

# Introduction to text mining and information retrieval applications for biology

**Patrick Ruch**

**HEG - HES-SO Genève & SIB Text Mining**

CUSO Workshop, June 3-5, 2015 CMU, H

[patrick.ruch@hesge.ch](mailto:patrick.ruch@hesge.ch)

# Workplan

## ⌘ AM – Focus on evaluations (Julien Gobeill)

- ☑ Evaluation metrics: Information Retrieval and Text Categorization
- ☑ Competitions: TREC 2014...

## ⌘ PM – Focus on applications (Patrick Ruch)

- ☑ Information Retrieval and related tasks, incl. Question-Answering
- ☑ Applications, incl. Question Answering

# Round table...

# Requirements

## ⌘ Data mining

- ☑ Distance

- ☑ Machine learning

  - ☑ Classifiers

  - ☑ Clustering

## ⌘ Database

- ☑ Data types

- ☑ Queries (SQL...)

- ☑ Indexes...

# Objectives

- ⌘ Being able to define the main concepts
  - ☑ supporting text mining applied to biology
  - ☑ to understand and to actively participate in day 2 & 3  
[https://docs.google.com/document/d/1Xa1KWZZowjsnR9gPybP\\_L3kMWe\\_nDxQfYGvwJJM9QEe0/edit?usp=sharing](https://docs.google.com/document/d/1Xa1KWZZowjsnR9gPybP_L3kMWe_nDxQfYGvwJJM9QEe0/edit?usp=sharing)
- ⌘ Being able to describe the main market players
- ⌘ Define the main tasks performed with text mining
- ⌘ Evaluation [AM]

# How important is search ?

Information retrieval ~ 25% of labour time 😊

Why\_you\_can\_t\_just\_Google\_for\_Enterprise\_Knowledge.pdf  
[Oct. 2013]

# Market

⌘ “Market share in information retrieval”

⌘ Market share: Google, Autonomy, Lucene, FAST...

Old and biased sources:

<http://www.ideaeng.com/enterprise-search-matrix-0206>

<http://www.domorewithsearch.com/microsoft-fast-versus-google-search-appliance-6-8/>,  
written by FAST !

# DBMS

## ⌘ According to Gartner (2008)

- ⌘ Oracle Database - 70%
- ⌘ Microsoft SQL Server - 68%
- ⌘ MySQL (Oracle Corporation) - 50%
- ⌘ IBM DB2 - 39%
- ⌘ IBM Informix - 18%
- ⌘ SAP Sybase Adaptive Server Enterprise - 15%
- ⌘ SAP Sybase IQ - 14%
- ⌘ Teradata - 11%













[http://en.wikipedia.org/wiki/Relational\\_database\\_management\\_system](http://en.wikipedia.org/wiki/Relational_database_management_system), octobre 2013



# DBMS

⌘ According to DB-Engines, the most popular systems are Oracle, MySQL, Microsoft SQL Server, PostgreSQL and IBM DB2.<sup>[3]</sup>

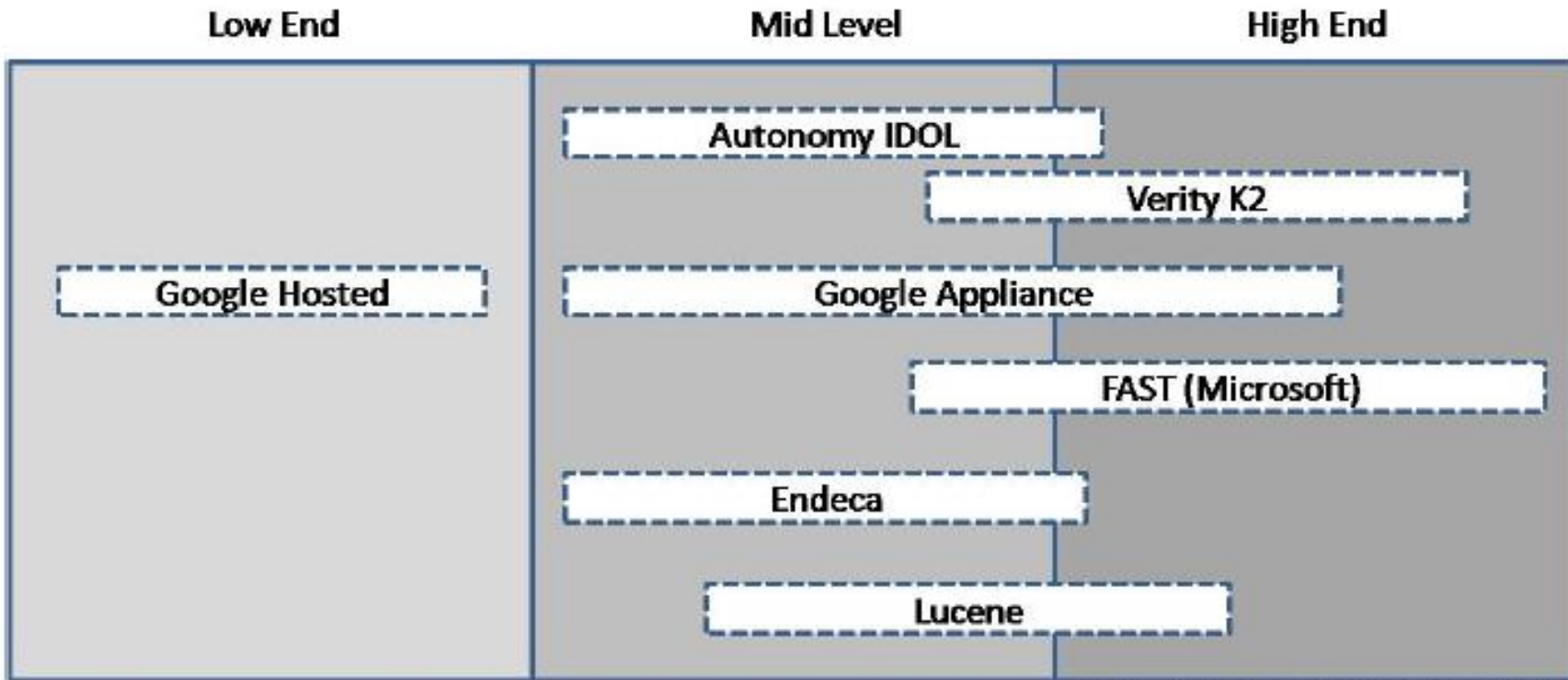
# Web search

Search engine ↕	Market share in May 2011 ↕		Market share in December 2010 <sup>[14]</sup> ↕	
Google	82.80%		84.65%	
Yahoo!	6.42%		6.69%	
Baidu	4.89%		3.39%	
Bing	3.91%		3.29%	
Yandex	1.7%		1.3%	
Ask	0.52%		0.56%	
AOL	0.3%		0.42%	

[http://en.wikipedia.org/wiki/Web\\_search\\_engine](http://en.wikipedia.org/wiki/Web_search_engine)

