

Developers' view

SAT implementation

Artur Šilić and Frane Šarić

Knowledge Technologies Laboratory

Faculty of Electrical Engineering and Computing

University of Zagreb

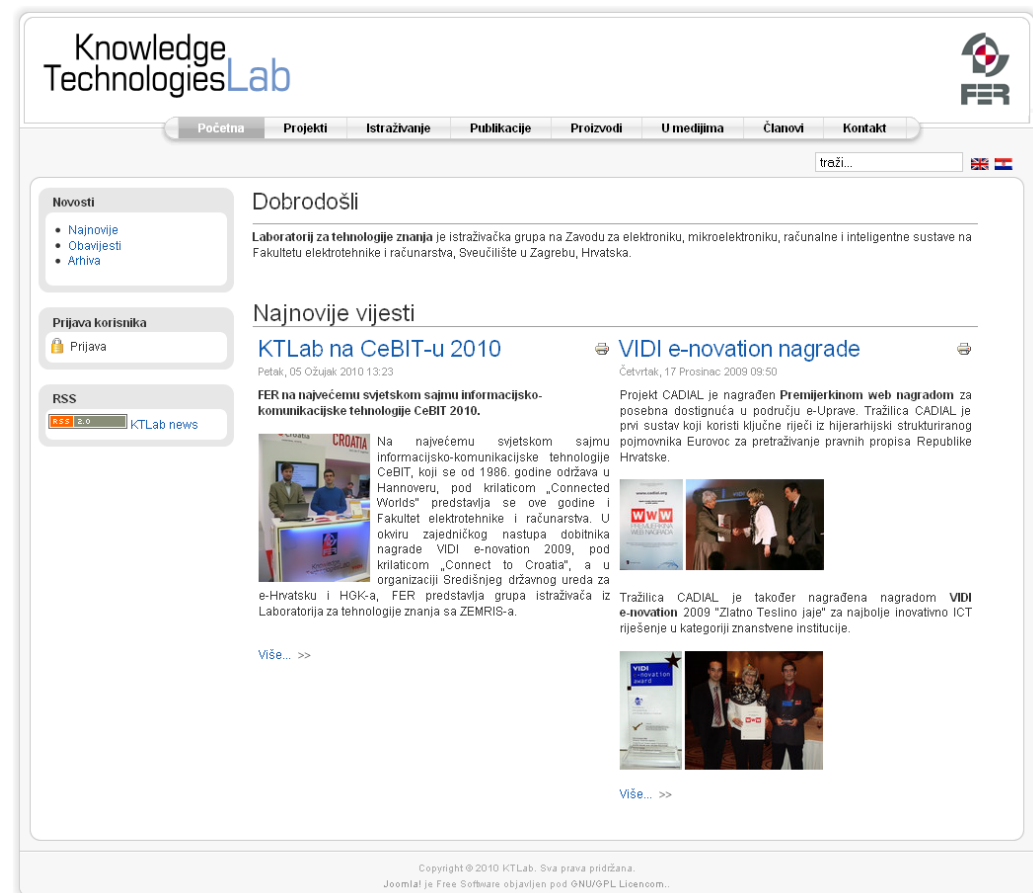


Knowledge
TechnologiesLab

About us

- Prof. Bojana Dalbelo Bašić
- Dr. Jan Šnajder
- Frane Šarić
- Artur Šilić
- Goran Glavaš
- Mladen Karan

- <http://ktlab.fer.hr/>

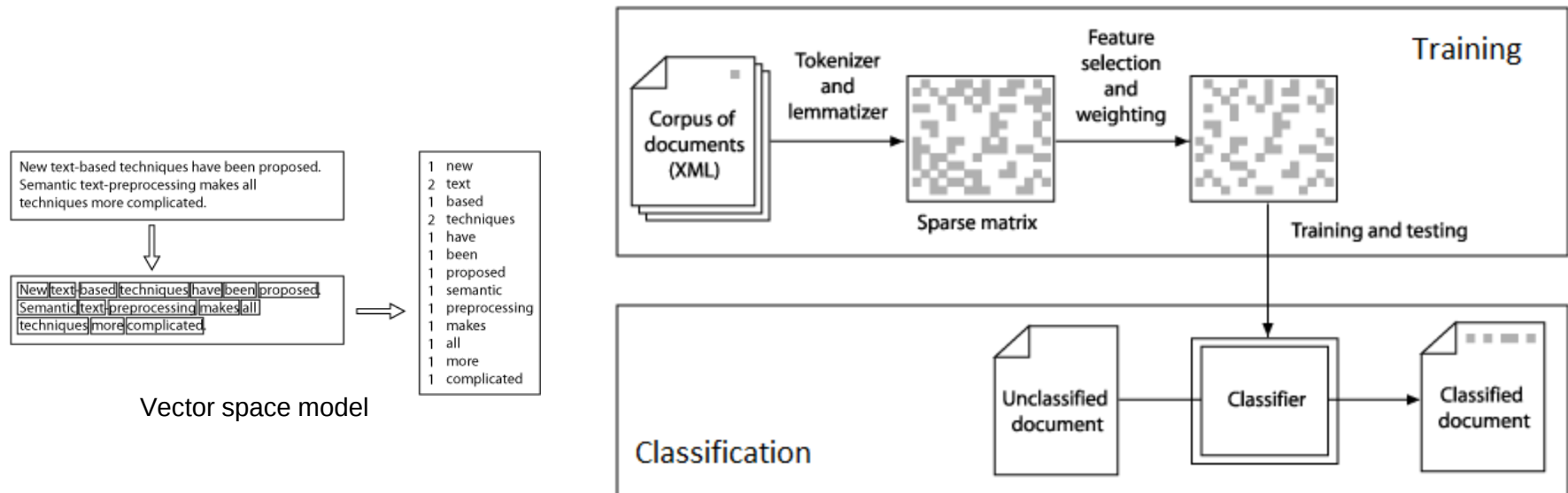


Classifier and trainer modules (KTN)

- Mini-outline:
 - Functionality
 - Architecture
 - Background layer
 - Client layer
 - Technologies
 - Implementation examples

Classifier and trainer modules

- Functionality
 - automatic and semi-automatic text classification
 - training classifiers for new categories
 - retraining classifiers for existing categories
 - system administration



Classification algorithm

- Linear logistic regression
 - Popular, very fast
 - LIBLINEAR (<http://www.csie.ntu.edu.tw/~cjlin/liblinear>), BSD license
 - Results similar to SVMs on linear text classification problems
 - L2 SVM loss function
 - One parameter – C
 - an easy optimization task

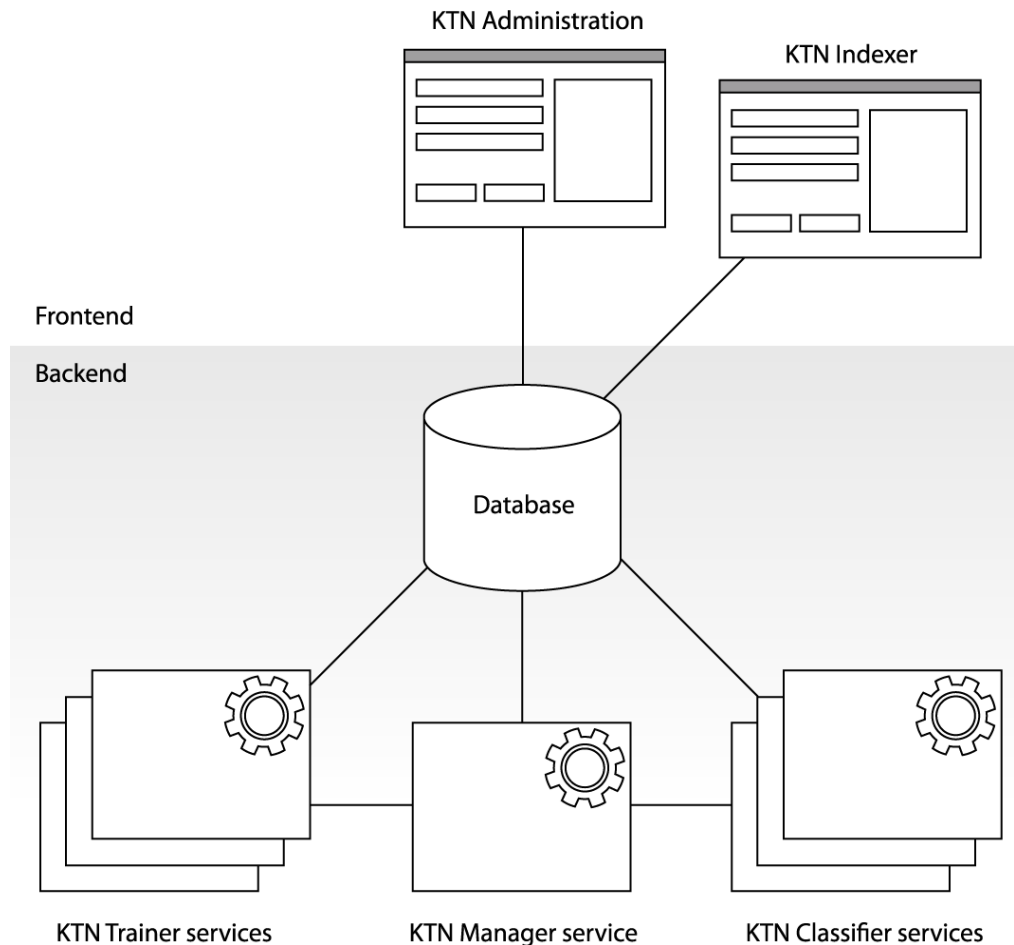
$$P(y = \pm 1 | \mathbf{x}, \mathbf{w}) = \frac{1}{1 + \exp(-y(\mathbf{w}^T \mathbf{x} + b))}$$

$$\min_{\mathbf{w}} f(\mathbf{w}) \equiv \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^l \log(1 + e^{-y_i \mathbf{w}^T \mathbf{x}_i})$$

Classifier and trainer modules (2)

■ Architecture

KTN Indexing System



Classifier and trainer modules (3)

- Background layer
 - Windows services that communicate via MS Message Queues
 - Trainer service trains new classifiers based on available data in database
 - Classifier service automatically classifies documents after their import into the system
 - The manager service coordinates both classifier and trainer services by giving necessary commands, also imports, preprocesses, exports documents, maintains a full-text index and answers search queries

Classifier and trainer modules (4)

- Client layer
 - KTN Administration provides the capability of collection and category maintenance, setting new train tasks for the trainer, commanding index building and collection export, as well as general system variables and user accounts

The screenshot displays the 'KTN Administration' window with the 'Collections' tab selected. The interface is divided into several sections:

- Clean collections (for training):** A table listing collections 7 through 11. Each row shows details like Name, Articles, Date created, Finished status, and sizes of manual and auto labels.
- Production collections:** A table listing collections 12 and 13 with similar details.
- Buttons:** Below the tables are buttons for 'Delete', 'Delete manual labels', 'Delete auto labels', 'Assign auto labels', and 'Export'.
- Collection info:** A section for 'Collection0008' with a 'Clean' checkbox and a 'Save' button.
- Database operations:** Buttons for 'Build indexes', 'Reset database', 'Finish all collections', 'Rebuild indexes', and 'Clear database'.
- Categories (for collection 'Collection0008'):** A panel on the right with radio buttons for 'Existence', 'Manual', and 'Auto'. It also includes a 'Show categories' button and a list of 'Sorted categories' with a checkbox for 'Unsorted categories'.

Classifier and trainer modules (5)

- In KTN Indexer the user can confirm automatic labels or spend an arbitrary amount of time to manually inspect classifications that are identified as uncertain by the system

Filter One category view

Showing category 'ZAPOŠLJAVANJE' from collection 'Collection0003'

Label distribution

Document text

za:1. odgojitelj - 9 djelatnika - rad na neodređeno vrijeme2. stručni suradnik defektolog - 1 djelatnik VSS - rad na 50% radnog vremena - neodređeno radno vrijemeUvjeti: Prema 61. 24. i 25. Zakona o predškolskom odgoju i naobrazbi (NN 10/97) i Pravilniku o vrsti stručne spreml stručnih djelatnika te vrsti i stupnju stručne spreml ostalih djelatnika u dječjem vrtiću (NN 133/97). Za radna mjesta odgojitelja mogu se prijaviti osobe oba spola koje ispunjavaju propisane uvjete, a prednost imaju djelatnici zatečeni na tim radnim mjestima.Prijave na natječaj podnose se na adresu Dječjeg vrtića u roku osam dana od dana objave natječaja u dnevnom tisku (s naznakom - prijava za natječaj za odgojitelja) Za navedena radna mjesta prijavi treba priložiti dokaz o stručnoj spremi, stručnom ispitu, radnom stažu, domovnici i životopis. O ishodu natječaja Dječji vrtić će izvijestiti kandidate u roku osam dana od dana donošenja odluke.

Relative impacts

Feature	Impact
radni	2.521
natječaj	1.364
stručan	1.265
prijava	1.135
sprem	0.705
prijaviti	0.693
neodređen	0.603
životopis	0.393
spremiti	0.363
vss	0.257
rok	0.220
rad	0.201
adresa	0.195
domovnica	0.180
ispit	0.178
uvjet	0.176
objava	0.171
kandidat	0.161
priložiti	0.145
spol	0.111
nn	0.110
dokaz	0.108
podnositi	0.108
mjesto	0.105
dan	0.103
ispunjavati	0.098
djelatnik	0.094
objaviti	0.092
suradnik	0.084
zakon	0.079
vrijemeuvjeti	0.063
prednost	0.056
vrijeme	0.043
osoba	0.040
stupanj	0.023
naveden	0.018
izvijestiti	-0.011
imati	-0.012
ostao	-0.011
moći	-0.012

Show documents

Labelled: 0 0 8

Unlabelled: 4959 3 30

Manual label	Auto output	Auto rule	Id
Yes	1.000	0	602168
Yes	1.000	0	601297
	1.000	0	604261
Yes	1.000	0	602338
	0.999	0	601280
Yes	0.998	0	601450
Yes	0.998	0	602170
	0.998	0	601376
No	0.998	0	
	0.997	0	
	0.997	0	
	0.997	0	
No	0.996	0	604206
	0.990	0	601279
Yes	0.989	0	602103
	0.989	0	601414
	0.989	0	601879
	0.987	0	601282

Label as... Positive

Label all with output less than x as... Negative

Label all with output greater than x as... Unlabeled

Label all as...

Save Cancel

Classifier and trainer modules (6)

- KTN Indexer enables defining new categories by methods of active learning
- The same procedure is used to efficiently clean existing collections from legacy databases

The screenshot displays the CLARA software interface for active learning. The main window is titled "Choose collections for category induction" and "Active learning a new category".

Left Panel: Document List

Search: zrakoplov
Found 46 documents.

Label	Snippet	Id
✓ Yes	...Narudžba se odnosi na verzije zrakoplova A350-800 i A350...	594096
✓ Yes	...incidentima koji uključuju isti tip zrakoplova", rekao je Danie...	594465
✓ Yes	...uprave tvrtke da prestane koristiti zrakoplove Dash Q400 ka...	594999
✓ Yes	...kvalitetom". Croatia Airlines kupuje četiri zrakoplova Dash 8 ...	595056
✓ Yes	...SAS je zaključio da ti zrakoplovi imaju probleme s kvaliteto...	594888
✓ Yes	...gdje god oni bili u zrakoplovu, da se pridržavaju standarda...	597258
✓ Yes	...SAS) odlučila izvan upotrebe staviti zrakoplove Dash Q400.	591462
✓ Yes	...koji su vodili ljubav u zrakoplovu dok leti na visini većoj...	596995
✓ Yes	...CL-415 spada među najučinkovitije zrakoplove za borbu pr...	594234
✓ Yes	...a vrlo vjerojatno i protupožarnih zrakoplova Cana-dair. doz...	592215
✓ Yes	...u jedinici za proizvodnju vojnih zrakoplova... ..je isporučila ...	594095
✓ Yes	...sezonu dok ne stignu novi zrakoplovi. Naime, Vlada je previ...	593889
✓ Yes	...koje to nisu. U svim zrakoplovima vrijede ista pravila.pr...	596818

Buttons: Add all as negative, Add all as positive

Bottom Left: Text Summary

Najveći svjetski proizvođač **zrakoplova**, tvrtka Airbus, od International Lease Financea dobila je narudžbu za 20 **zrakoplova** A350 vrijednih ukupno 4 milijarde dolara. Narudžba se odnosi na verzije **zrakoplova** A350-800 i A350-900 s opcijom da tvrtka može odabrati i verziju A350-1000. U zasebnom je priopćenju za javnost International Lease Finance (ILFC) objavio kako isporuke tih **zrakoplova** očekuje između 2014. i 2017. godine. ILFC je promijenio prvotnu narudžbu od 16 **zrakoplova**, a najnovija je ipak manja od one koju su očekivali analitičari.

Center Panel: Labeled documents (10)

Manual label	Auto output	Article Id
✓ Yes	0.922	599384
✓ Yes	0.914	599307
✓ Yes	0.915	598180
✓ Yes	0.923	601191
✓ Yes	0.961	592215
✓ Yes	0.920	597366
✓ Yes	0.957	594999
✓ Yes	0.974	595056
✓ Yes	0.967	594465
✓ Yes	0.950	597258

Buttons: 1/1, 1/6

Bottom Center: Unlabeled documents (10003)

Auto output	Article Id	Certainty
0.661	594096	0.161
0.297	592663	0.203
0.211	591513	0.289
0.159	591464	0.341
0.143	591935	0.357
0.101	591914	0.399
0.096	591933	0.404
0.095	591553	0.405
0.095	591995	0.405

Buttons: 1/6

Right Panel: Classifier construction

Learn ☒ Unlabeled are negative
Random negatives ratio: 3

Label distribution

C=64.00

☒ Show labeled ☒ Show unlabeled

Finish category

Reset Label all unlabeled as negative

To finish, please label all articles.

Relative impacts

Feature	Impact
zrakoplov	3.492
airbus	0.216
tvrtka	0.140
priopćenje	0.043
narudžba	0.025
opcija	0.024
godina	0.019
proizvođač	0.019
vrijedan	0.019
svjetski	0.018
dobiti	0.014
moći	0.011
javnost	-0.007
nov	-0.003
odabrati	-0.003
očekivati	-0.024
odnos	-0.008
ukupan	-0.005
objaviti	-0.013
privatan	0.000
dolar	-0.002
verzija	0.000
odnosti	-0.006
international	0.000
analitičar	-0.019
promijeniti	0.003
isporuka	-0.011
najveći	-0.004
manja	0.000
zaseban	-0.005
financ	0.000
leasti	0.000

Graph control

Y-scale ☐ Smoothness ☐

Classifier and trainer modules (7)

- Technologies
 - services and client interfaces:
 - .NET 2.0 (C#), Microsoft Message Queues
 - search, vectorization, classification engine:
 - C++
 - database:
 - Microsoft SQL Server (or compatible)

Demo

- Classifier and trainer modules

Key phrase extraction module

- Vague problem definition
- High level description: summarizing an article with just a few key phrases

Key phrase extraction – index.hr

Vlada zbog duga iz 1991. podiže kredit od 260 milijuna eura

Piše: J.C.
Objavljeno prije 50 minuta

 Recommend

– Tekst +

Foto: Davor Pongračić / Cropix

HRVATSKA vlada u Erste i Hypo banci podiže novih 260 milijuna eura kredita, uz 5,5 posto kamata. Odlučeno je to na današnjoj sjednici Vlade. Kredit se podiže zbog otplate dugova koji datiraju još iz 1991. godine.

Tada su, naime, izdane obveznice za restrukturiranje gospodarstva u svrhu sanacije bankarskog sustava. Manji dio duga, u iznosu od 482,3 milijuna kuna, pokrit će se novcem iz državnog proračuna. Ostatak duga sanirat će se novim sredstvima koja država uzima na kredit.

VEZANO

VIJESTI | BLOGOVI | TAGOVI

sanacija banaka

zaduživanje

banke

kredit

Vlada

SJEDNICA VLADE

Vlada digla kredit da bi vratila dug od sanacije banaka '91.

Vlada je na svojim internetskim stranicama objavila tekst ugovora o pristupanju Hrvatske EU.

Piše: pmž/VLM, vmo/VLM

Vlada se opet odlučila zadužiti. Najnoviji kredit podiže u Erste i Hypo banci, i to za plaćanje obveznica za restrukturiranje gospodarstva iz 1991. godine koje su izdane za sanaciju bankarskog sustava. Inače, imatelji tih obveznica i jesu uglavnom banke.

Nepovoljni uvjeti

Prema riječima ministrice financija Martine Dalić, riječ je o povratu duga iz 1991. godine čija se glavnica kontinuirano uvećavala svakih šest mjeseci za indeks proizvođačkih cijena u industriji.

TAGOVI

Martina Dalić,
kreditno zaduživanje,
Moody's, Jadranka
Kosor

Key phrase extraction module (2)

- Non-functional requirements
 - Fast extraction (should process thousands of documents per day)
 - Quick response ($< 1s$)
 - Reliability
 - Low maintenance cost
- Functional requirements
 - Key phrases should enable horizontal, cross-category linkage between related documents
 - The system should automatically discover new relevant key phrases

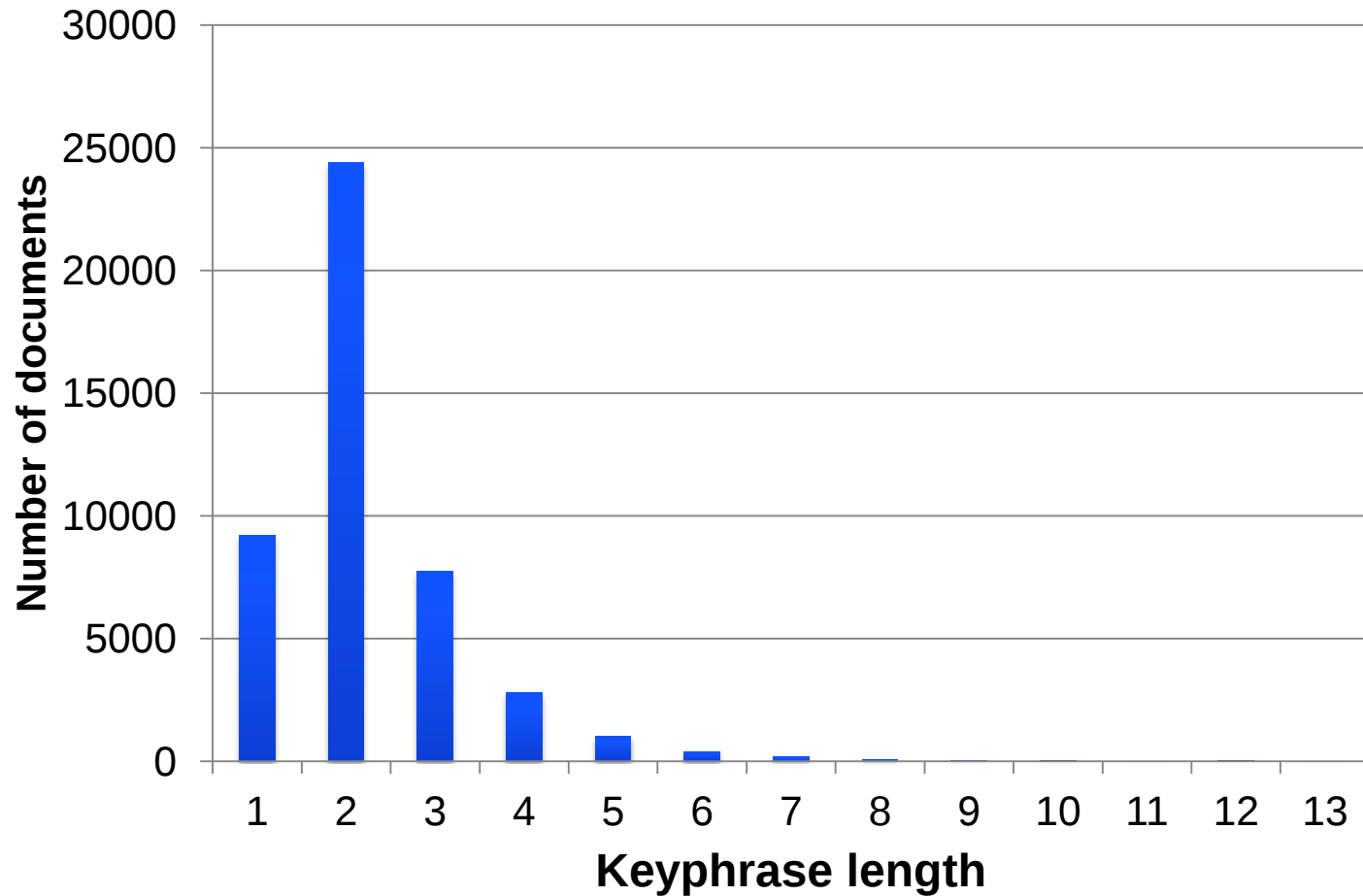
Key phrase extraction module (3)

- Assignment vs. extraction
- Supervised vs. unsupervised approach
- Tradeoffs: speed & reliability vs. complexity & quality

Key phrase extraction module (4)

- Analysis of sample of 39684 documents:
 - On average 4.88 key phrases per document.
 - 194 000 assigned key phrases
 - 59% appear in document
 - In 6% of documents none of assigned key phrases appear in the document
 - 46 000 different key phrases
 - 31 000 assigned only once

Key phrase extraction module (5)



Key phrase extraction module (6)

- Specification refinement
 - The system will extract key phrases consisting of up to four words.
 - Extracted key phrases should be explicitly mentioned in the text.
 - We will make the best effort on producing grammatically correct keywords, but the system might be wrong sometimes.

Key phrase extraction module (7)

- Idea:
 - Generate candidate keyphrases
 - Rank them and select the best ones
- Candidate extraction
 - Extract all 1-, 2-, 3- and 4-grams
 - Removing invalid key phrases
 - POS filtering (e.g. key phrases can't contain a verb)
 - MSD filtering

Key phrase extraction module (8)

- POS filtering
 - ANN
 - NAN
 - ...
- MSD filtering
 - (Ax N?) Ng (psihološka ispitivanja regruta)
 - (Ax N?) Nd Ng (besplatni pristup bazi podataka)
 - Nn N? Ng (internet ispitivanja regruta)
 - ~50 rules

Key phrase extraction module (9)

- Term frequency
 - Simple online computation
 - $tf(i)$ = normalized key phrase frequency
- Inverse document frequency
 - $idf(i) = \log(\text{number of documents} / \text{number of documents where term } i \text{ appears})$
- $score(term(i)) = tf(i) * idf(i)$

Key phrase extraction module

- Evaluation
 - Key phrase extraction - highly subjective task
 - *World Championships in Athletics, Championships in Athletics, Athletics, Athletic Championship*
 - F2 measure – favors recall over precision
$$F2 = 5PR / (4P + R)$$
 - Low inter-annotator agreement
 - Average human F2: 63.9%
 - This system: 52.2%
 - Lowest performing annotator: 51.0%
 - Supervised system: 56.9%

Wrapper and debugging modules

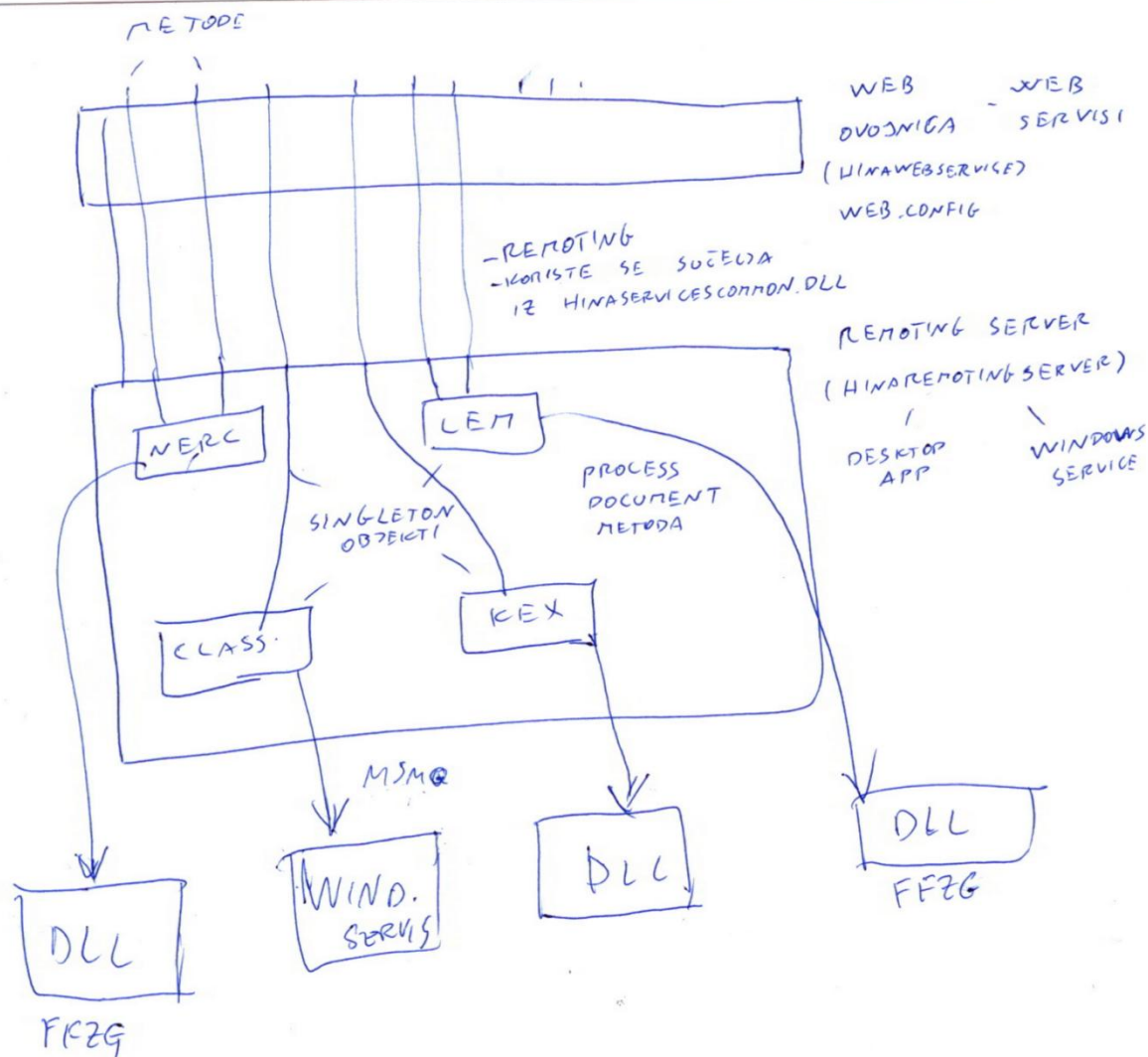
- We need to coordinate other modules (lemmatization, NERC, classification and keyword extraction)
- Goals:
 - Easy administration
 - Quick error detection
 - System robustness
 - Error recovery

Wrapper and debugging modules

- Easy administration:
 - The wrapper is a Windows service
- Error detection:
 - All actions are logged
 - We can track down (the series of) actions that lead to error
- System robustness:
 - Protect modules from invalid input
- Error recovery:
 - Try to restart failed modules
 - Try to restart the whole service
 - Send alerts

Wrapper and debugging modules

- Documentation



HINA OVOJNICA ARCHITEKTURA

Morphological normalization in KTN

Morphological normalization

- When converting texts to vector representation, we want to normalize word forms -> remove flectional differences
- Croatian language is morphologically rich
 - {brzinom, brzine, brzinama,...} -> brzina (speed)
 - {brz, brzina, brzinski,...} -> brz (fast)
- Normalization is used during:
 - text classification
 - information retrieval

belgija belgija N=fpg
belgijama belgija N=fpd
belgijama belgija N=fpl
belgijama belgija N=fpi
belgije belgija N=fsg
belgije belgija N=fpn
belgije belgija N=fpa
belgije belgija N=fpv
belgiji belgija N=fsd
belgiji belgija N=fsl
belgijo belgija N=fsv
belgijom belgija N=fsi
belgiju belgija N=fsa
belgijska belgijski Aspnpn
belgijska belgijski Aspnpa
belgijska belgijski Aspnpv
belgijska belgijski Aspfsn
belgijska belgijski Aspfsv
belgijske belgijski Aspmpa
belgijske belgijski Aspfsf
belgijske belgijski Aspfpn
belgijske belgijski Aspfpv
belgijske belgijski Aspfpv
belgijski belgijski Aspmsn
belgijski belgijski Aspmsa
belgijski belgijski Aspmsv
belgijski belgijski Aspmpn
belgijski belgijski Aspmpv
belgijskih belgijski Aspmpg
belgijskih belgijski Aspnpv
belgijskih belgijski Aspfpv
Belgijanac belgijanac N=msn
Belgijanac belgijanac N=msa
Belgijanaca belgijanac N=mpg
Belgijanca belgijanac N=msg

belgijanac N=mpa
Belgijancem belgijanac N=msi
Belgijanci belgijanac N=mpn
Belgijanci belgijanac N=mpv
Belgijancima belgijanac N=mpd
Belgijancima belgijanac N=mpl
Belgijancima belgijanac N=mpi
Belgijancu belgijanac N=msd
Belgijancu belgijanac N=msl
Belgijance belgijanac N=msv
belgijanaka belgijanka N==pg
Belgijanci belgijanka N==sd
Belgijanci belgijanka N==sl
Belgijanci belgijanka N=fsd
Belgijanci belgijanka N=fsl
Belgijanka belgijanka N==sn
Belgijanka belgijanka N==pg
Belgijanka belgijanka N=fsn
Belgijanka belgijanka N=fpg
Belgijankama belgijanka N==pd
Belgijankama belgijanka N==pl
Belgijankama belgijanka N==pi
Belgijankama belgijanka N=fpd
Belgijankama belgijanka N=fpl
Belgijankama belgijanka N=fpi
Belgijanke belgijanka N==sg
Belgijanke belgijanka N==pn
Belgijanke belgijanka N==pa
Belgijanke belgijanka N==pv
Belgijanke belgijanka N=fsg
Belgijanke belgijanka N=fpn
Belgijanke belgijanka N=fpa
Belgijanke belgijanka N=fpv
Belgijanki belgijanka N==sd
Belgijanki belgijanka N==sl

Belgijanki belgijanka N==sl
Belgijanki belgijanka N=fsd
Belgijanki belgijanka N=fsl
Belgijanko belgijanka N==sv
Belgijanko belgijanka N=fsv
Belgijankom belgijanka N==si
Belgijankom belgijanka N=fsi
Belgijanku belgijanka N==sa
Belgijanku belgijanka N=fsa
Belgijac belgijac N=msa
Belgijaca belgijac N=mpg
Belgijca belgijac N=msg
Belgijce belgijac N=msv
Belgijce belgijac N=mpa
Belgijcem belgijac N=msi
Belgijci belgijac N=mpn
Belgijci belgijac N=mpv
Belgijcima belgijac N=mpd
Belgijcima belgijac N=mpl
Belgijcima belgijac N=mpi
Belgijcom belgijac N=msi
Belgijcu belgijac N=msd
Belgijcu belgijac N=msl
Belgijce belgijac N=msv

Croatian morphology

- High level of affixation, numerous phonological alterations
 - Infexion
 - nouns: declination (7 cases, number)
 - verbs: conjugation (tense, person, number, gender)
 - adjectives: declination (7 cases, number, gender), comparison (3 degrees), definiteness
 - Formation (derivation)
 - lots of patterns (verb->noun, noun->noun, noun->verb, noun->adjective, ...)

Croatian morphology (2)

- Example of inflexion:
 - adjective: *brz, brza, brzi, brzima, brzih, brzoj, brze, brzim, brzog, brzoga, brz, brza, brzo, brzom, brzomu, brži, bržeg, brža, brži, bržima, bržih, bržoj, brže, bržim, bržem, bržima, najbrži, bržeg, najbrža, najbržima, najbržih, najbrže, najbržim, najbrži, najbržoj, ...*
 - noun: *brzina, brzinom, brzine, brzinama, brzinu, brzina, brzini*
 - adjective: *brzinski, brzinskom, brzinske, brzinskih, brzinska, brzinskoj, brzinsko, brzinskog, brzinskoga, ...*
 - adverb: *brzo, brže, najbrže, brzinski*
- Example of formation:
 - *brz -> brzina -> brzinski*

Our approach

- 1. Acquisition of inflexional lexicon
 - input: morphological model and an untagged corpus
 - output: inflexional or inflexional/derivational morphological lexicon
- 2. Word form normalization
 - inflexional (lemmatization)
 - inflexional/derivational (close to stemming)
- Advantages
 - can be used as a lemmatizer and as a stemmer
 - high coverage tailored according to a corpus
 - economical price of construction

Model of morphology

- Example: noun paradigm
- *vojn*ik (solider)

case	singular	plural
N	vojn	vojni <i>c</i> -i
G	vojn	vojn
D	vojn	vojni <i>c</i> -ima
A	vojn	vojn
V	vojni <i>č</i> -e	vojni <i>c</i> -i
L	vojn	vojni <i>c</i> -ima
I	vojn	vojni <i>c</i> -ima

Model of morphology (2)

- **Inflexional** and **derivational** patterns are defined
- Model is based on functional programming concepts
- **Functions** are building blocks:
 - condition functions
 - transformation functions
- Functions are defined indirectly with higher order functions:
- E.g.
 - sfx
 - sfx('a')
 - sfx('a')('vojnik') = 'vojnika'
 - sfx('e') alt(pal)
 - (sfx('e') alt(pal))('vojnik') = 'vojniče'

Model of morphology (3)

case	singular	plural
N	vojn ^{ik} -Ø	vojni ^c -i
G	vojn ^{ik} -a	vojn ^{ik} -a
D	vojn ^{ik} -u	vojni ^c -ima
A	vojn ^{ik} -a	vojn ^{ik} -e
V	vojni ^č -e	vojni ^c -i
L	vojn ^{ik} -u	vojni ^c -ima
I	vojn ^{ik} -om	vojni ^c -ima

$(\lambda s.\text{ends}('k','g','h')(s) \rightarrow \text{cgr}(s), \{\text{null}, \text{sfx}('a'), \text{sfx}('u'),$
 $\text{sfx}('om'), \text{sfx}('e')\text{alt}(\text{pal}), \text{sfx}('i')\text{alt}(\text{sib}), \text{sfx}('ima')$
 $\text{alt}(\text{sib}), \text{sfx}('e')\})$

Model of morphology (4)

- Advantages:
 - Close to morphology descriptions in traditional grammars
 - Highly expressive formalism
 - Adaptable to morphologically similar languages
- Implemented in Haskell
 - Purely functional programming language
 - Model implementation does not require programming skills
- *If you want to create a Haskell morphological model for your language using our formalism, do send us an email (ktlab@fer.hr)*

Lexicon acquisition

- In order to get the lemmas we use the morphological model and an untagged corpus
 - frequencies of word forms in corpus + heuristics
 - very high linguistic precision is not a major goal
 - manual intervention not needed

Šnajder, J., Dalbelo Bašić, B., Tadić, M. (2008) Automatic Acquisition of Inflectional Lexica for Morphological Normalisation. *Information Processing and Management*, 44 (5), 1720–1731.

