	0.108	114	0.052	12
Weimar Re...	Weimar Republic	77	0.035	32	0.015	45
Edward VI ...	Edward VI of E...	70	0.032	31	0.014	39
1988	1988	291	0.134	146	0.067	14
Medieval c...	Medieval com...	18	0.008	6	0.003	12
Bodyline	Bodyline	24	0.011	6	0.003	18
Political in...	Political integr...	59	0.027	3	0.001	56
1827	1827	117	0.054	61	0.028	56
Battle of t...	Battle of the Li...	21	0.01	9	0.004	12

Análise de Redes

❖ Filtrando os top 5 vértices de maior grau total

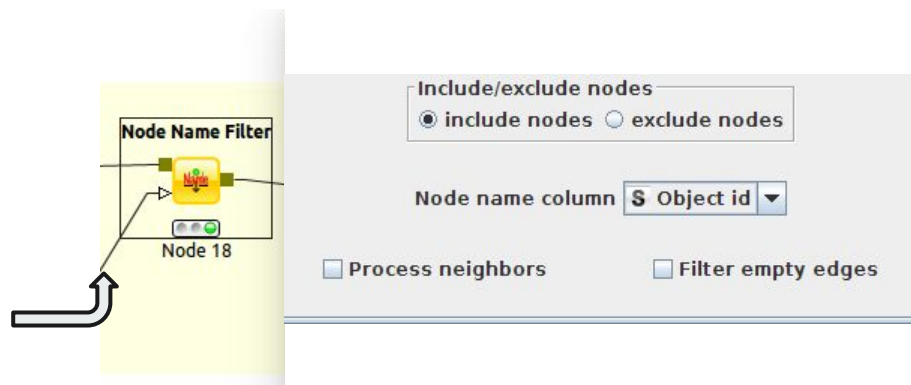
Ordene os vértices por grau decrescente e **filtre** os 5 maiores



Análise de Redes

❖ Retirando vértices

Com o **Node Name Filter** podemos excluir da rede os vértices por nome. Nesse caso vamos incluir na rede apenas os vértices que filtramos anteriormente.



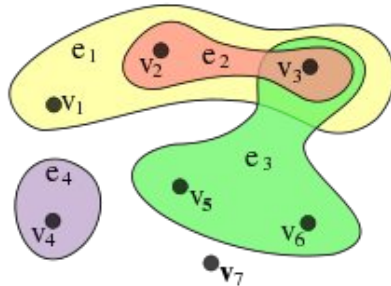
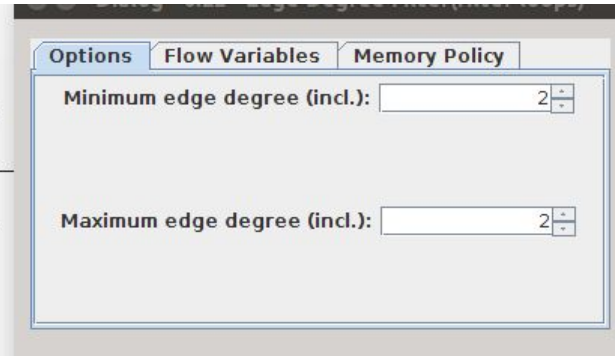
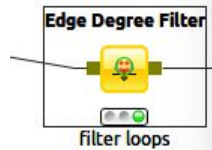
Obs: é preciso conectar também a rede original

Nota:

Não esqueça que “Node” em grafos significa vértice ou nó da rede e não um node do knime

❖ Hiperarestas

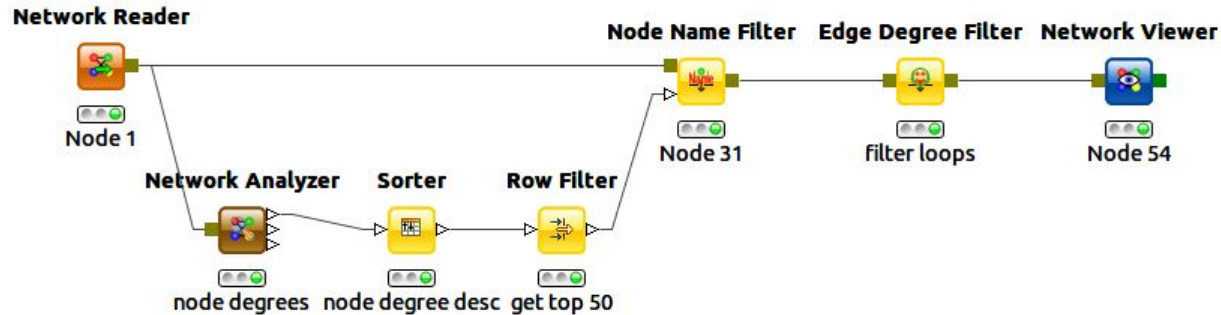
Com o **Edge Degree Filter** eliminamos hiperarestas, ou seja, selecionamos arestas que conectam no máximo dois vértices