Oct 2013

# Non-web search

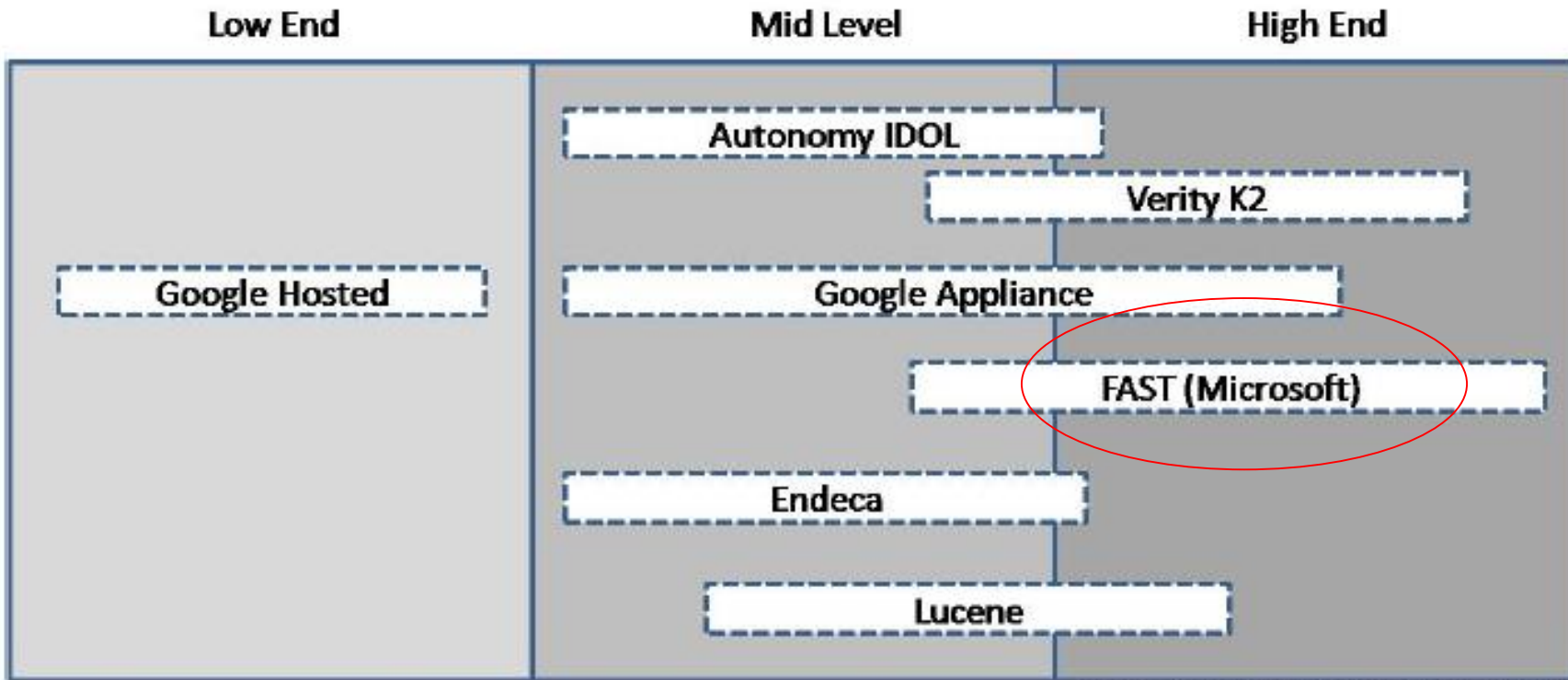


*Source: New Idea Engineering*

<http://www.ideaeng.com/enterprise-search-matrix-0206>

Oct 2013

# Bullshit...



Source: New Idea Engineering

<http://www.ideaeng.com/enterprise-search-matrix-0206>

Oct 2014

# My opinion

## ⌘ Web search : Google

Good for web contents and **only** web content  
[Cf. GE 😊]

## ⌘ Entreprise search

- ☑ Google box

- ☑ Autonomy (HP)

- ☑ Fast (MS)

- ☑ they include EDM + CMS..., e.g. Sharepoint

# Professional – yet – open source tools

⌘ Lucene ecosystems (Hadoop/HDFS, Lucidwork...)

☑ SOLR, ElasticSearch...

⌘ Other (with better ranking functions)

☑ Lemur

☑ Terrier

☑ Indri

☑ MG4J...

☑ Weka, Clementine... ok for toy collections [Data mining platform]

# Other players

## ⌘ Search engine/Analytics + **Contents**

### ☒ Business

☒ Thomson Reuters

☒ LexisNexis

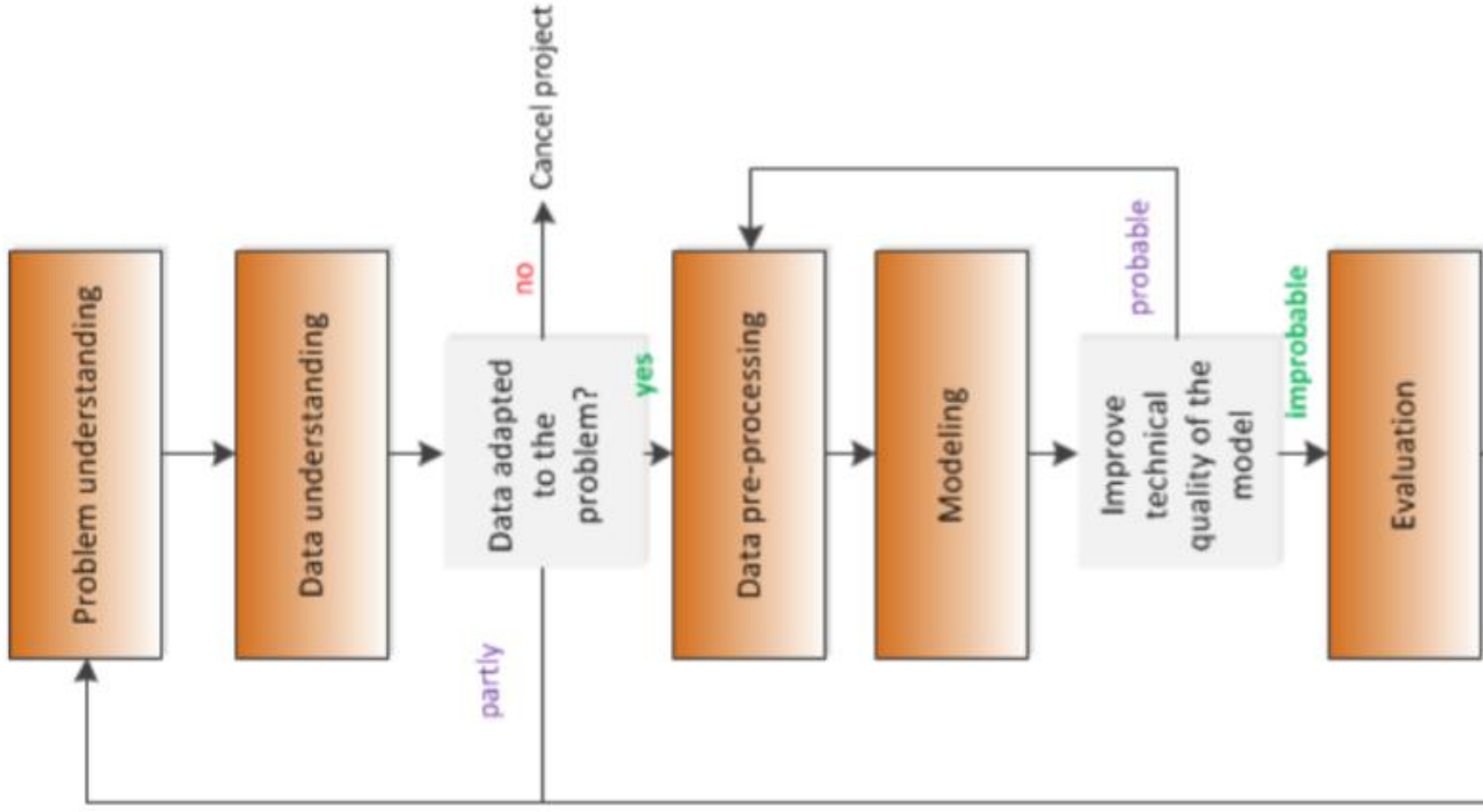
☒ Bloomberg

### ☒ Sciences

☒ Elsevier

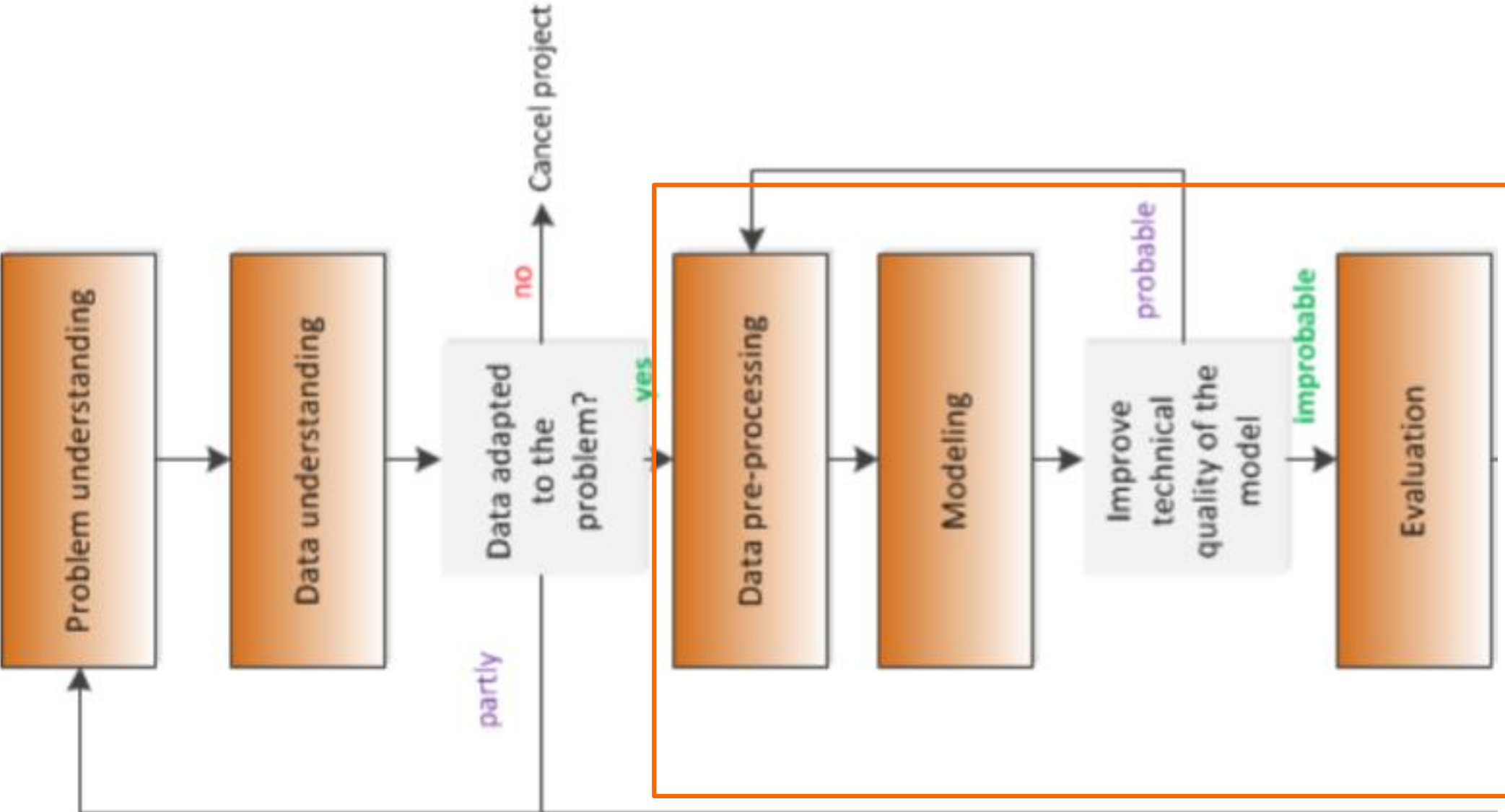
☒ Springer

# Text mining





# Text mining



# Text Mining

## ⌘ Data mining for unstructured data

☑ Small storage → Small Data → Relational DB

☑ Data mining

☑ Large storage capacity, Text collection, Web

☑ **Big data created to address text mining challenges !**

# Distances

⌘ Euclid and non Euclid distances

[http://fr.wikipedia.org/wiki/Distance\\_%28math%C3%A9matiques%29](http://fr.wikipedia.org/wiki/Distance_%28math%C3%A9matiques%29)

# Similarity

=  $1/\text{distance}$

So I will often use either distance or similarity measures

# Information retrieval

Determine the best distance/similarity to rank documents depending to a query

# Machine learning

- ⌘ Compute methods to perform classification and regression tasks

Sentiment:

[http://www.csc.ncsu.edu/faculty/healey/tweet\\_viz/tweet\\_app/](http://www.csc.ncsu.edu/faculty/healey/tweet_viz/tweet_app/)

# Training/tuning data

- ⌘ What is needed to learn or tune a task
- ⌘ Text mining → Training
- ⌘ Information retrieval → Tuning

# Example

⌘ Annotated or automatically annotated

Sentiment...

e.g. 😊 → Positif

☹️ → Negatif

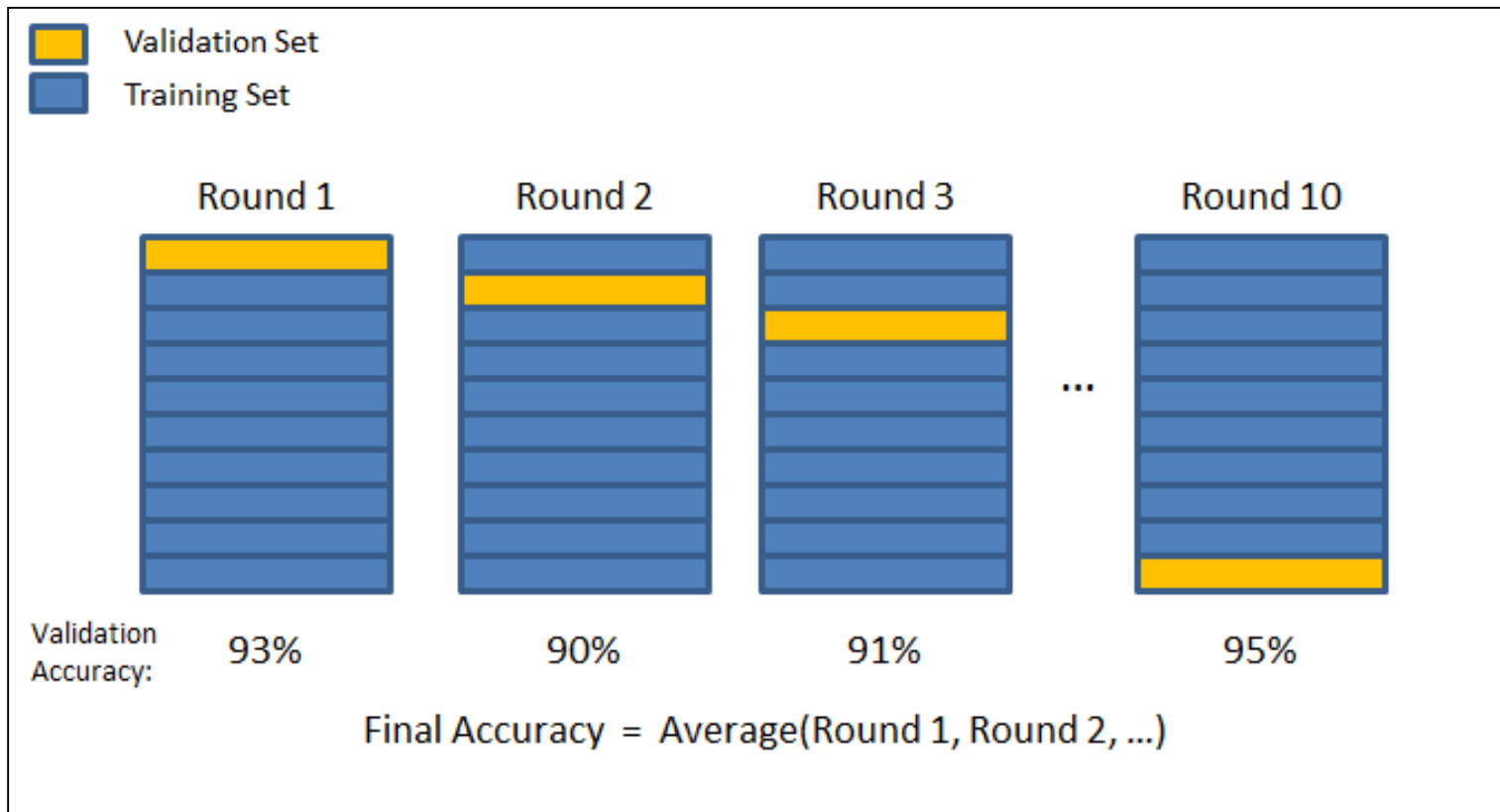


# Pipeline

⌘ Training set

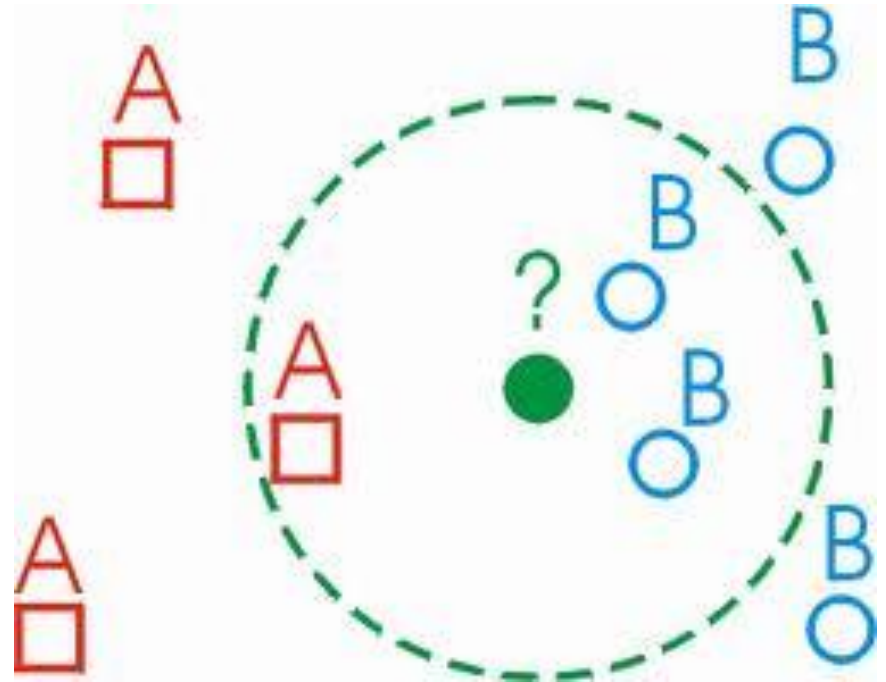
⌘ Test set

⌘ Validation/**Cross-validation**/Leave-one out



# Different classifiers

- ⌘ **K nearest neighbors**
- ⌘ Support Vector Machine
- ⌘ Neuronal networks
- ⌘ Naive Bayes
- ⌘ Decision tree



# Mainly two tasks

⌘ Information retrieval → Cranfield paradigm !

[Metrics: AM]

⌘ Classification

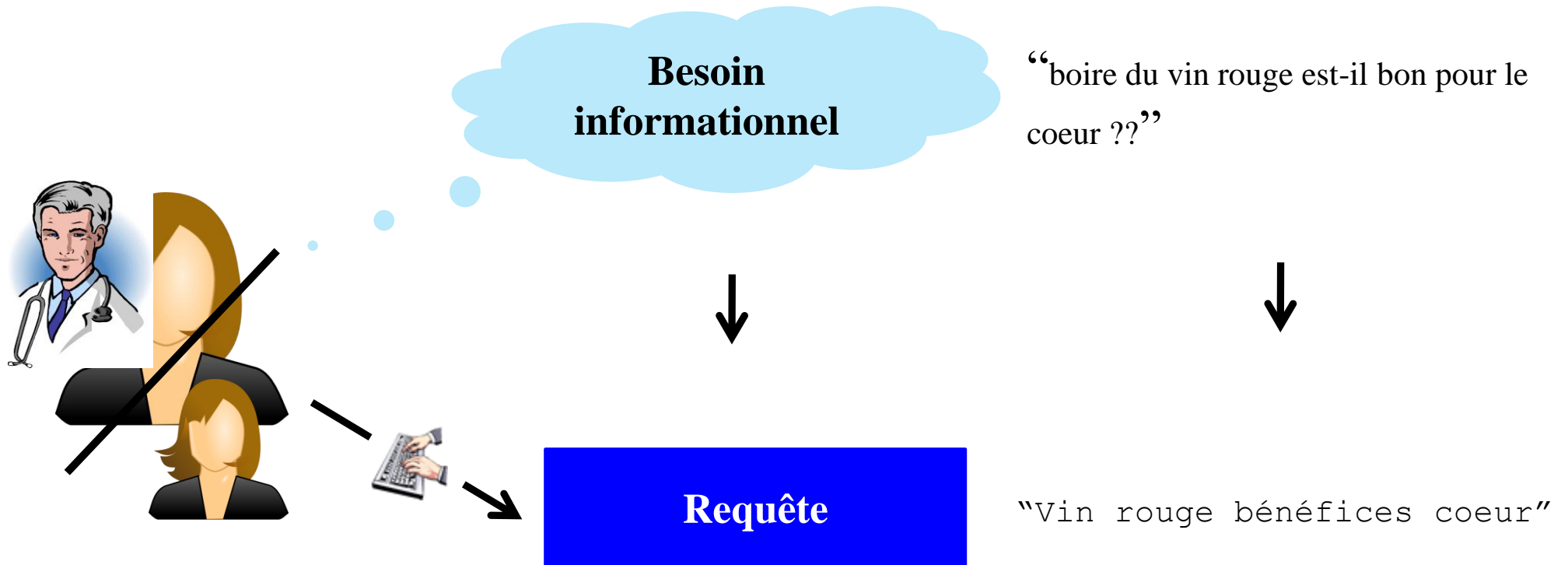
☑ binary

☑ multiclass

# Pertinence

⌘ Dictionary : « what is appropriate » !

⌘ Information needs :



# What is the most pertinent ?

**Femme  
Actuelle.fr**

**Les bienfaits du vin**  
*par Claire Frayssinet,*



Consommé avec modération, le vin est un véritable élixir de jeunesse. Il aurait des propriétés anticancéreuses et protégeraient des maladies cardiovasculaires. Aujourd'hui, boire un petit coup n'est plus forcément tabou.

De nombreuses enquêtes d'épidémiologie réalisées au cours des 35 dernières années dans les pays industrialisés ont démontré que les populations consommatrices de vin présentaient des taux bas de mortalité pour les maladies cardiovasculaires. Certaines études suggèrent que le vin pourrait diminuer de 40% les risques d'infarctus du myocarde et de 25% les risques de thromboses vasculaires cérébrales.

**PERTINENT**

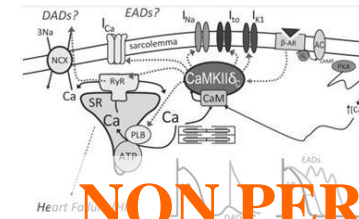
**NON PERTINENT**



Journal of  
**Cardiovascular  
Pharmacology**

**Grapes, Wines, Resveratrol and Heart Health**  
*Bertelli, Alberto, PhD.; Das, Dipak K PhD, ScD*

Epidemiological and experimental studies have revealed that a mild-to-moderate drinking of wine, particularly red wine, attenuates the cardiovascular, cerebrovascular, and peripheral vascular risk. However, the experimental basis for such health benefits is not fully understood. The



cardioprotective effect of wine has been attributed to both the alcoholic portion and more importantly, the alcohol-free portion containing antioxidants. Wines are manufactured from grapes, which also contain a large variety of antioxidants including resveratrol, catechin, epicatechin and proanthocyanidins. Resveratrol is mainly found in the grape skin, while proanthocyanidins are found only in the

**NON PERTINENT**

**PERTINENT**

?



?

