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

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_L3kMWenDxQfYGvwJJM9QEe0/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, octobre 2013

DBMS

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

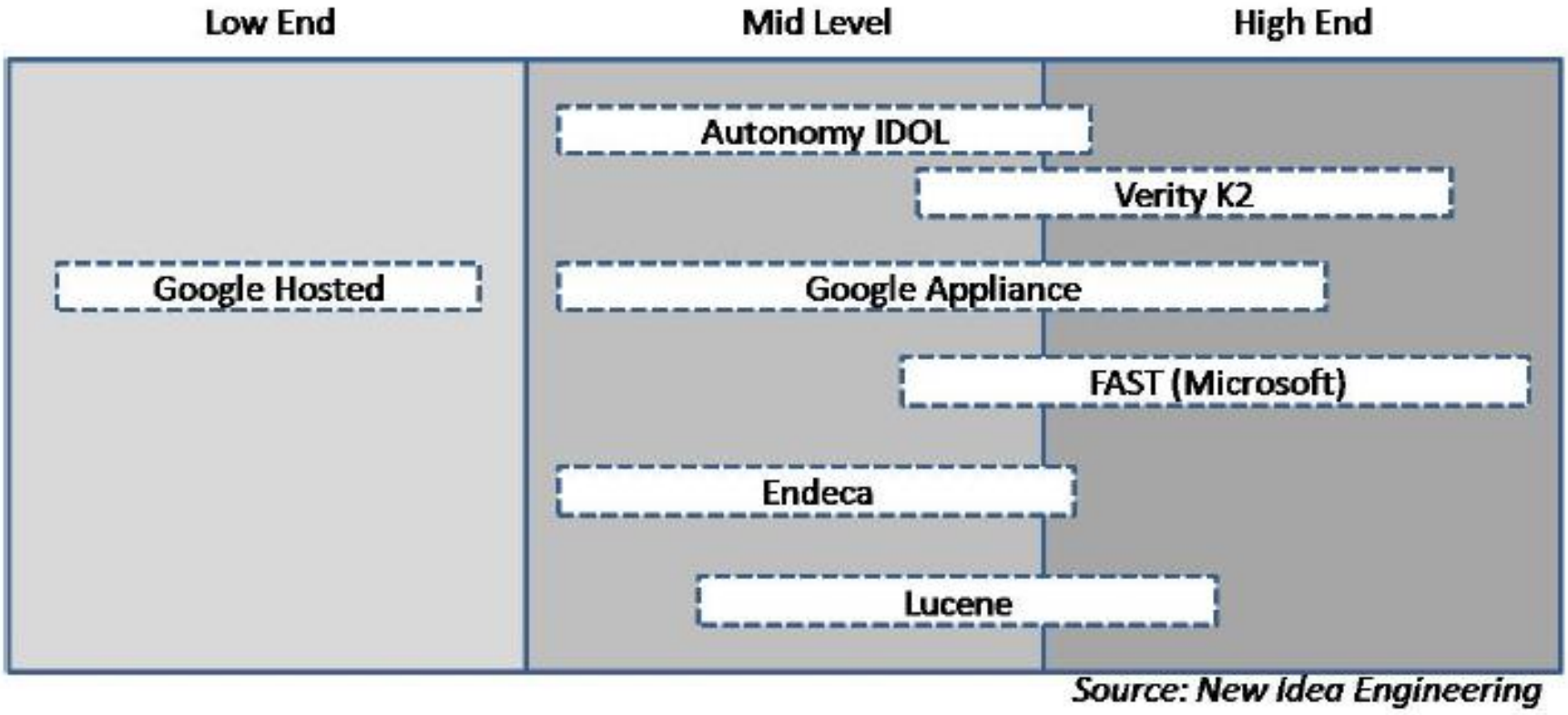
Web search

Search engine	Market share in May 2011	Market share in December 2010[14]
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