Isso é necessário pois o visualizador do knime não aceita hipergrafos

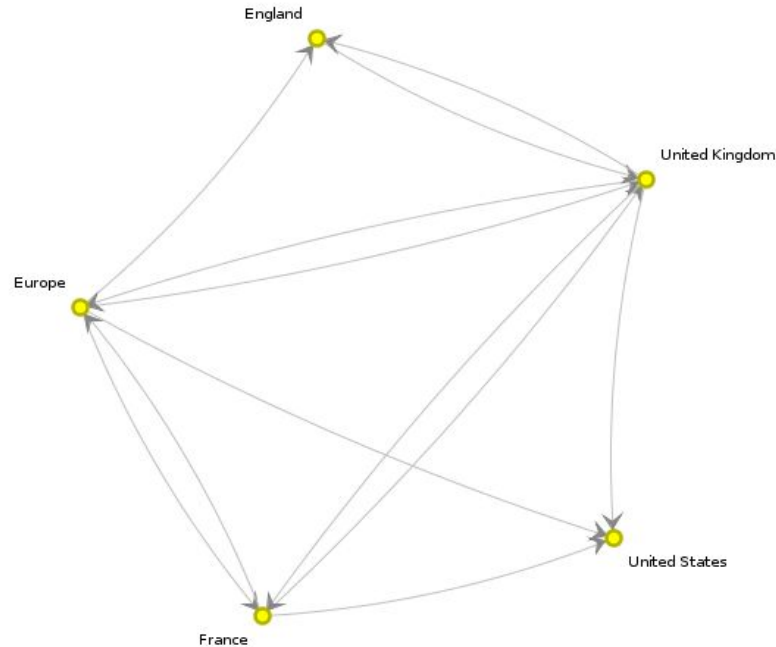
Análise de Redes

- ❖ Grafo formado pelos 5 vértices de maior grau de entrada



Análise de Redes

- ❖ Grafo formado pelos 5 vértices de maior grau de entrada (**Resultado**)



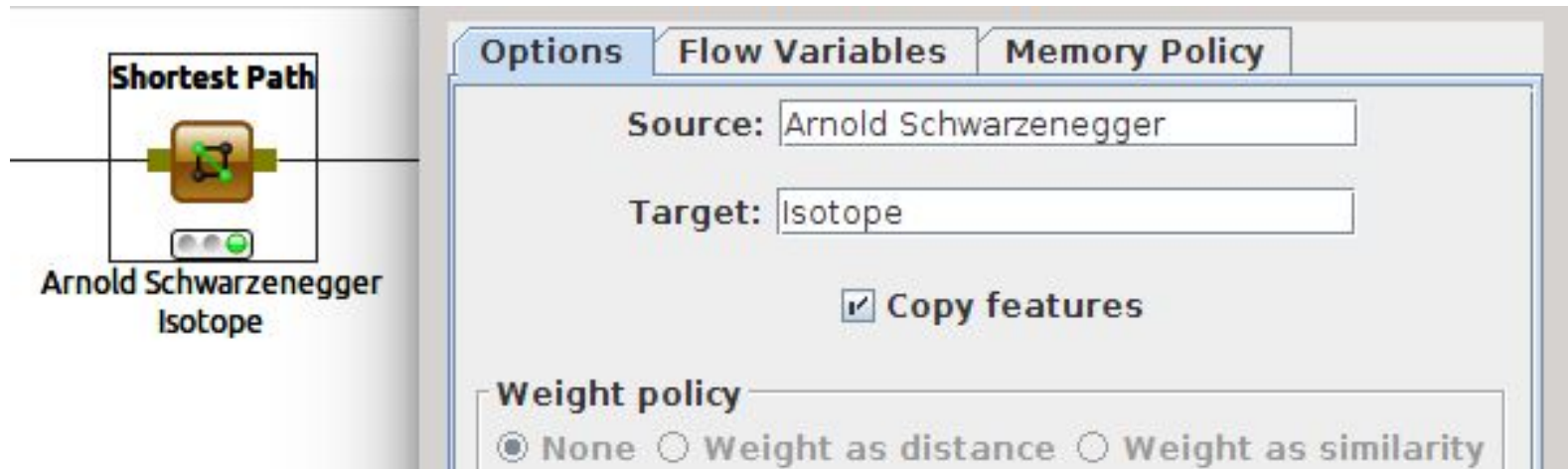
Network Analysis - SchoolsWiki



- Fluxo: **NetworkAnalysisSchoolsWiki-pratica-menor-caminho**
- Menor caminho entre vértices

Análise de Redes

- ❖ Qual o menor caminho entre **Schwarzenegger** e **Isótopo**?