← “python” →

```
VIM - ~/work/popub/poptool.py
File Edit Window IDE Syntax Help

# query the POP-3 host, grab the key headers and print them to stdout
def getPopInfo(host, user, passwd, port = poplib.POP3_PORT):
    M = poplib.POP3(host, port)
    M.user(user)
    M.pass_(passwd)

    numMessages, boxSize = M.stat()

    print "\nRetrieving info from POP mailbox at %s (port %d) for %s \\"
          % (host, port, user)
    print "%d messages (%dk)" % (numMessages, boxSize/1024)

    for j in range(1, numMessages + 1):
        info = {}
        for field in HEADER_LIST:
            info[field] = field

        print
        header = M.top(j, 0)
        for line in header[1]:
            fields = string.split(line)
            try:
                keyWord = fields[0]
                if keyWord in HEADER_LIST:
                    info[keyWord] = line
            except IndexError:
                pass

        print "Message #%d" % (j)
        for field in HEADER_LIST:
            if (info[field] == field) and (SKIP_NULL):
                continue
            else:
                print "%s%s" % (SPACER, info[field])

84,0-1
```

→ User-specific meaning

# Solution

## ⌘ Benchmark

- ☑ Queries (min. 25-40)
- ☑ Collection of documents
- ☑ A list of associations, so-called relevance judgements

# Relationship between Information Retrieval and Automatic text categorization

- ⌘ Indexing of annotated collection
- ⌘ Building of a k-NN
- ⌘ Automatic text categorization



# Information retrieval

## ⌘ Theoretical definition

- ☒ Find documents which correspond to a user information need

## ⌘ Practical definition

- ☒ A task performed by users in order to find access some information content

# Problems

⌘ Lexical ambiguity (polysemy), cf. “Python”

English: *Time flies like an arrow; fruit flies like a banana*

French: *La belle ferme le voile !*

⌘ Synonyms

*hepatitis = liver inflammation*

# Several tasks

Defined by the Text Retrieval Conferences

- ⌘ Ad hoc
- ⌘ Related doc search
- ⌘ Passage retrieval
- ⌘ Automatic text categorization
- ⌘ Known-item search [specific metrics]
- ⌘ Information extraction
- ⌘ Question-answering

# Ad hoc: natural language → documents

« fusion »



≠



fusion dans l'industrie automobile - Recherche Google - Mozilla Firefox

Fichier Édition Affichage Historique Marque-pages Outils ?

Google fusion dans l'industrie automobile Rechercher Recherche avancée Préférences

Rechercher dans :  Web  Pages francophones  Pages : Suisse

Web Résultats 1 - 10 sur un total d'environ 214 000 pour fusion dans l'industrie automobile (0,28 secondes)

[Mondialisation automobile: Une industrie automobile en "FUSION ...](#)  
On parle donc de **fusions** strictement d'affaires pour les deux parties. Si l'on remonte au début des années 70, l'**industrie automobile** mondiale ne comptait ...  
[www.guideauto.com/nouvelles/actualites/2531 auto45/mondialisation-automobile-industrie-automobile-fusion... - 39k - En cache - Pages similaires](#)

[Utilisation d'installations de fusion et de revenu innovantes pour ...](#)  
Utilisation d'installations de **fusion** et de revenu innovantes pour la réalisation de pièces en aluminium haut de gamme pour l'**industrie automobile** ...  
[cat.inist.fr/?aModele=afficheN&cpsid=16311553 - Pages similaires](#)  
de H JOHNEN - 2004

[Regroupements et séparations d'entreprises | Michel Freyssenet](#)  
Il revisite pour ce faire l'histoire de l'**industrie automobile** depuis la ... L'article relève que les grandes **fusions** ou alliances transnationales de la ...  
[freyssenet.com/?q=node/446 - 14k - En cache - Pages similaires](#)

Liens commerciaux

[Unternehmensfusion](#)  
Merrill DataSite™ optimiert Ihre Prozesse bei Fusionen.  
[www.merrilldatasite.ch/fusion](#)

[Gamme Fusion Encounter](#)  
Chez Mecatechnic, toute la gamme **Fusion** Encounter pour votre voiture  
[www.mecatechnic.com](#)

# Ad hoc

<http://casimir.hesge.ch:8088/solr/collection1/browse/>

# Variation: Cross-lingual IR !

Babel Fish (Systran pour AltaVista, 2007)

alcoholic fatty liver → foie gras alcoolique

[Stéatose hépatique alcoolique]

## ExGallia - La Boutique des Français du Monde

... vos abonnements partout dans le monde chez vous !

**Foie-gras** truffé du Lot, confit de canard à l'... Apéritif 100% naturel obtenu par la fermentation **alcoolique** de 6 à 700 fleurs par bouteilles. A ...

[www.exgallia.com/produits-francais02.htm](http://www.exgallia.com/produits-francais02.htm) • [Translate](#)

→ Ratio = 60% → 92%

# Image search

- ⌘ Often: Search in descriptions associated with images (flickr)
  - ☑ Meta-data, ~ authors, volumes, ...
- ⌘ Content-image retrieval
  - ☑ Color/grey histograms

	image <sub>Query</sub>	image <sub>1</sub>	image <sub>2</sub>	image <sub>3</sub>
R	100	25	99	100
V	1	200	0	1
B	50	67	49	40

# Related article search

Related Articles for PubMed (Select 18689271) - PubMed Results - Mozilla Firefox

Fichier Édition Affichage Historique Marque-pages Outils ?

1: [Hassanein NM, Talaat RM, Bassiouny K, Hamed MR.](#) Related Articles, Links

Influence of protein malnutrition on induction and treatment of inflammation in mice.  
Egypt J Immunol. 2006;13(2):49-60.  
PMID: 18689271 [PubMed - indexed for MEDLINE]

2: [Vigil SV, de Liz R, Medeiros YS, Fröde TS.](#) Related Articles, Links

Efficacy of tacrolimus in inhibiting inflammation caused by carrageenan in a murine model of air pouch.  
Transpl Immunol. 2008 Apr;19(1):25-9. Epub 2008 Jan 24.  
PMID: 18346634 [PubMed - indexed for MEDLINE]

3: [Spiller F, Alves MK, Vieira SM, Carvalho TA, Leite CE, Lunardelli A, Poloni JA, Cunha FO, de Oliveira JR.](#) Related Articles, Links

Anti-inflammatory effects of red pepper (*Capsicum baccatum*) on carrageenan- and antigen-induced inflammation.  
J Pharm Pharmacol. 2008 Apr;60(4):473-8.  
PMID: 18380920 [PubMed - indexed for MEDLINE]

4: [Agha AM, El-Khatib AS, Kenawy SA, Khayyal MT.](#) Related Articles, Links

The influence of carbon tetrachloride-induced liver damage on the inflammatory reaction elicited by carrageenan and its treatment with diclofenac.

**Medline, patent libraries...**



# KWIC key-word in context / Passage Retrieval

Web [Images](#) [Vidéos](#) [Maps](#) [Actualités](#) [Groupes](#) [E-mail](#) [plus](#) ▼

   [Recherche](#) [Préférences](#)

Rechercher sur le Web  Rechercher les pages en français

---

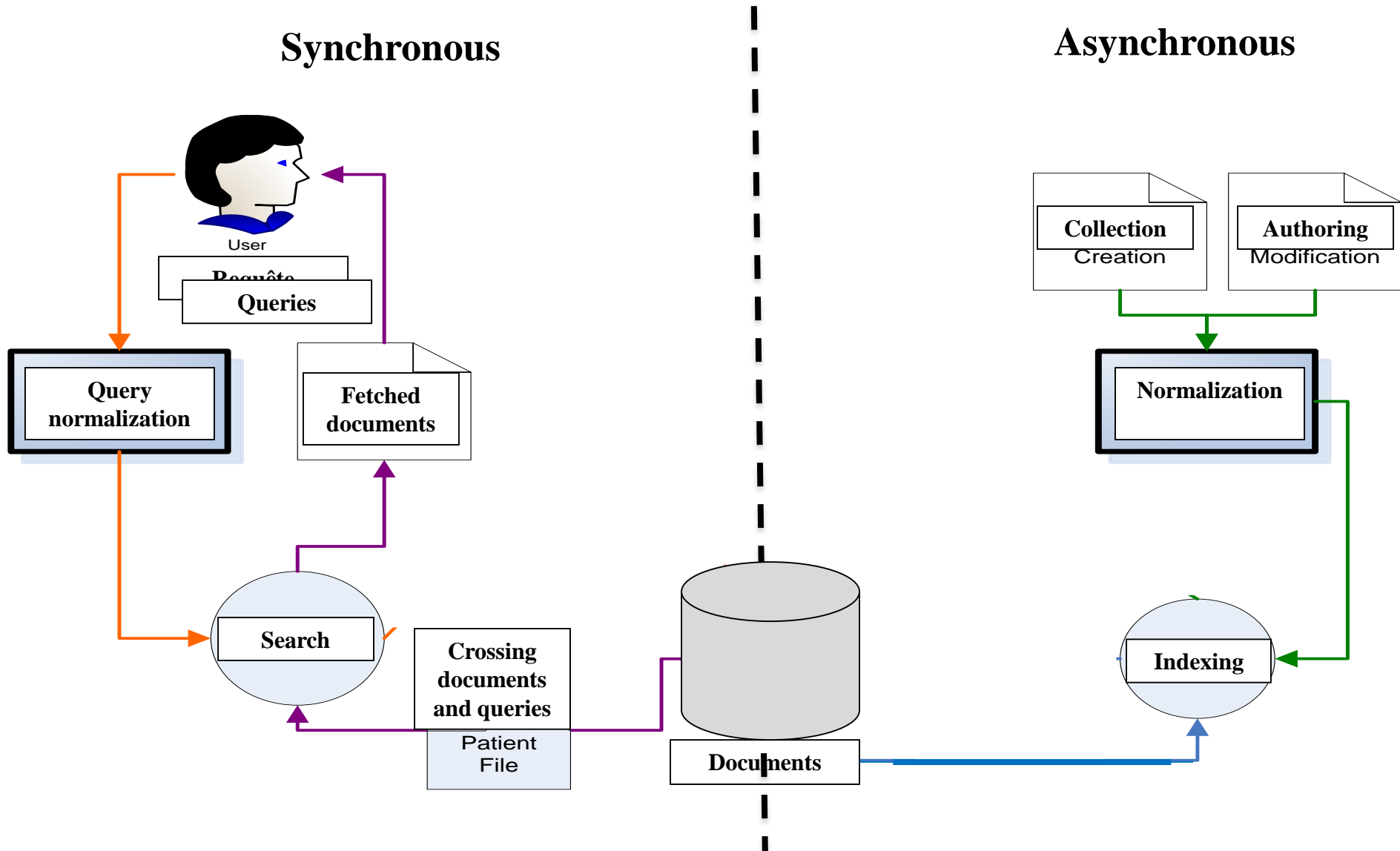
Web Résultats 1 - 10 sur un t

[Le vin rouge protège le coeur](#)  
Boire deux à trois verres de **vin rouge** par jour est bon pour le **coeur**.  
[www.medicms.be/dt3/vinrouge.htm](http://www.medicms.be/dt3/vinrouge.htm) - [En cache](#) - [Pages similaires](#) -   

[Health and Food - article : Coup de coeur pour le coup de rouge](#)  
La saga du **vin rouge** et de ses effets « cardioprotecteurs » se poursuit, ... d'une alimentation saine pour le **coeur**, qu'elle soit méditerranéenne ou pas. ...  
[www.healthandfood.be/.../vin\\_coeur\\_prevention.htm](http://www.healthandfood.be/.../vin_coeur_prevention.htm) - [En cache](#) - [Pages similaires](#) -   

[Extenso | Échelle de crédibilité scientifique](#)  
Cela dit, deux substances sont souvent mentionnées comme étant à l'origine de l'effet bénéfique du **vin rouge** sur les maladies du **coeur** : l'alcool comme tel, ...  
[www.extenso.org/echelle\\_credibilite/.../1327](http://www.extenso.org/echelle_credibilite/.../1327) - [En cache](#) - [Pages similaires](#) -   

# Architecture I



# Architecture II

## ⌘ Synchronous: Processing of query

- ☑ Tokenization of query (word splitting)
- ☑ Filtering of stop words/Normalisation (stemming)
- ☑ Weighting (of words)
- ☑ Ranking of fetched documents

## ⌘ Asynchronous

- ☑ Tokenization of document
- ☑ Filtering of stop words/Normalization (stemming)
- ☑ Indexing
  - ☑ Inverted file
  - ☑ Building of statistical model

# Models of RI

## ⌘ Boolean

☑  $x \text{ AND } y \text{ AND } z; x \text{ OR } y \text{ OR } z$

## ⌘ Vector-space (probabilistic, statistical...)

☑  $X \ Y \ Z \sim X \text{ OR } Y \text{ OR } Z + \text{Ranking}$

☑ TF (term frequency): the more frequent in a document, the more strongly associated with the document...

☑ IDF (inverse document frequency): the more frequent in the collection, the less strongly discriminant for a document...

## ⌘ Hybrid system (+ hyperlinks)

☑ Web search engines

# Indexes

## ⌘ ... Shakespeare

- 📦 *tokens* : « Hong Kong », « O'Neill », Food, Fed...
- 📦 *stop words* : remove « the », « are », « but »...
- 📦 *normalisation (stemming)* : {« rarely », « rare »...} -> « rare »
- 📦 *weighting*

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth	...		Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth	...
Antony	1	1	0	0	0	1		Antony	9	2	0	0	0	1	
Brutus	1	1	0	1	0	0		Brutus	2	5	0	1	0	0	
Caesar	1	1	0	1	1	1		Caesar	6	12	0	2	1	2	
Calpurnia	0	1	0	0	0	0		Calpurnia	0	3	0	0	0	0	
Cleopatra	1	0	0	0	0	0		Cleopatra	8	0	0	0	0	0	
mercy	1	0	1	1	1	1		mercy	4	0	2	3	3	5	
room	1	1	1	1	1	1		room	21	29	32	25	14	33	
...								...							



→ **TF (local to document)**

→ **IDF (collection-wide)**

# Weighting schema

<http://nlp.stanford.edu/IR-book/pdf/01bool.pdf>

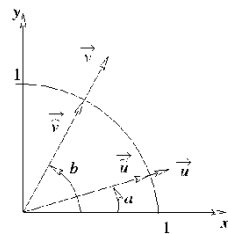
⌘ To go further: query « Mercy of Brutus for Caesar »

⌘ *Vector space:*

⌘ document vectors

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth	...
Antony	9	2	0	0	0	1	
Brutus	2	5	0	1	0	0	
Caesar	3	12	0	2	1	2	
Calpurnia	0	3	0	0	0	0	
Cleopatra	8	0	0	0	0	0	
mercy	4	0	2	3	3	5	
worser	21	29	32	25	14	33	
...							

⌘ ... cosine and other distances :



# Weighting schema

example, Okapi BM25: **tf, df, dl**

$$Score(D_i, Q) = \sum_{t_j \in q} qtf \cdot \log \left( \frac{n - df_j}{df_j} \right) \cdot \frac{(k_1 + 1) \cdot tf_{ij}}{K + tf_{ij}}$$

where  $K = k_1 \cdot [(1 - b) + b \cdot (l_i / avdl)]$

# Clustering/faceting

Query:

3637 results

## Facets

**url\_job**

- twitter.com (491)
- twitter.com (331)
- twitter.com (289)
- twitter.com (179)
- twitter.com (153)

**source\_type**

- twitter (3623)
- rss (14)

## Facets Date

**last\_modified**

- 2012-10-13 (0)
- 2012-10-14 (0)
- 2012-10-15 (0)
- 2012-10-16 (0)
- 2012-10-17 (0)
- 2012-10-18 (0)

**date**

- 2012-10-13 (27)
- 2012-10-14 (53)
- 2012-10-15 (39)
- 2012-10-16 (29)
- 2012-10-17 (21)
- 2012-10-18 (0)

## Clusters

- [Buy Cialis](#)
- [Generic Viagra](#)
- [Other Topics](#)

**"... If Viagra or**

Last modified: Tue Sep 11 20:33:07 CEST 2012

... "... If **Viagra** or Cialis have let you down..." #commercial ...

<https://twitter.com/thejameslynch/statuses/245590812732317696>

**Viagra online: Viagra online.**

Last modified: Sun Sep 02 20:14:12 CEST 2012

... **Viagra** online: **Viagra** online. , Generic cialis, >:-), **Viagra**, =-PP, Cialis online, 70697, **Viagra** ...

<https://twitter.com/viagraviagras/statuses/242324560811286528>

**Generic viagra: Generic viagra.**

Last modified: Fri Sep 07 23:20:43 CEST 2012

... Generic **viagra**: Generic **viagra**. , Online **viagra**, 5660, **Viagra**, nwlx, Cialis without a prescription ...

<https://twitter.com/viagraviagras/statuses/244183438603714561>

**Generic viagra: Generic viagra**

Last modified: Fri Sep 07 23:22:20 CEST 2012

... Generic **viagra**: Generic **viagra** , Online **viagra**, 5660, **Viagra**, nwlx, Cialis without a... Buy Cialis ...

<https://twitter.com/CialisCiali/statuses/244183845103075328>

**Generic viagra: Generic viagra**

Last modified: Fri Sep 07 23:37:07 CEST 2012

... Generic **viagra**: Generic **viagra** , Online **viagra**, 5660, **Viagra**, nwlx, Cialis without a Buy Cialis ...

<https://twitter.com/oncialisline/statuses/244187565496864768>



# Information extraction with “is a”, “is an”...

Example: create wikidictionary !

Your question was : *what is retinoblastoma ?*, reformulated as *retinoblastoma*



*A malignant tumor arising from the nuclear layer of the retina that is the most common primary tumor of the eye in children. The tumor tends to occur in early childhood or infancy and may be present at birth. The majority are sporadic, but the condition may be transmitted as an autosomal dominant trait. Histologic features include dense cellularity, small round polygonal cells, and areas of calcification and necrosis. An abnormal pupil reflex (leukokoria); NYSTAGMUS, PATHOLOGIC; STRABISMUS; and visual loss represent common clinical characteristics of this condition. (From DeVita et al., Cancer: Principles and Practice of Oncology, 5th ed, p2104)*

Keyword in context :

ToxiCat on the selected articles (Beta) :

Score



## High-Resolution MR Imaging of the Orbit in Patients with Retinoblastoma.

[Rauschecker AM](#) , [Patel CV](#) , [Yeom KW](#) , [Eisenhut CA](#) , [Gawande RS](#) , [O'Brien JM](#) , [Ebrahimi KB](#) , [Daldrup-Link HE](#)  
*Radiographics*. 2012 Sep; 32(5): 1307-26  
Pmid : 22977020

PubMed

Retinoblastoma is the most common intraocular childhood malignancy, with a prevalence of one in 18,000 children younger than 5 years old in the United States ... In 80% of patients, retinoblastoma is diagnosed before the age of three, and in 95% of patients, retinoblastoma is diagnosed before the age of five.

Score

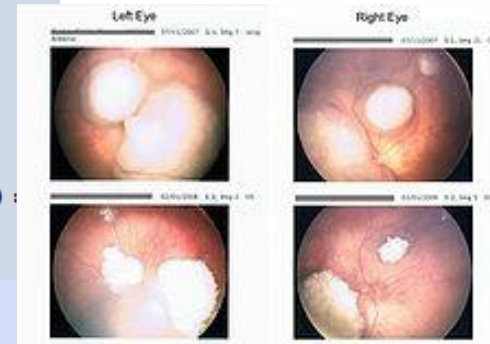


## Current therapy and recent advances in the management of retinoblastoma.

[Meel R](#) , [Radhakrishnan V](#) , [Bakhshi S](#)  
*Indian J Med Paediatr Oncol*. 2012 Apr; 33(2): 80-8  
Pmid : 22988349

PubMed

Retinoblastoma is the most common intraocular malignancy in children ... The survival in case of extraocular retinoblastoma is still low, and the reported survival rate ranges between 50% and 70%.



# Question-Answering

EAGLi: <http://eagl.unige.ch/EAGLi/>

The screenshot shows the EAGLi website interface within a Mozilla Firefox browser window. The browser's title bar reads "EAGLi: the EAGL project's biomedical question answering and information retrieval interface - Mozilla Firefox". The address bar shows the URL "http://eagl.unige.ch/EAGLi/". The website features a header with the EAGLi logo, which includes an eagle's head and the text "EAGLi Engine for question-Answering in Genomics Literature". A search bar is present with a dropdown menu for "Search Engine examples" and a "Query" input field. To the right, a "Question-Answering examples" dropdown menu is open, displaying a list of sample questions such as "what diseases are associated with brca1?". Below the search bar, there are radio buttons for "EAGLi" and "PubMed", with "PubMed" selected. A link for "How to formulate your question?" is also visible. At the bottom of the page, there are logos for "UNIVERSITÉ" and "FNSNF". The browser's taskbar at the bottom shows various system icons and the Windows taskbar.

# Pre-processing

# Indexing unit

- ⌘ An arbitrary representation of the content of the document as stored in an index or inverted file
  - ☑ In books: keywords, authors...

# Indexing unit

☒ Word → *bag of words*

Hypothesis: words are independent !

e.g. *peas*

☒ Le stem

e.g. *pea*

☒ N-grams  $1 < n < 5$

e.g. *green peas*

☒ Keywords

e.g. *banana split*

→ Compensate for the independence hypothesis

# Stemming (English)

Token	Type of Normalization			
	Lovins	Porter	Plural Stemming	Lemmatization*
genetic	genet	genet	genetic	genetic
genetically	genet	genetically	genetically	genetically
genetics	genet	genet	genetic	genetics
gene	gene	gene	gene	gene
genes	gene	gene	gene	gene
homogeneous	gene	homogen	homogeneous	homogeneous
plaid	plai	plai	plaid	play
play	plai	plai	play	play

**\*Lemmatisation: normalisation based on syntactic information  
(verb, noun, adjective, adverb...)**

# Stemming français (J. Savoy)

**On parvient néanmoins à libérer la vésicule du lit vésiculaire de manière rétrograde et à individualiser le canal cystique . Mise en place d'une ligature au niveau proximal du canal cystique . Section partielle du canal cystique . Il est impossible de cathétériser le canal cystique . On décide de ne pas continuer les manoeuvres vu le risque de plaie au niveau de la voie biliaire et , d'autre part , la cholangiographie rétrograde préopératoire , qui s'est avérée normale . Ligature au niveau du cystique.**

# Result: 80 → 51 → 24 formes !

biliair  
canal  
cholangiograph  
continu  
cystiqu  
liber  
Ligatur  
lit  
manoeuvr  
mis  
niveau  
normal

part  
partiel  
plac  
plai  
preoperatoire  
proximal  
retrograd  
risqu  
section  
vesicul  
voi

Ratio = 1/3 !



# Stemming

<http://albator.hesge.ch:8984/solr/#/collection1/analysis>

# Select stemmer, e.g. "Deutsch"



- Dashboard
- Logging
- Core Admin
- Java Properties
- Thread Dump
- collection1
- Overview
- Analysis**
- Dataimport
- Documents
- Files
- Ping
- Plugins / Stats
- Query
- Replication
- Schema Browser

Field Value (Index)  
 Vielen Lösungen

Field Value (Query)  
 Vielen Lösungen

Analyse Fieldname / FieldType:  ⓘ

Verbose Output

Analyse Values

ST	text	Vielen	Lösungen
	raw_bytes	[56 69 65 6c 65 6e]	[4c c3 b6 73 75 6e 67 65 6e]
	start	0	7
	end	6	15
	positionLength	1	1
	type	<ALPHANUM>	<ALPHANUM>
	position	1	2
LCF	text	vielen	lösungen
	raw_bytes	[76 69 65 6c 65 6e]	[6c c3 b6 73 75 6e 67 65 6e]
	start	0	7
	end	6	15
	positionLength	1	1
	type	<ALPHANUM>	<ALPHANUM>
	position	1	2
SF	text	vielen	lösungen
	raw_bytes	[76 69 65 6c 65 6e]	[6c c3 b6 73 75 6e 67 65 6e]
	start	0	7
	end	6	15
	positionLength	1	1
	type	<ALPHANUM>	<ALPHANUM>

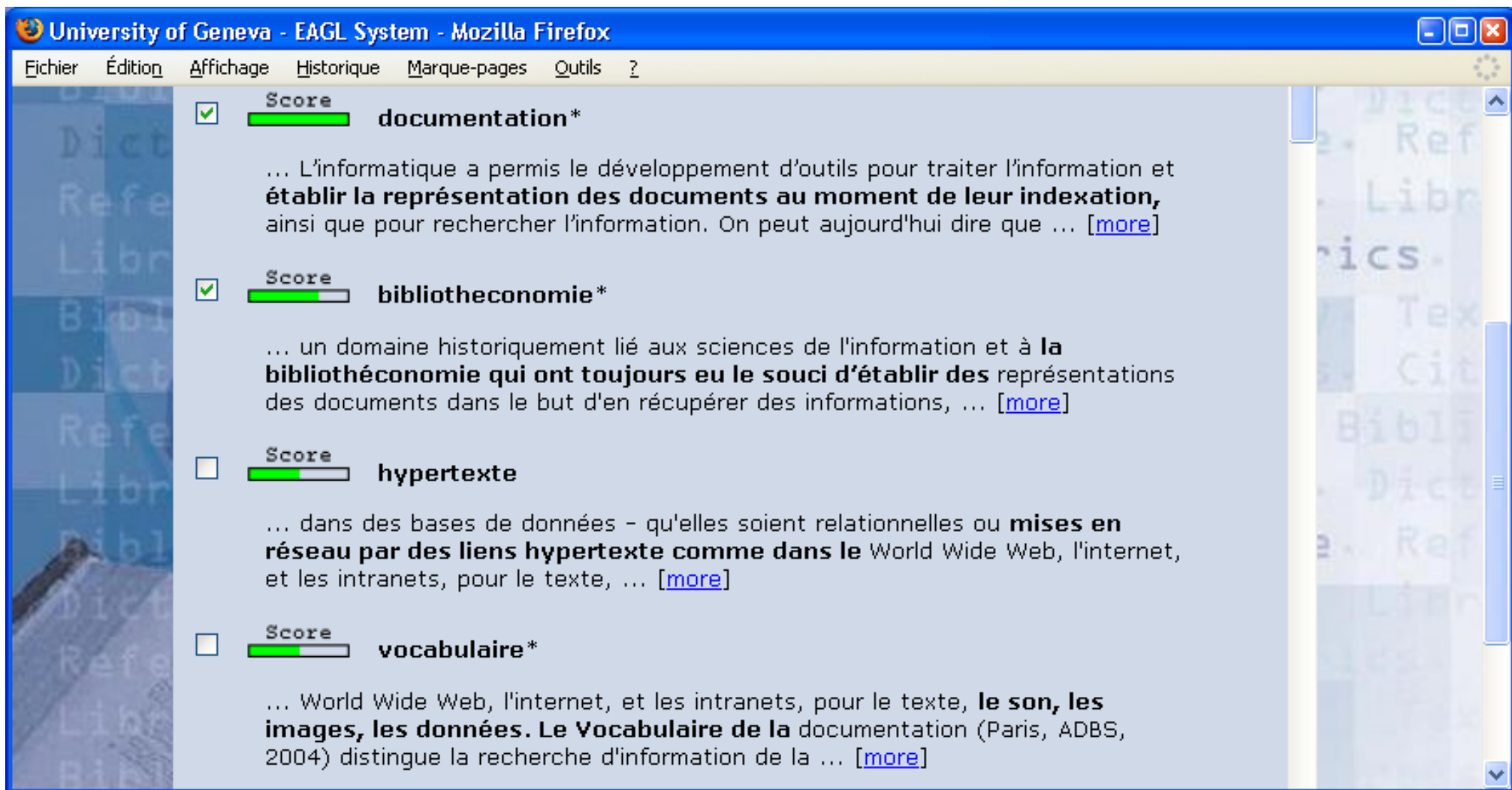
ST	text	Vielen	Lösungen
	raw_bytes	[56 69 65 6c 65 6e]	[4c c3 b6 73 75 6e 67 65 6e]
	start	0	7
	end	6	15
	positionLength	1	1
	type	<ALPHANUM>	<ALPHANUM>
	position	1	2
LCF	text	vielen	lösungen
	raw_bytes	[76 69 65 6c 65 6e]	[6c c3 b6 73 75 6e 67 65 6e]
	start	0	7
	end	6	15
	positionLength	1	1
	type	<ALPHANUM>	<ALPHANUM>
	position	1	2
SF	text	vielen	lösungen
	raw_bytes	[76 69 65 6c 65 6e]	[6c c3 b6 73 75 6e 67 65 6e]
	start	0	7
	end	6	15
	positionLength	1	1
	type	<ALPHANUM>	<ALPHANUM>

# Binary classification

- ⌘ Twitter Sentiment <http://www.sentiment140.com/>
- ⌘ Often supervised (need annotated data)

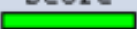



# ATC with no supervision but with lexical similarities

Source: [http://fr.wikipedia.org/wiki/Recherche\\_d%27information](http://fr.wikipedia.org/wiki/Recherche_d%27information)



The screenshot shows a Mozilla Firefox browser window titled "University of Geneva - EAGL System - Mozilla Firefox". The browser's menu bar includes "Fichier", "Édition", "Affichage", "Historique", "Marque-pages", and "Outils". The main content area displays search results for the term "Recherche d'information".

The search results are listed as follows:

- Score**  **documentation\***  
... L'informatique a permis le développement d'outils pour traiter l'information et **établir la représentation des documents au moment de leur indexation**, ainsi que pour rechercher l'information. On peut aujourd'hui dire que ... [\[more\]](#)
- Score**  **bibliotheconomie\***  
... un domaine historiquement lié aux sciences de l'information et à **la bibliothéconomie qui ont toujours eu le souci d'établir des** représentations des documents dans le but d'en récupérer des informations, ... [\[more\]](#)
- Score**  **hypertexte**  
... dans des bases de données - qu'elles soient relationnelles ou **mises en réseau par des liens hypertexte comme dans le** World Wide Web, l'internet, et les intranets, pour le texte, ... [\[more\]](#)
- Score**  **vocabulaire\***  
... World Wide Web, l'internet, et les intranets, pour le texte, **le son, les images, les données. Le Vocabulaire de la** documentation (Paris, ADBS, 2004) distingue la recherche d'information de la ... [\[more\]](#)

# Hierarchical classification

⌘ Using  $n$  to classify  $n+1$  is ineffective

⌘ Example: TWINC, Yahoo <http://dir.yahoo.com/>

## Organisms

- NT1 Eukaryotes
  - NT2 Animals
    - NT3 Aquatic animals
      - NT4 Aquatic mammals
        - NT5 Marine mammals
          - NT6 Whales
            - NT7 Baleen whales
              - NT8 Blue whale
              - NT8 Gray whale
              - NT8 Humpback whale
              - NT8 Minke whale
              - NT8 Right whales
                - NT9 Bowhead whale
                - NT9 Northern right whale
                - NT9 Southern right whale
            - NT7 Fossil whales
            - NT7 Toothed whales
              - NT8 Beaked whales
              - NT8 Beluga (Whale)
              - NT8 Dolphins (Mammals)
                - NT9 Bottlenosed dolphins
                - NT9 Killer whales
                - NT9 Pilot whales
                - NT9 River dolphins
              - NT8 Narwhal
              - NT8 Porpoises
              - NT8 Sperm whale

# ATC evaluation

## ⌘ Precision

Ratio between pos categ and total generated categ

## ⌘ Recall

Ratio between pos categ and expected categories

⌘ Law:  $F(\text{Précision}) = 1/F(\text{Rappel})$

⌘ Combination:  $F_{\text{score}}$

# Exercise

⌘ Recall and precision ?

Expected categories

Generated categories

Diabetes

Female

1. Human

2. Female

3. Type II Diabetes

4. New born babies

# Solution

⌘ Precision = 0.25 (1/4) ; Recall = 0.5 (1/2)

Expected	Generated
Diabetes Female	1. Human 2. Female 3. Type II Diabetes 4. New born babies



# Feed-back

⌘ Twinc: <http://casimir.hesge.ch/ChemAthlon/index.html#>

⌘ Similar to query expansion

☒ Normalization (stemming)

☒ Other methods: thesaurus-driven

☒ Tend to improve recall and degrade precision !

→ To be evaluated

# Example

## ⌘ Stemming/thesaurus:

☑ +Recall:

*vésicules, vésicule, vésiculaire → vésicul*

☑ -Precision

foi (faith), foie (liver) → foi

# Limitations of benchmarks

⌘ Information retrieval: OK

⌘ Categorisation:

☑ How reliable is the source annotation ?

# Benchmarks for categorization

## ⌘ Methods

- ☒ Annotations = *ground truth, gold standard...*
- ☒ Agreement between curators (librarians, biologists...)
- Precision = 100% is impossible...
- Baseline = Inter-rater agreement (Kappa)  
MeSH Major, 22'000 descriptors: ~ 40% !  
GO ~ 40-50%...

# Conclusion

1. IR/TM = wide set of different tasks
2. Each task require specific metrics
3. Each task require specific instruments → multiply applications instead of one size fit all !
4. Each development requires labor-intensive evaluation: 80/20