Oct 2013

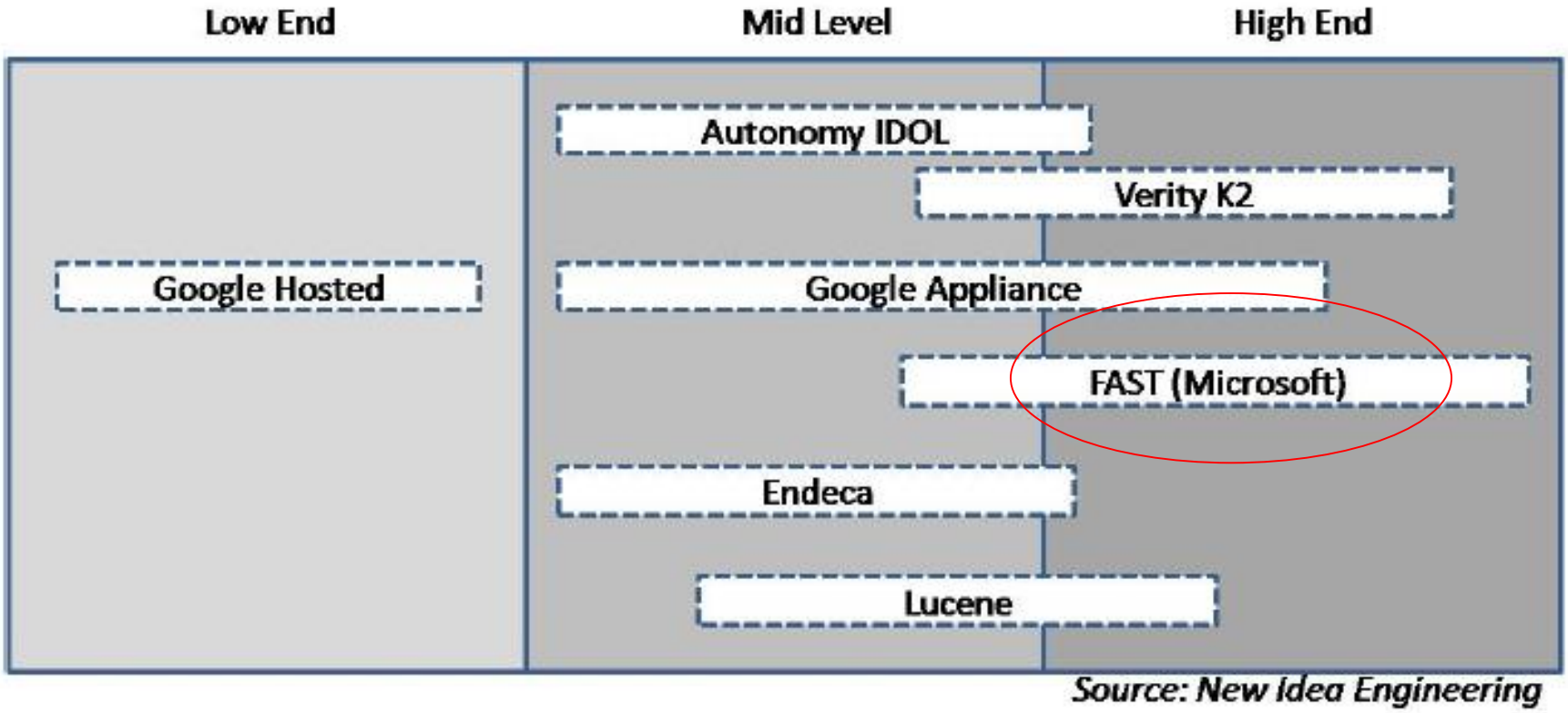
Non-web search



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

Oct 2013

Bullshit…



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 ☺]

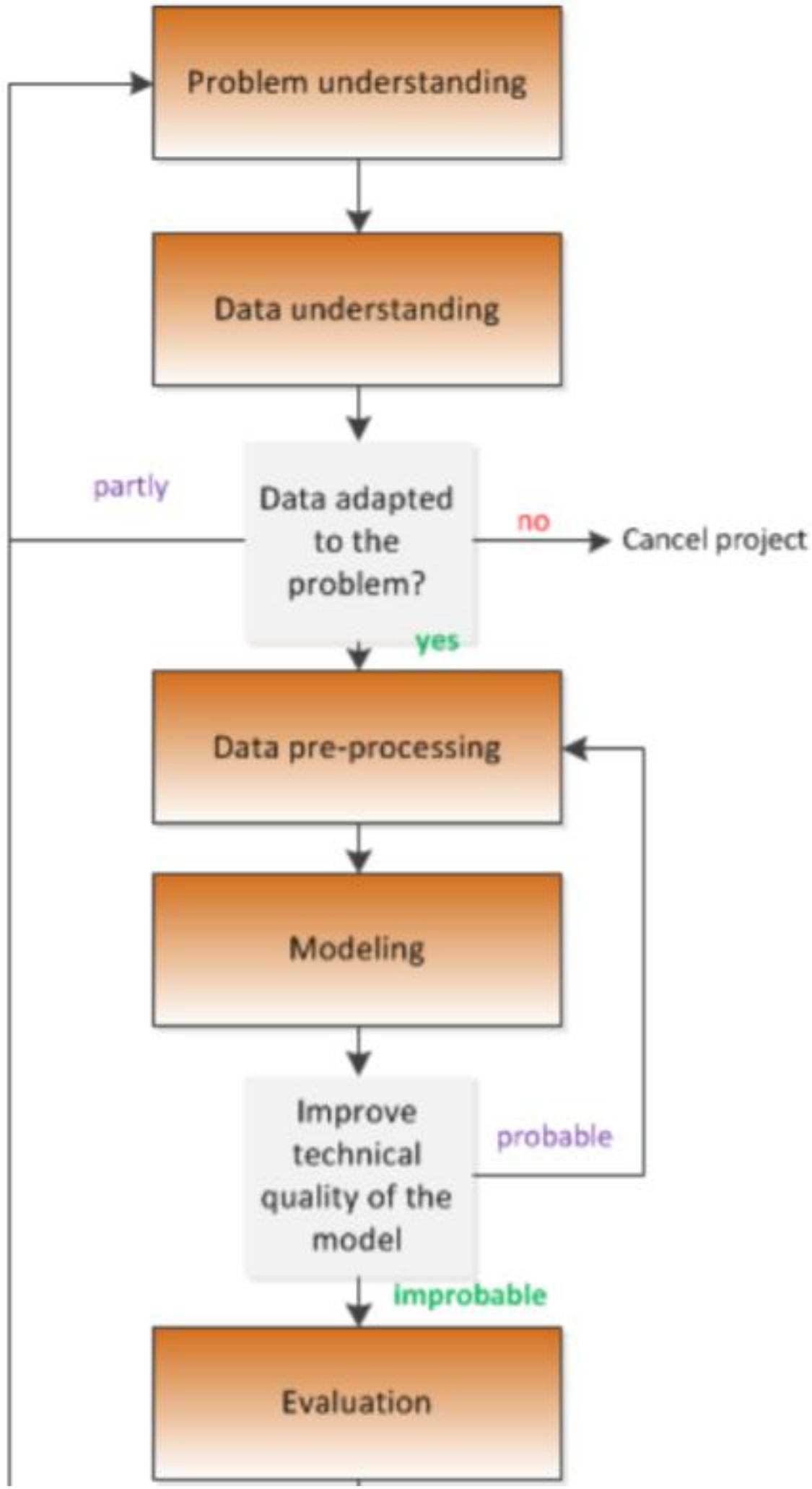
- Entreprise search
  - Google box
  - Autonomy (HP)
  - Fast (MS)
  - they include EDM + CMS…, e.g. Sharepoint

Professional – yet – open source tools

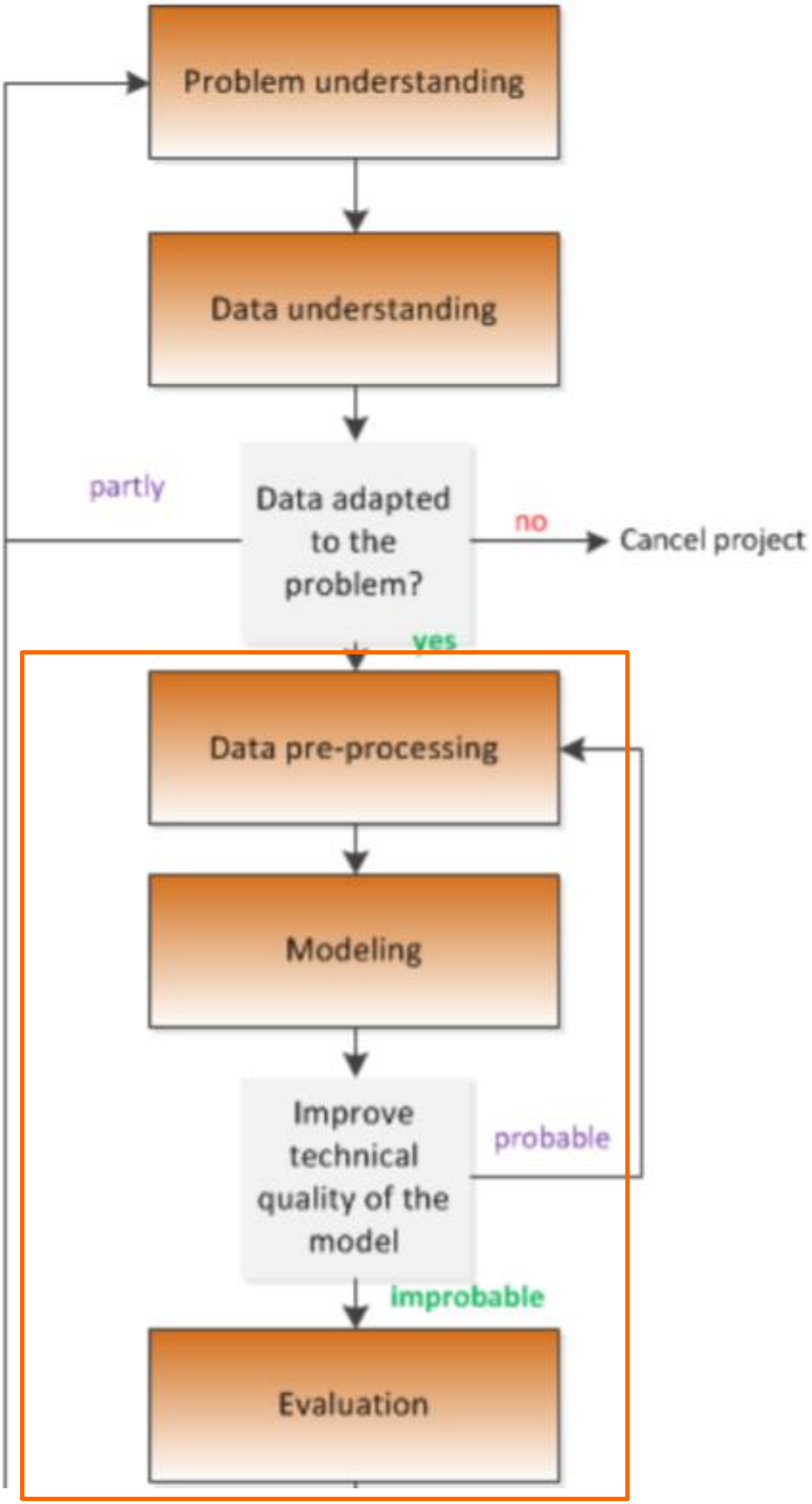
- Lucene ecosystems (Hadoop/HDFS, Lucidwork…)
  - SOLR, ElasticSearch…
- Other (with better ranking functions)
  - Lemur
  - Terrier
  - Indri
  - MG4J…
  - Weka, Clementine… ok for toy collections [Data mining platform]

Other players

- Search engine/Analytics + **Contents**
  - Business
    - Thomson Reuters
    - LexisNexis
    - Bloomberg
  - Sciences
    - Elsevier
    - Springer


Problem understanding
Data understanding
partly
Data adapted to the problem?
no
Cancel project
yes
Data pre-processing
Modeling
Improve technical quality of the model
probable
improbable
Evaluation

Text mining


Problem understanding
Data understanding
Data adapted to the problem?
partly
no
Cancel project
yes
Data pre-processing
Modeling
Improve technical quality of the model
probable
improbable
Evaluation

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 adress text mining challenges !

Distances

- Euclid and non Euclid distances

http://fr.wikipedia.org/wiki/Distance_%28math%C3%A9matiques%29

Similarity

= 1/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/

Training/tuning data

- What is needed to learn or tune a task

- Text mining → Training
- Information retrieval → Tuning

Example

- Annotated or automatically annotated

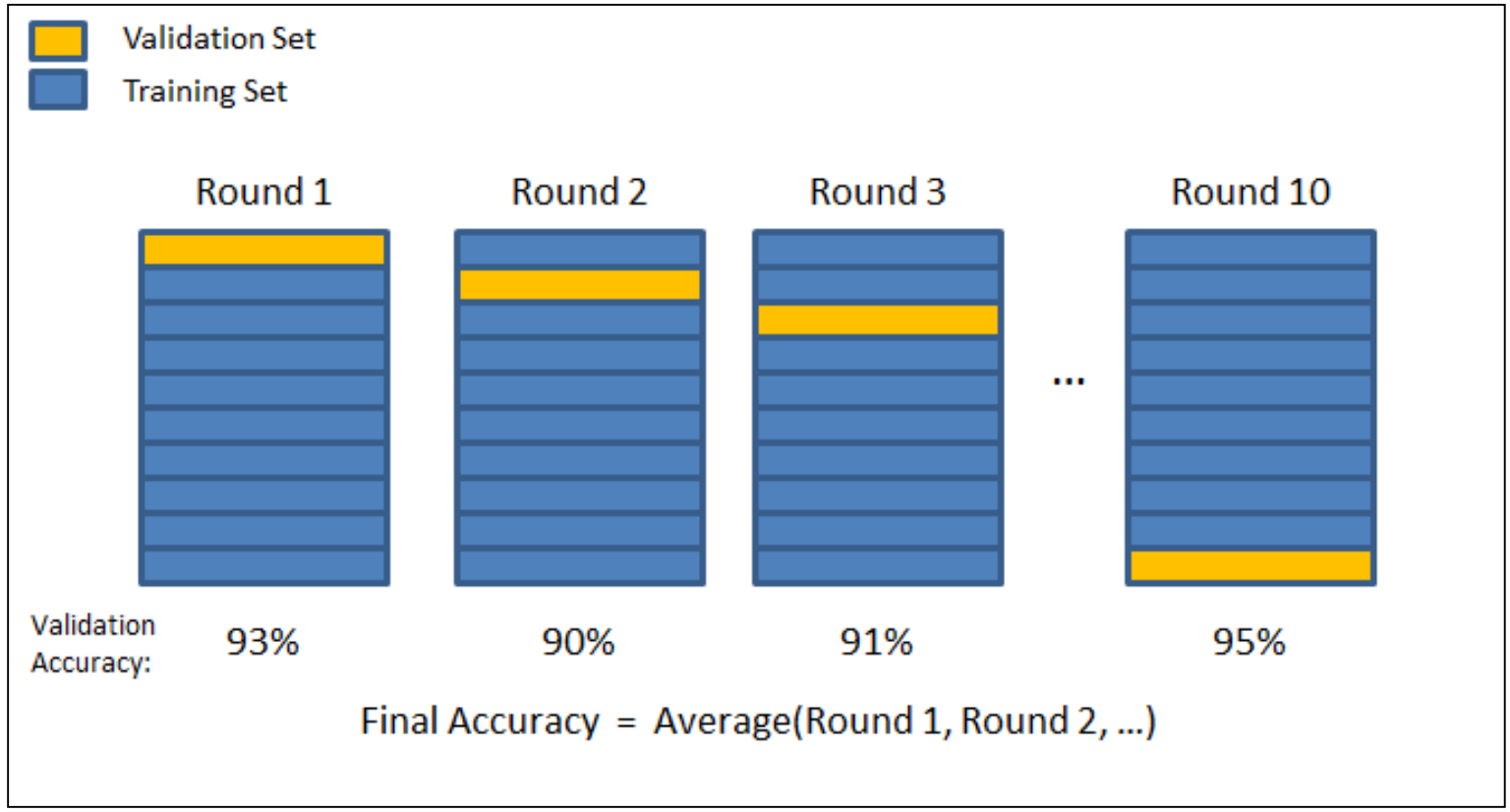
  Sentiment…

  e.g. ☺ → Positif

  ☹ → Negatif

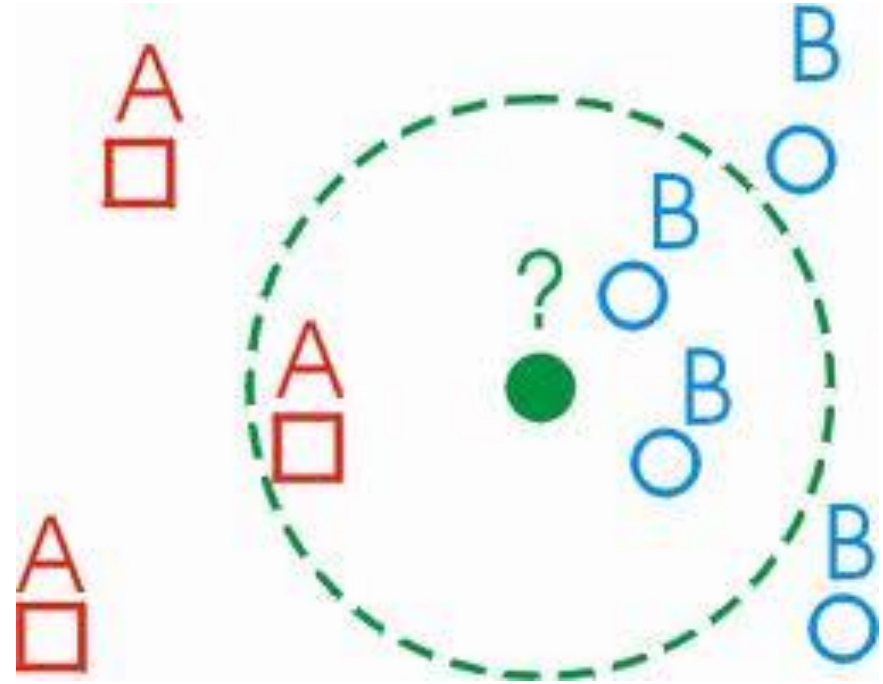
Pipeline

- Training set
- Test set
- Validation/**Cross-validation**/Leave-one out



Different classifiers

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



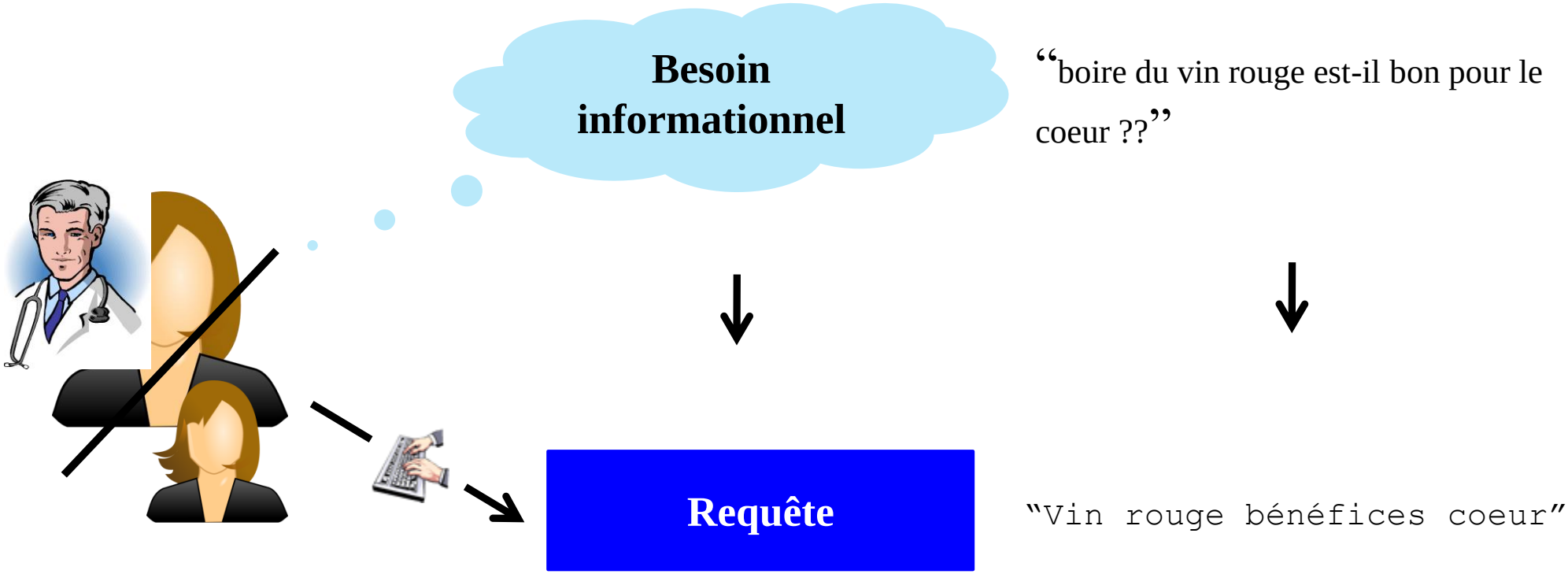
Mainly two tasks

- Information retrieval → Cranfield paradigm !
  [Metrics: AM]

- Classification
  - binary
  - multiclass

Pertinence

- Dictionary : « what is appropriate » !
- Information needs :



What is the most pertinent ?

FemmeActuelle.fr

Les bienfaits du vin
par Claire Frayssinet,

Consommé avec modération, le vin est un véritable élixir de jouvence. 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 confirmé 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

.....

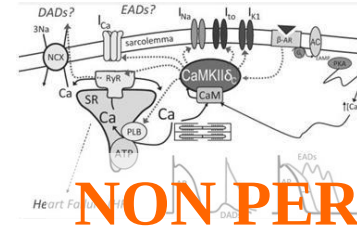
.....

.....

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 components of wine; 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

.....

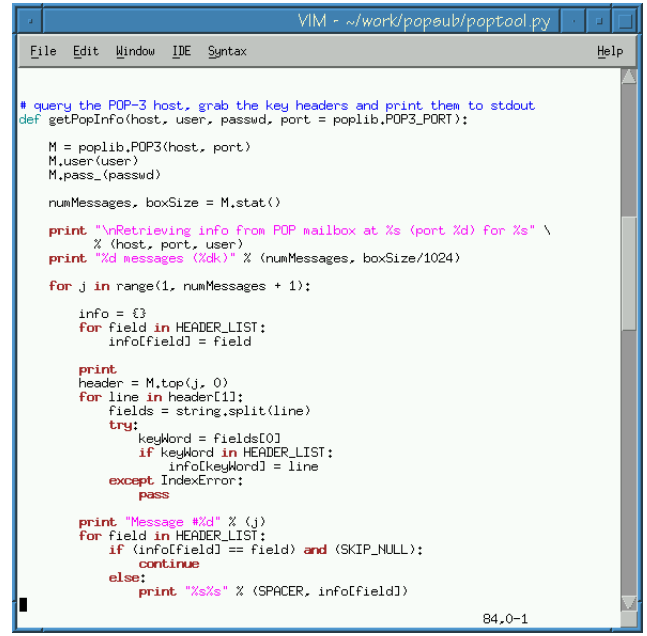
.....

.....

?

?

← “python” →



→ 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.guide**auto**.com/nouvelles/actualites/2531**auto**45/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&cpsidt=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 Francais 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 • Translate

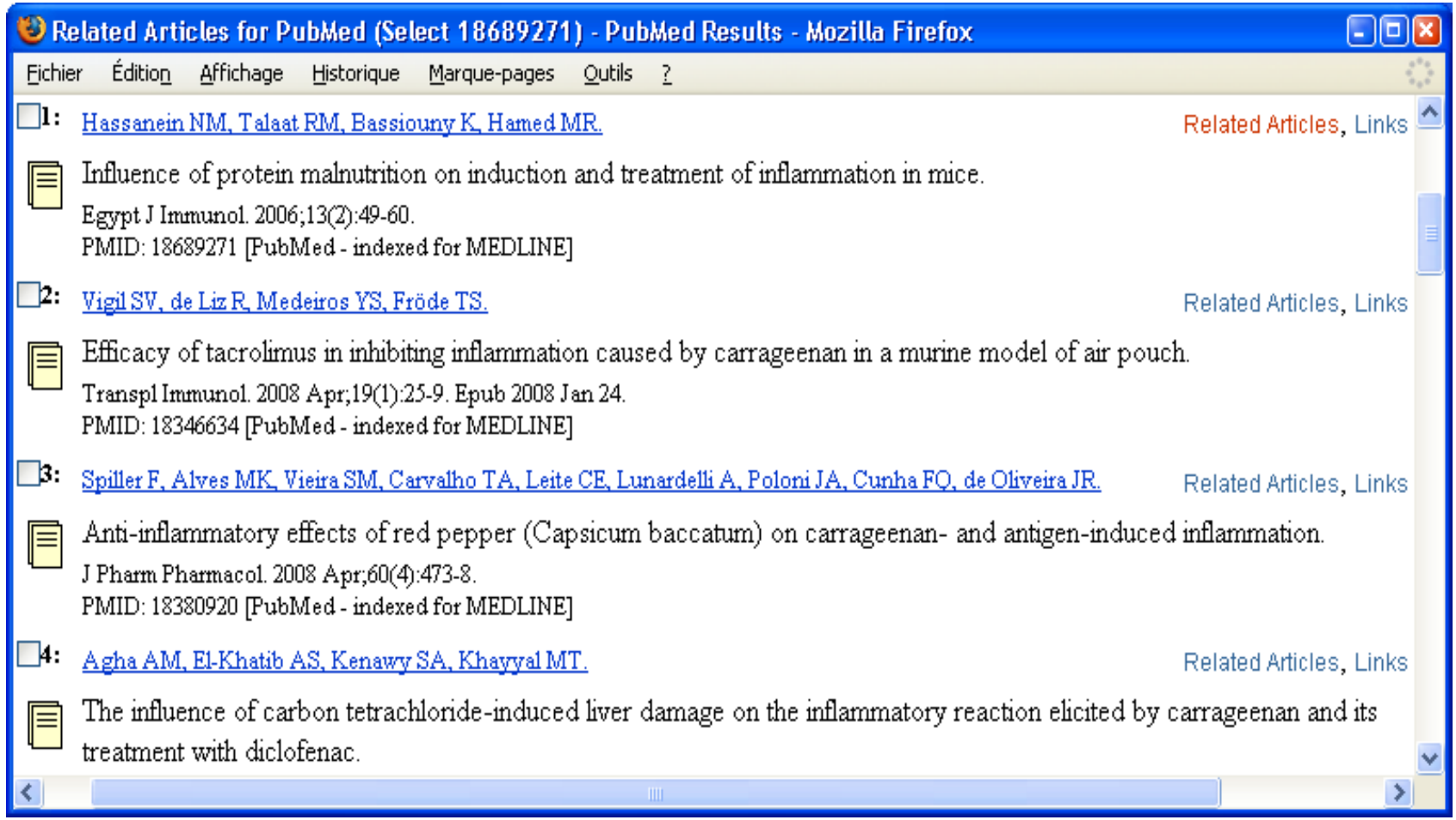
→ Ratio = 60% → 92%

Image search

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

	$image_{Query}$	$image_1$	$image_2$	$image_3$
R	100	25	99	100
V	1	200	0	1
B	50	67	49	40

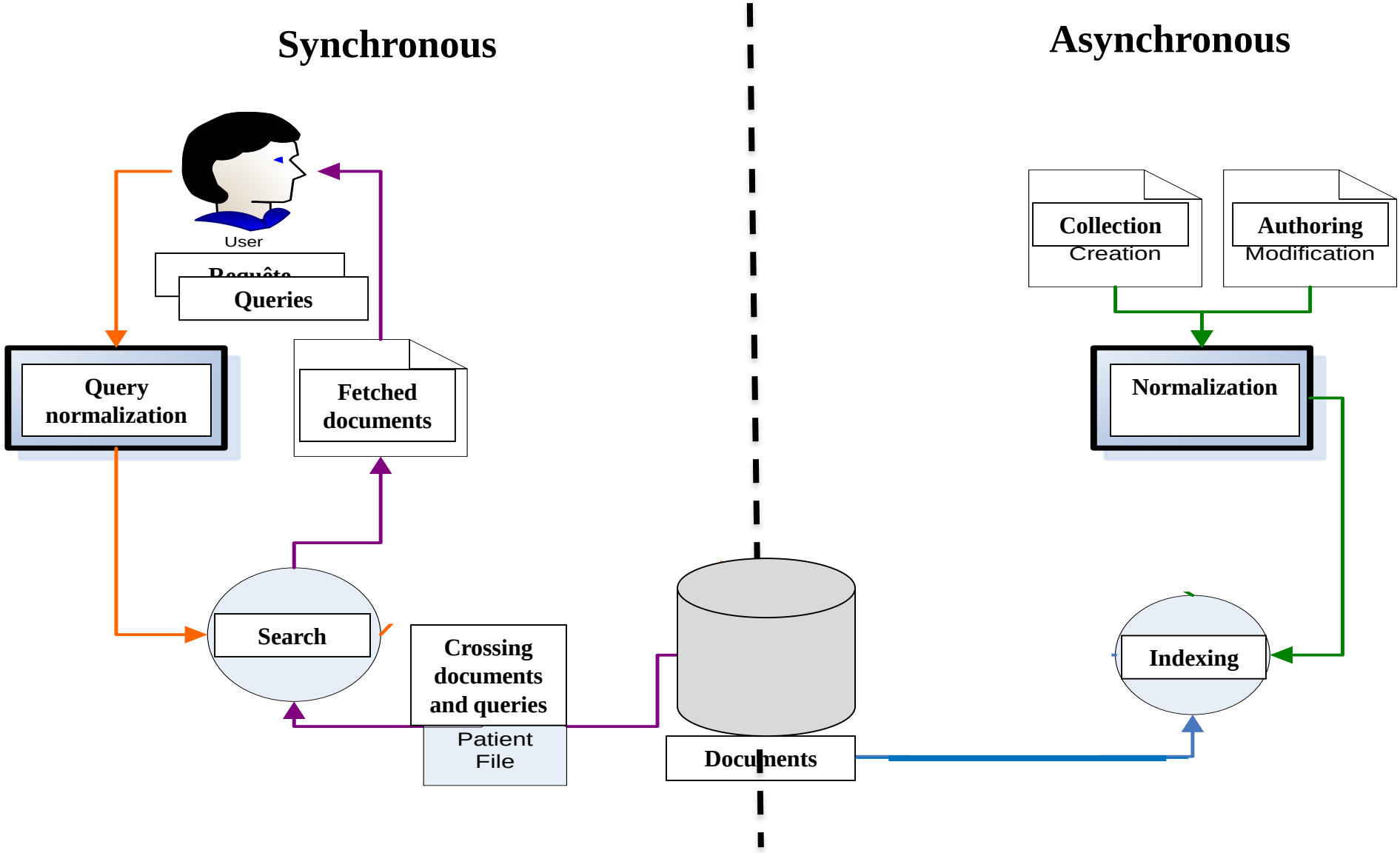
Related article search



Medline, patent libraries…

KWIC key-word in context / Passage Retrieval

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
    - Buidling of statistical model

Models of RI

- Boolean
  - x AND y AND z; x OR y OR z
- Vector-space (probabilistic, statistical…)
  - X Y Z ~ X OR Y OR Z + 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 discriminent 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	1	1	0	0	0	1	
Brutus	1	1	0	1	0	0	
Caesar	1	1	0	1	1	1	
Calpurnia	0	1	0	0	0	0	
Cleopatra	1	0	0	0	0	0	
mercy	1	0	1	1	1	1	
room	1	1	1	1	1	1	
…							

	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	6	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	
room	21	29	32	25	14	33	
…							

→ **TF (local to document)**

→ IDF (collection-wide)

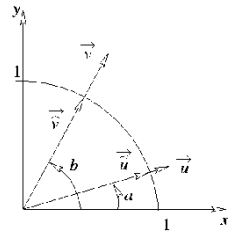
Weighting schema

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 :

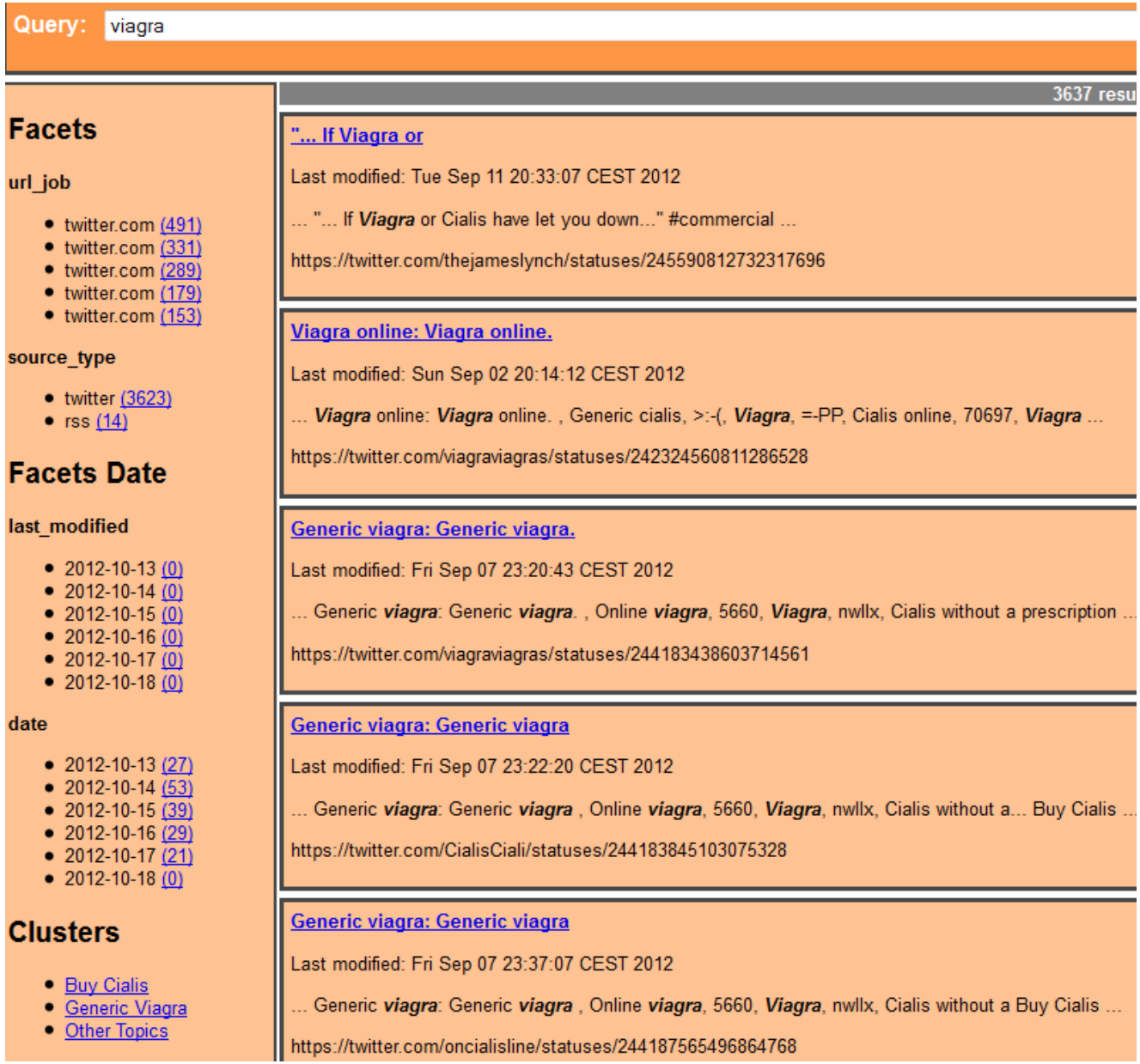


Weighting schema example, Okapi BM25: tf, df, dl

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

$$\text{where } K = k_1 \cdot \left[(1 - b) + b \cdot (l_i / avdl)\right]$$

Clustering/faceting



Information extraction with "is a", "is an"...

Example: create wikidictionary !

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

EAGLictionary MeSH :

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/

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)

	Type of Normalization			
Token	**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 neanmoins a liberer la vesicule du lit vesiculaire de maniere retrograde et a 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 catheteriser le canal cystique . On decide de ne pas continuer les manoeuvres vu le risque de plaie au niveau de la voie biliaire et , d'autre part , la cholangiographie retrograde preoperatoire , qui s'est averee normale . Ligature au niveau du cystique.

Result: 80 → 51 → 24 formes !

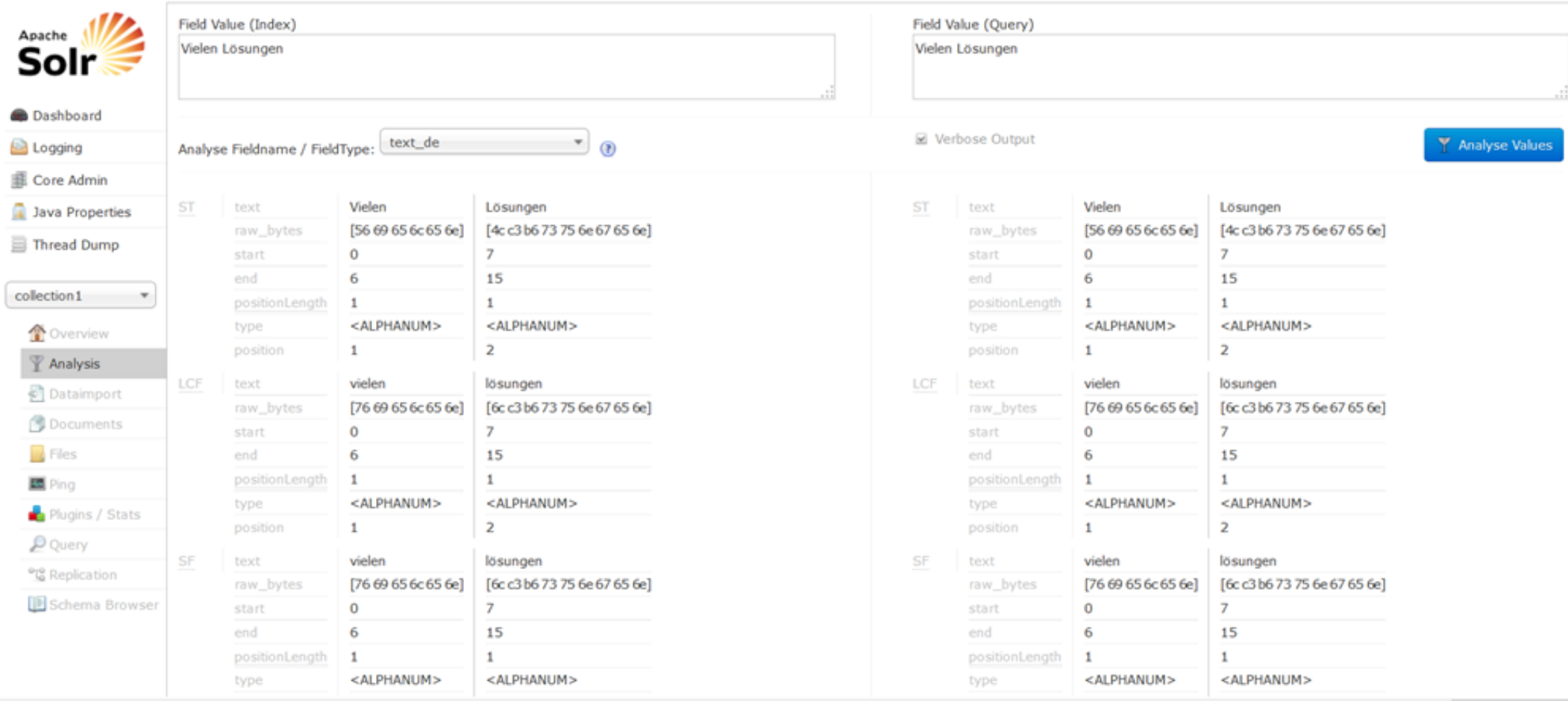
biliair
canal
cholangiograph
continu
cystiqu
liber
Ligatur
lit
manoeuvr
mis
niveau
normal
part
partiel
plac
plai
preoperatoir
proximal
retrograd
risqu
section
vesicul
voi

Ratio = 1/3 !

Stemming

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

Select stemmer, e.g. “Deutsch”

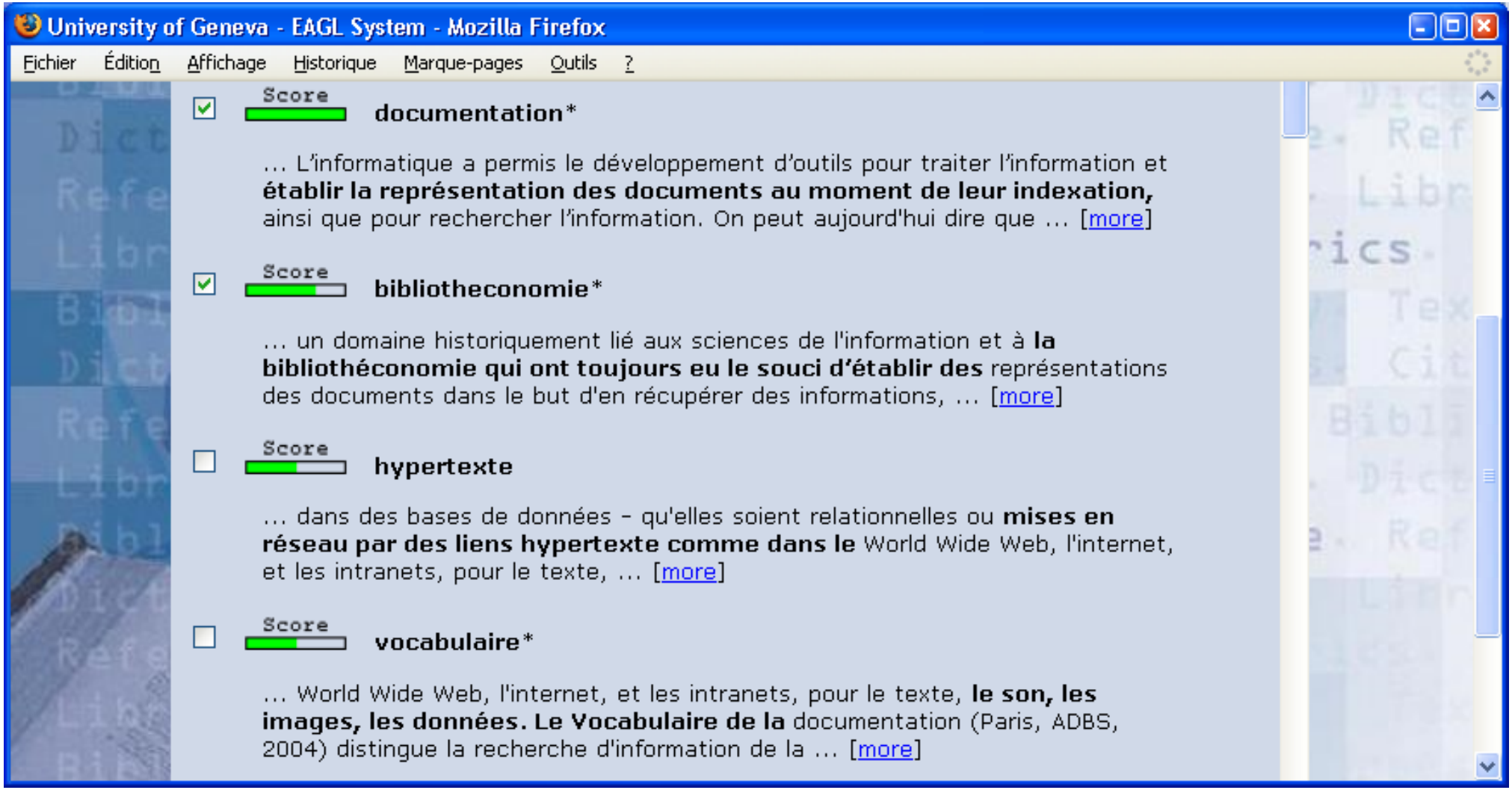


Binary classification

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

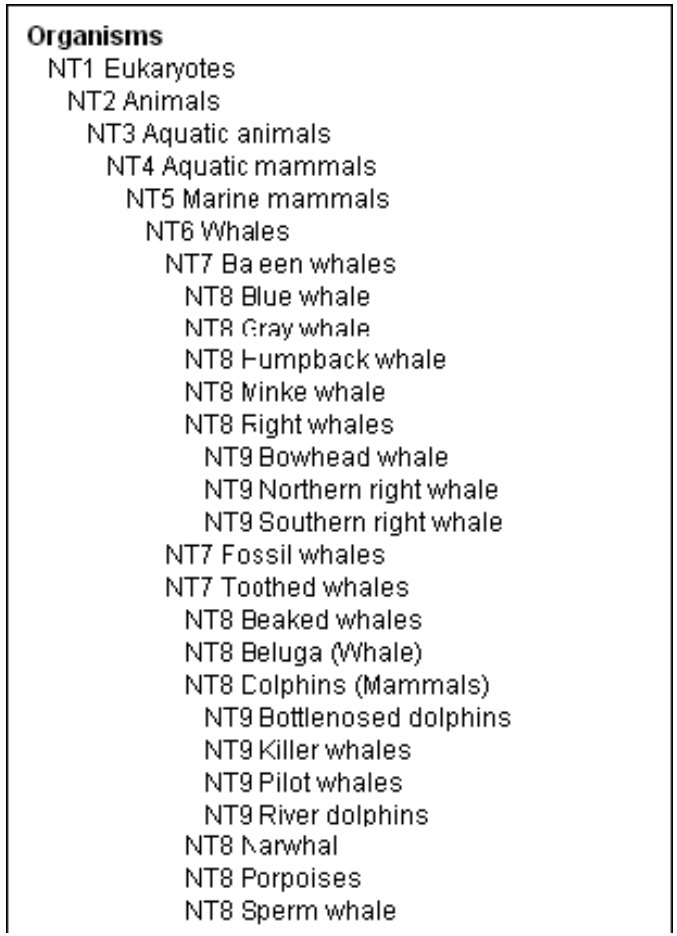
ATC with no supervision but with lexical similarities

Source: http://fr.wikipedia.org/wiki/Recherche_d%27information



Hierarchical classification

- Using n to classify n+1 is ineffective
- Example: TWINC, Yahoo http://dir.yahoo.com/



ATC evaluation

- Precision

  Ratio between pos categ and total generated categ
- Recall

  Ratio between pos categ and expected categories
- Law: F(Précision) = 1/F(Rappel)
- Combination: F_{score}

Exercice

- 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