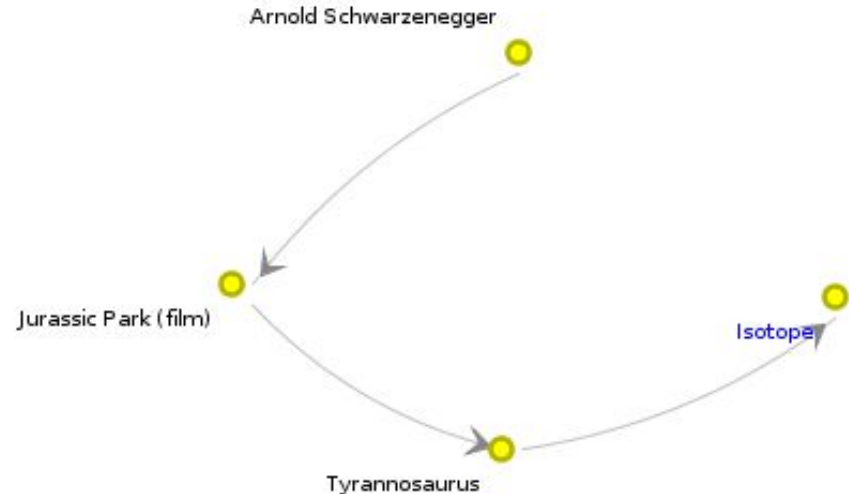
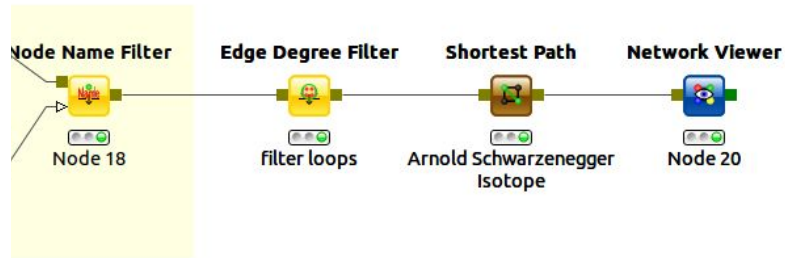


A medida do menor caminho é implementada pelo node **Shortest Path**

Análise de Redes

❖ Qual o **menor caminho** entre **Schwarzenegger** e **Isótopo**?

Para visualizar uma rede utilizamos o **Network Viewer**



Análise de Redes

- ❖ Como assim? Jurassic Park easter egg





A disciplina - RI

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. 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.
- **Unidade 03:** Tipos de coleta, arquitetura e componentes de coletores Web. Ferramenta e prática de coleta de dados na Web.

❖ Plano de Ensino



- **Unidade 04:** Aprofundando na mineração de texto e linguagem natural. Algoritmos e soluções para a análise da informação presente nas redes sociais online e em sites de conteúdo. Ferramenta e prática de mineração de texto.
- **Unidade 05:** Caracterização de redes sociais: Tipologia, características e representações gráficas. Algoritmos estocásticos, análise de redes complexas. Ferramenta e prática de mineração de redes complexas.

A disciplina - RI

❖ Teórico e Prático



O conteúdo estudado será exercitado em práticas utilizando ferramentas de mineração de texto e busca.

As aulas práticas serão avaliadas e em cada prática uma tarefa deverá ser realizada de maneira autônoma. **40 pontos.**

O Projeto Final será formado por conceitos discutidos e aplicados nas aulas, com adaptações individuais para um caso de uso real. O resultado das tarefas práticas poderá ser reaproveitado. **60 Pontos**

Condições para duplas**

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**
 - 15° dia após a última aula às 23:59hrs

Tópicos

◆ Mineração de Tópicos em Rede (Desafio)

Embora seja possível criar tagclouds que mostrem os tópicos por cores, o que você acha de criar um grafo de tópicos? Ficaria muito legal no projeto final :)

Tópicos em grafos
seria algo assim:

