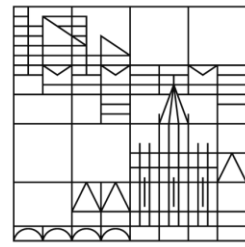




**Nycomed Chair for  
Bioinformatics and Information Mining**

Universität  
Konstanz



# **KNIME Textprocessing Feature**

Kilian Thiel

05.10.2011

Overview



Where ?

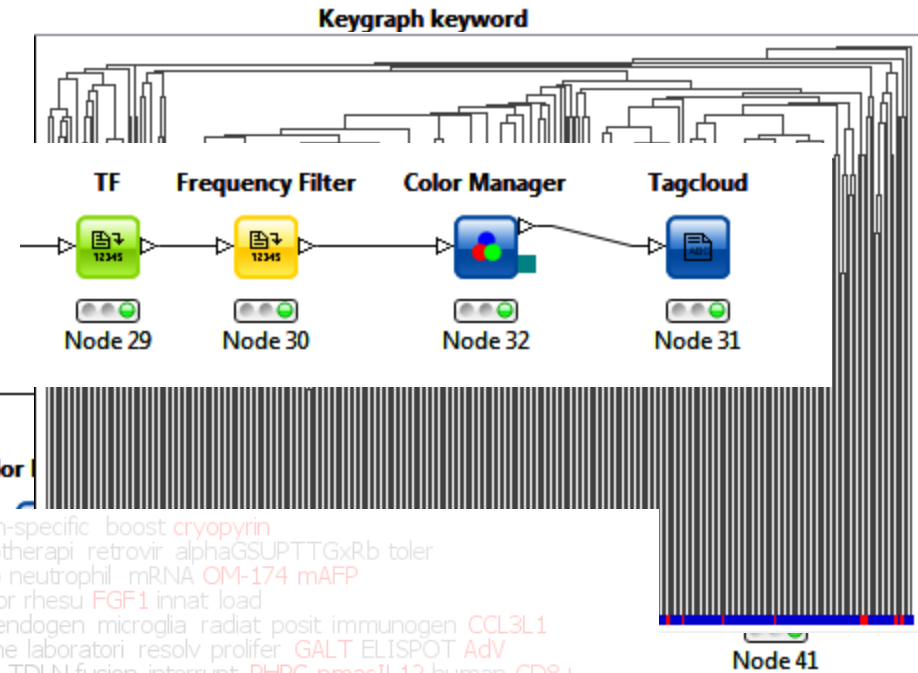
# Textprocessing

<http://labs.knime.org>



# What ?

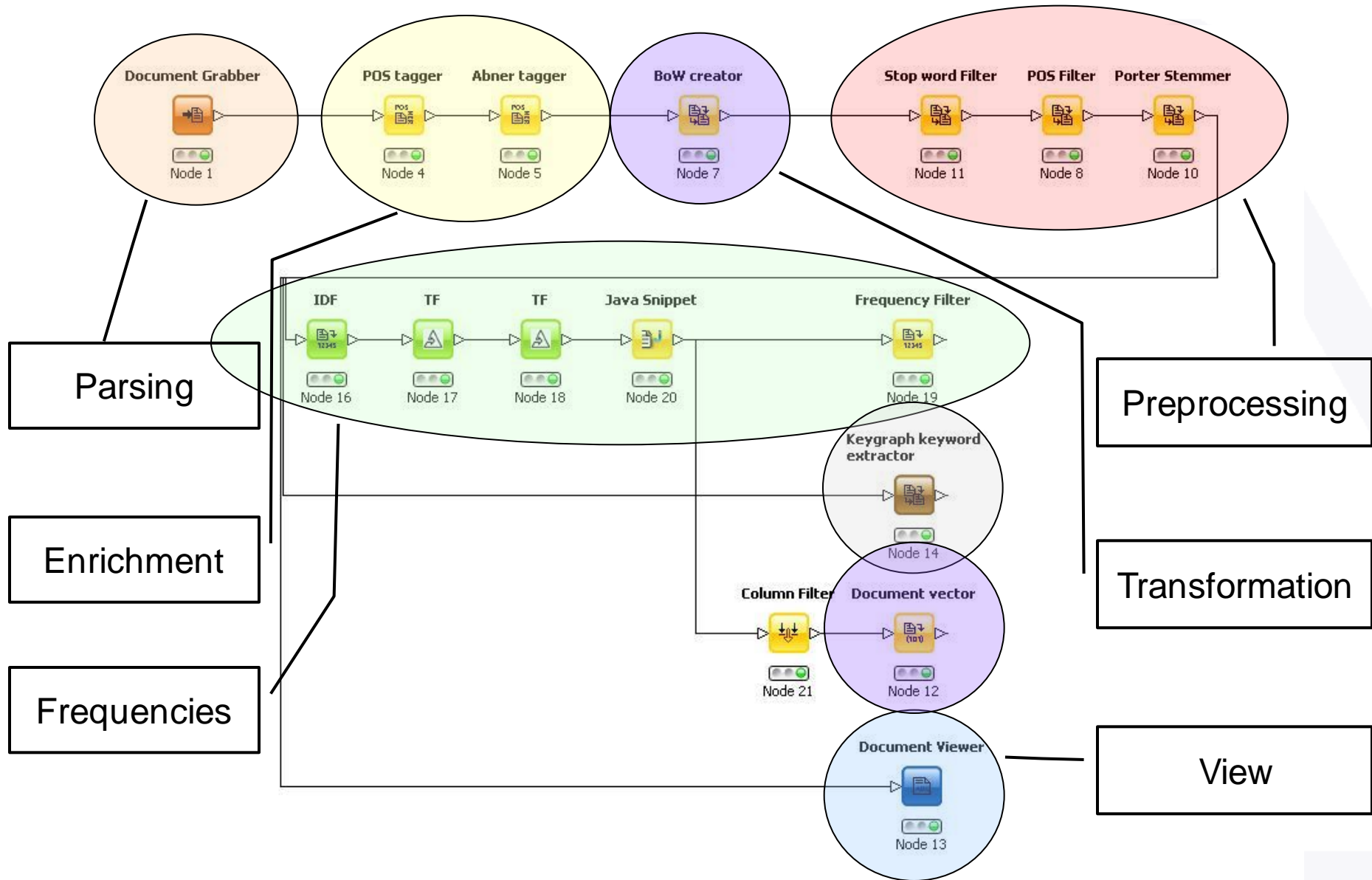
1. Extract Keywords
2. Find named entities
3. Cluster / classify documents
4. Visualization
5. ...



A-induced dissemin origin AIDS-CMV antigen-specific boost cryopyrin  
 IL-15 mechan Mc12 NKLAK calcium-pterin radiotherapi retrovir alphaGSUPTTGxRb toler  
 gliosi CD4+ Pro ROI lymphoma angiogenesi tattoo neutrophil mRNA OM-174 mAFP  
 irradi IRAK thymocyt pancreat tgAAC09 SOCS-1 adult LM8 restor rhesu FGF1 innat load  
 perforin mous OVHM B16-C215 Ad5 breast CXCL10 mtDNA recogn endogen microglia radiat posit immunogen CCL3L1  
 Ets-1-deficient polymorph apoptosi BMT agreem MDM IGRA SIV pgml IL-17 gene laboratoru resolv prolifer GALT ELISPOT AdV  
 viremia lesion Gag-specific B16 NYVAC-C Thd Dunn MAC alpha-GalCer week TDLN fusion interrupt PHPC pmacIL12 human CD8+  
 HIV-specific B16F10 chang method gammadelta IL-18R CD69 interferon cancer transfact local LpICE IL-18bp-Fc PDGF InR amino  
 tumorigenesi IgA HCV-specific cultur GM-CSF-- host Th1 count monkei der gain PBMC beta-sitosterol X-irradiation AcPL pylori IL-1beta ChR acid  
 albican ligand MVA DC-IL18-Lysate HIV-2 stimul skin TST hepat assai Medizin alert us SeV DNA-C popul tumour A11CD40L model IRD  
 plaqu dai GM-CSF metastat abl HIV-1-specific melanoma epitop CD8 control MMW Non-responders scienc escap donor HDC result casei  
 hypersensit anti-CD40 detect BCG Candida level caspase-1 T-cell immun DNA therapi Fortschritt HIVHCV gene-gun SCF exposur QFT-G ag  
 INF-gamma transplant cytokin IL-10 Salmonella bLF CD4 infect respons mice tuberculosi viru vector Assess dermatolog osteosarcoma  
 vCP1452 SPC infant EL-4 viral treatment macrophag HIV tumor cell IL-18 HIV-1 antitumor develop signal secret combinatori QFT  
 glioma neopterin Colon inhibitori resist progress test express cell IL-18 HIV-1 antitumor develop signal secret combinatori QFT  
 intestin interleukin-18 subtyp IFI function HIV-infected CTL IL-12 vaccin activ NKT liver administr sequenc Journal CSF-1 period respond  
 CD40L rIL-18 diagnosi type CMV-specific IL-2 anti-tumor product patient pituitari coinfect NKDC inject IL-23 transform astrocyt antimetastat  
 adapt murin IFNgamma GBV-C fibroblast migrat correl pleural dose individu DC PTTG subject IL-18 cDNA transactiv sarcoma IFN- $\alpha$  VSV  
 plu annual person therapeut radiosensit PEC declin children protect HCV cytotox latent metastasi effus B7-1 Talpha1 baboon stromal  
 cryopreserv B16M associ receptor ASC pregnanc wound rate HSE combin diseas growth CD72 N-2a Fgfr2b acquisit LKR-13 SHIV-896P  
 hypoplasia mucos 26IL-12 colon rmlIL-18 Listeria CD11c IL-21 speci acut liposom Lp18 vitamin macaqu pro-IL-18 studi stage gene-modified  
 process CpG complex RenCaIL-12 tuberculosis-specific CXCL8 toxin trial fungal advanc OK-PSA efficacit intak pEgr-IL-18-B71 DCJ558IL-18  
 low CFP-10 PCR predict regul mortal IRF-1 QFT-2G FB1 D-Fraction IL-7Ralpha IL-27 bacteria EGT T-SPOTTB  
 mDC ESAT-6 blood specif MHV matur conjug AGM treat vervet dual modifi QTF  
 promiscu cellular dendrit inhibit induc C215Fab-SEA hybrid bLFH extract mutat neoforman



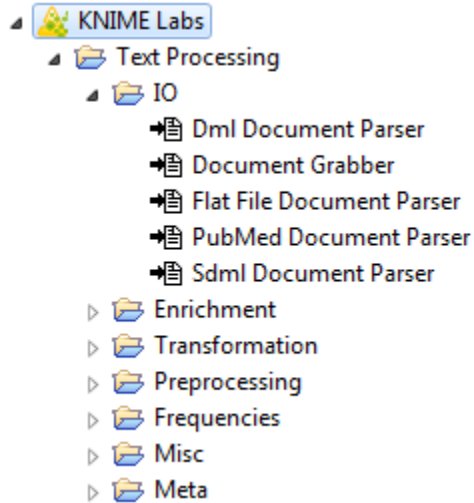
# How ?



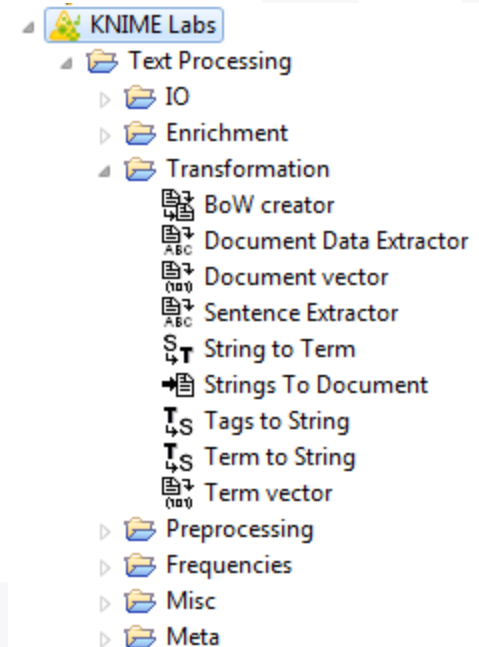


# Nodes

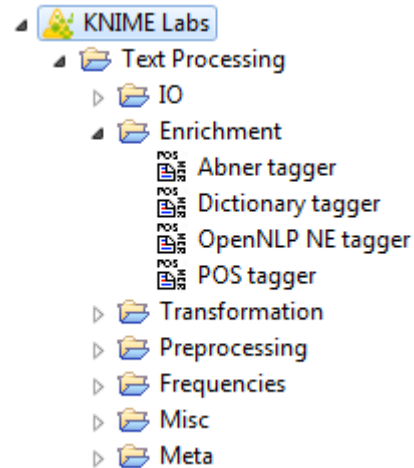
- IO / Parsing



- Transformation



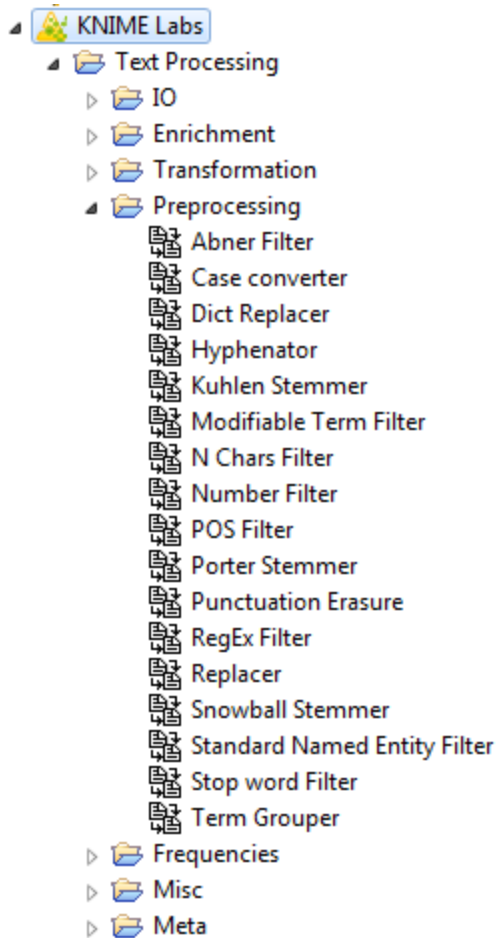
- Enrichment



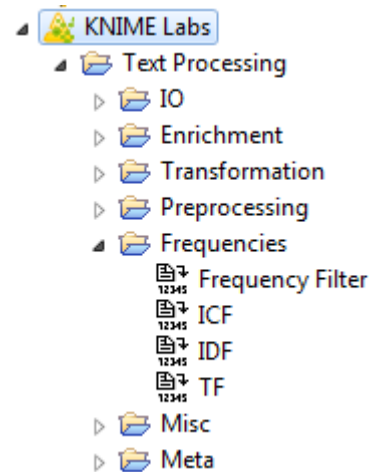


# Nodes

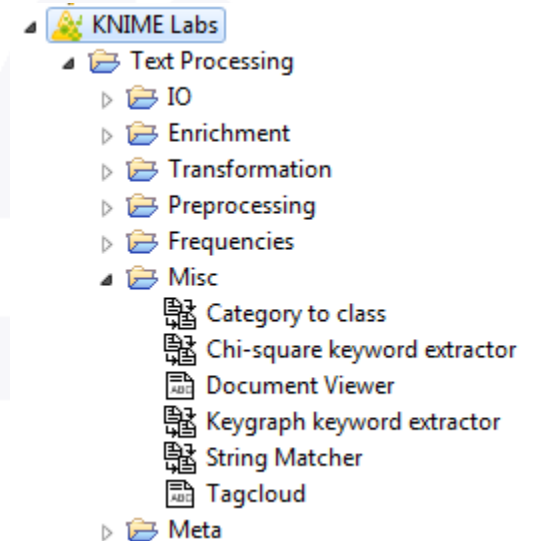
- Preprocessing



- Frequencies

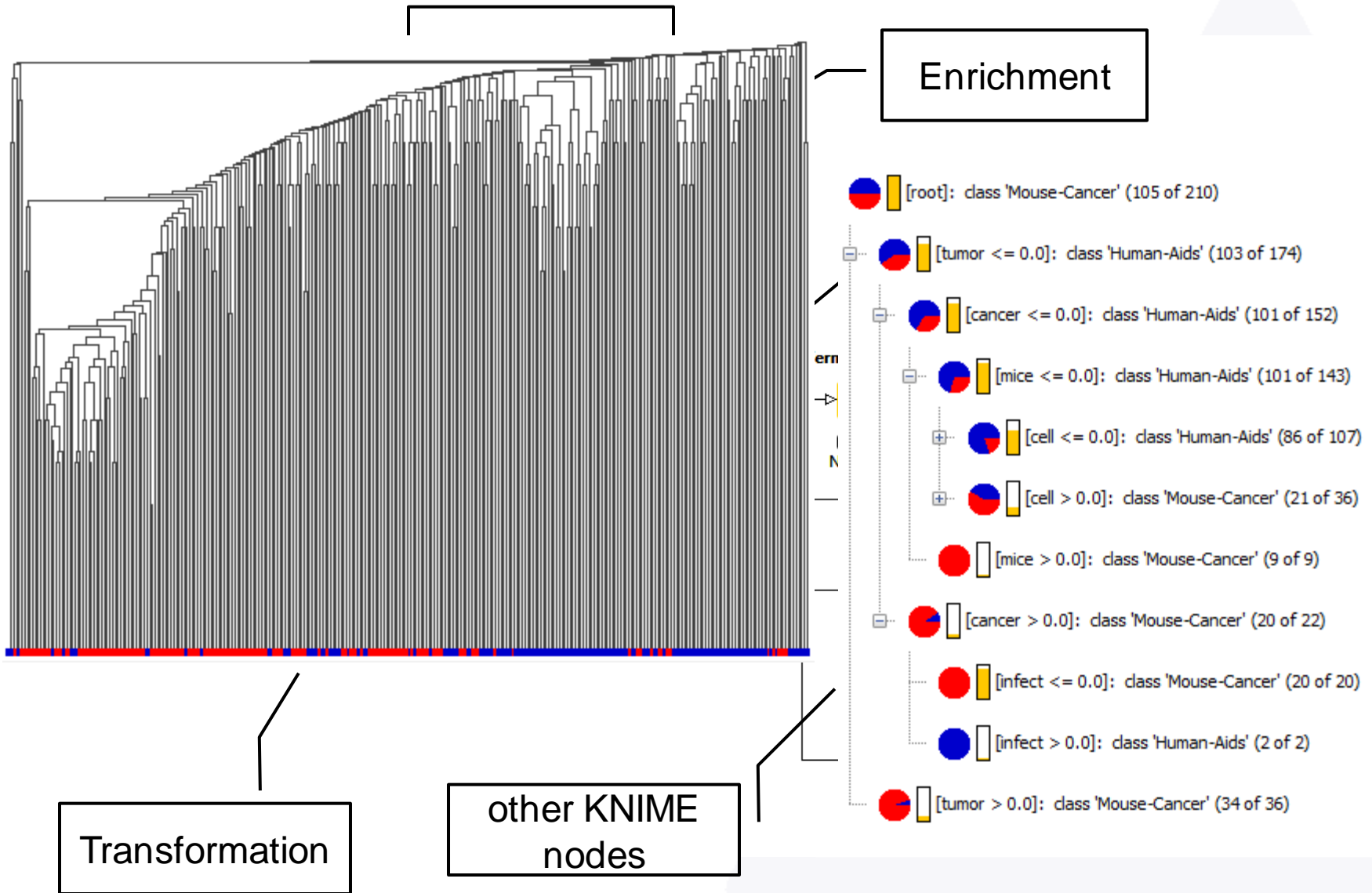


- Misc





# Example





# Example

Document Grabber

Brazil neurolog sex brain B71 P0001 plu exposur CIMT  
 contraceptp005 social cardiomyocyt costunolid PRESS sound CPT-11 lymphocyt abacavir CD3  
 immort E76K vitro condom evalu adenoviru anim seroconvers miR-155 microsporidia impact RiD strategi  
 incid EAC articulatin-D overweight malaria analysi behaviour viral platinum-resistant suppress belief women sheal poverti imag  
 p210BcrAbl VSV sugar psychiatric CD4-negative novel UHL ABC3TC septic OTBC ovarian review CTC preval dendrit metabol gp130  
 meta-analysis LC-MSMS role cytotoxag hormon EFV load Infectioneduc CPT-loadedline estim Amazon level commun lipid cells L  
 detectsignificanti pigment individu steroid melanoma skin Journalreport Immunodeficien advertis -bisabolol complex econom lesion NGO arm  
 D surfac Anip973NVBlymphoma T-cells MV rate methylat PEDIATR TCR CCA STAT1Illness Treglog protein billion ascorb injuri associ capabl  
 Trastuzumab ATvr interventprogress repair Servic HAART studi CHOP system efavirenz Human isol technologi unr-- Unite carcinoid growth  
 inhibitor RNA procoagul F3CE access Special AANATEDNRB CEA Adult diagnosi Parkin ESC Healthcar ecosystemcomposit neuron sequenc  
 cleavag chemokin diagnos 4-HR Transit us Novel microphon HpRecA bone IL-6 Elf5 CD4+ HSV-TK Chronic ecolog carcinogenesi B7-H1 reflect  
 MMTVel Women Viru HIV-infectedexplain PKC pregnanc TXNIP th rapeutique RRM1 platelet caregiv transmit syndrome cisplatin HHV-6 PAD  
 music million pediater GMME1 success ACTH increas Global drug DNA Umschau Ptk1 Poloxin HCV prostat pep attendDuraS popul GPC3  
 transplant CMV AAC cDC woundRat-9 LMP1 p190A HIVAIDS CD26 GM3 HIV-1 Tuberculosis leukemia liver BP1 PHI vulner databas mechan  
 PCEC radiat Pax6 Foxp3 DT-IgG EMILIN1 SKLB703 health PI3K AIDS express HHV8 Therapeutisch test ART cultur MYB famili DIO  
 pDC GATA-3antibodi pancreat MUC4 activ England patient HIV p53 therapi COX-2 Revu research CagA symptom tRNA BMSC  
 SHARPIN muscl aid servic Slit2 model HPV medicin vaccin cell tumor diseas women York MART-1 adult effect PAF  
 chromosom treatmentXMRV PAI-1 lung breast mammari PP5 SMO Scienc colon Ciz1 noncommunic depress c-Myc  
 microspher peptid NYAP initi Advanc Ptpn11 cost A112 immun cancer c-Met Tnp3 resist FWGE replicatur NVB progesteron content  
 gland infant assess antivir birth countConverg hear Neu Sirt1 journal respons infect CD4 start experi arthriti EDIA rectal transmiss  
 MSM HCC mesothelioma CD8 defectCYLD Develop develop Snail1 mice viru NAT1 reward diseases--lessons cluster AhR anal mtDNA primat  
 TA2MSWM in-the-ear mutat bladder inhibit tuberculosis NP MG132 human DRO1 finish development--improvingchildren metastasi smoke CTL  
 megakaryocyt genit intellig ethic ganciclovir differenti apoptosi PA sexual confer HIV-AIDS tumour EMT red cachexia syndemic tamibaroten  
 ileal cytokeratin stigma shortag triterpen Africa AIF SIV risk antiretrovir BRCA1 functionburden pathologiFBHW IL-13R 2 efficien brush  
 tester microbiota endotheli SSAT damag thyroid CVD PFP lineag SSA controlbind her gene Latino gemcitabin mitochondri non-human  
 cellsmm train protectfailur HLA knee cat American G-MDSC integr PnA1 gun microRNA tamoxifen cyclin E characterist nociceptor MSW  
 trimest bite benefitnonrandom format segreg specimen site HIV-2 G-MDSCs Mous intestin PCT melatonin USBHW propoli non-APL MT1MT2  
 updat culture-negativecalcium PBMC pericyt paclitaxel Japan pain MT1-MMP Fansidar surgeri geriater Hispan acut virus plasma virion Warburg  
 statu hUCMSC expectlactat esophag prodomain emodin mesothelin counsel live HIV-LD xenograft follow-up neoplasia memori design J-coupling  
 strand IBM fever 1-integrin inform receptor mast repeatErcc1- combin capsid intraepitheli neural poli  
 feed metastas invert 1-integrins loss repress wear vitamin SHS deriv SKP care Ad enAFP D55-SOCS3  
 mother stratum synthesi agent valuat virolog worker infiltr threatvascular deficien

ing

Frequencies



- Two new data types
  - DocumentBlobCell
    - Encapsulates a document
    - Is implemented as BlobCell
    - Serialization via XML
  - TermCell
    - Encapsulates a term
    - Is implemented as a usual DataCell

| Document                |
|-------------------------|
| "Expression of hum...   |
| "Enrollment fluid st... |

| T Term             |
|--------------------|
| acid[NN(POS)]      |
| analytes[NNS(POS)] |



- Data Table structures

- Document Table

- List of documents

| Row ID | Document   |
|--------|--|
| 1      | "Sonoporation of the Minicircle-VEGF(165) for Wound Healing of Diabetic Mic...     |
| 2      | "Structural basis for midbody targeting of spastin by the ESCRT-II I protein C...  |
| 3      | "Functional Consequences of the Human Leptin Receptor (LEPR) Q223R Tra...          |
| 4      | "Apolipoprotein E Highly Correlates with AbetaPP- and Tau-Related Markers i...     |
| 5      | "The role of 5gk-1 in the upregulation of transport proteins by PPAR-(gamma...     |
| 6      | "Comparison of 3 ad libitum diets for weight-loss maintenance , risk of cardio...  |
| 7      | "Evaluation of various doses of recombinant human thyrotropin in patients w...     |
| 8      | "Inhibition of renal glucose reabsorption : a novel strategy for achieving gluc... |
| 9      | "Attenuation of diabetes-induced retinal vasoconstriction by a thromboxane         |
| 10     | "Cathepsin E regulates the presentation of tetanus toxin C-fragment in PMA         |

- Bag of words

- Tuples of documents and terms

| Row ID | Term              | Document                           |
|--------|-------------------|------------------------------------|
| 1      | acid[NN(POS)]     | "Fluorescence sensors for monos... |
| 2      | analytes[NN...]   | "Fluorescence sensors for monos... |
| 3      | on[IN(POS)]       | "Fluorescence sensors for monos... |
| 4      | spectral[JJ](...] | "Fluorescence sensors for monos... |
| 5      | methods/pr...     | "Fluorescence sensors for monos... |
| 6      | affords[NNS...]   | "Fluorescence sensors for monos... |
| 7      | upon[IN(POS)]     | "Fluorescence sensors for monos... |

- Term / Document vectors

- Vectors representing Documents / terms

| Row ID | Document                | D P<0.01 | D cultur | D RT-PCR | D blood | D NIH3T3 | D Dulbecco | D coloclysi | D Fast |
|--------|-------------------------|----------|----------|----------|---------|----------|------------|-------------|--------|
| 1      | "Expression of hum...   | 0.062    | 0.027    | 0.018    | 0.019   | 0.124    | 0.021      | 0.062       | 0.041  |
| 2      | "Enrollment fluid st... | 0        | 0        | 0        | 0       | 0        | 0          | 0           | 0      |
| 3      | "Hyperinsulinemia F...  | 0        | 0        | 0        | 0       | 0        | 0          | 0           | 0      |
| 4      | "Portal Venous Don...   | 0        | 0        | 0        | 0       | 0        | 0          | 0           | 0      |
| 5      | "The environmenta...    | 0        | 0        | 0        | 0       | 0        | 0          | 0           | 0      |
| 6      | "Angiopietin-1 , b...   | 0        | 0        | 0        | 0.013   | 0        | 0          | 0           | 0      |
| 7      | "CD4( +)CD25 ( + ...    | 0        | 0        | 0        | 0.008   | 0        | 0          | 0           | 0      |
| 8      | "Overexpression o...    | 0        | 0        | 0.019    | 0       | 0        | 0          | 0           | 0      |
| 9      | "Diabetes and hear...   | 0        | 0        | 0        | 0       | 0        | 0          | 0           | 0      |



# Outlook

- Tackle memory issues
- Easy integration of third party tag types
- More nodes and features (i.e. chem tagger)