Results

- Example of a small lexicon:
 - News corpus, 20 Mw
 - 100 inflexional + 300 derivational patterns
 - contains 42.000 lemmas / 500.000 word forms
 - Quality
 - coverage: **96% over word forms, 98% over tokens**
 - understemming < 5%
 - overstemming < 4%
 - Can be improved with manual editing

Implementation of normalization

- lexicon lookup, 957K unique word forms
- uncompressed size: 225 MB
 - not practical to hold in main memory
(at least it was not some years ago)
- compressed size: <1MB
 - minimal finite state automaton
 - C++ implementation
 - equal speed of lookup ($O(\#letters)$)

TermeX

- <http://ktlab.fer.hr/termex>
- A tool for automatic collocation extraction and terminology lexis construction
- Using 14 different association measures to rank candidate collocations
- Similar ideas used for keyword extraction

TermeX - GUI

TERMEX

Projekt: Korpus Kolokacije Pomoć

Mjera: DICE N-Gram: 3-gram Količina: 10.000

<untitled>
listopad_4457.txt

PM 2 DI 2 G1 2 G2 2 G3 2 G4 2 G5 2 G6 2 FR 2 G* 3 DI 3

Rang	N-Gram	Vrijednost	Kolokacija
8309	Ferenc Gyurcsany	0,522	☐
8310	Giuliano Amato	0,522	☐
8311	Haris Ćorbić	0,522	☐
8312	Silvije Degen	0,522	☐
8313	Slobodana Miloševića	0,522	☒
8314	Condoleezza Riće	0,52	☐
8315	Romano Prodi	0,519	☐
8316	Višnja Samardžija	0,519	☐
8317	divljih svinja	0,519	☒
8318	kvadratnog korijena	0,519	☐
8319	prirodnih bogatstava	0,516	☒
8320	kaznene prijave	0,516	☒
8321	kriminalističku obradu	0,515	☐
8322	Josi Mraoviću	0,514	☐
8323	tržišnog natjecanja	0,514	☐
8324	zločinački pothvat	0,513	☒
8325	europsku unijuKnjiga	0,508	☐
8326	Ika Šarić	0,507	☐
8327	prometnoj pećnici	0,506	☐

Spreman

Označene kolokacije

- Slobodana Miloševića
- admiralu Brovetu
- divljih svinja
- istražnih radnji
- kaznene prijave
- kriznog štaba
- optuženičku klupe
- organiziranog kriminala
- pilom rezao
- pokušavši orobiti
- prirodnih bogatstava
- smrtnom posljedicom
- spolnim odnošajem
- task force
- višesatnom pretresu
- vox populi
- vukovarskoj trojci
- zakopali leševe
- zamlatiti čekićem
- zapadnog Balkana
- zloćudni tumor

Rang	Pozicija	Prefix	N-Gram	Sufix
10	318557	granici, nisu mogli ništa prigovoriti Petraču. Istražni je sudac po službenoj dužnosti podigao	kaznenu prijavu	protiv Petrača te je protiv njega počeo sudski proces koji je završio krajem 2002. oslobađajućo
11	356256	terijalne štete, zadarski protorej Petar Jovanović odlučio je da neće ići a proces pokretanja	kaznene prijave	Pretpostavlja se da napad na Sv. Iliju nije izoliran incident, nego se naslanja na niz počinjenih
12	561509	ov otac, obavljeno u uobičajenom roku. Kako je, pak, policija odradila svoj posao nakon prvih dviju	kaznenih prijava	potvrđuje nalaz njihova unutarnjeg istražnog tijela, na čijem je čelu načelnik interventnog vo
13	590289	hički bolesnik zadržan na liječenju. Zbog više kaznenih djela, protiv nasilnika će se podnijeti	kaznene prijave	Općinskom državnom odvjetništvu. (čr)Nakon brzo izvedene akcije, okončane bez težih posljedica
14	803268	ogu najaviti privođenje u četvrtak ujutro te tiskovnu konferenciju na kojoj ćemo iznijeti detalje	kaznene prijave	Slučaj smo odradili ozbiljno i predano, te vrlo brzo došli do počinitelja- kazao nam je nače
15	852996	ju biti kvalitetno pripremljeni tako da se u što kraćem roku može početi s njihovom provedbom.	Kaznena prijava	protiv KuzmanićaPULA - Općinsko državno odvjetništvo u Bujama protiv dviju fizičkih i dviju
16	853380	anju i krivotvorili službene isprave. Ministarstvo financija i pazinska porezna uprava podnijeli su	kaznenu prijavu	protiv Kuzmanića i još tri osobe. Tužiteljstvo je prošli mjesec zaprimilo prijavu kojom se pulsk
17	877375	orati platiti 15.000 kuna. Nakon policijske obrade, saslušao ih je istražni sudac. Pustio ih je uz	kaznenu prijavu	zbog krađe. Osoblje kafića"Blues Brothers"nije uspjelozaustaviti dvojackoji je iznio
18	888346	ran je priveden u postajute je nakon kriminalističke obrade oko 17 sati priveden istražnom sucu uz	kaznenu prijavu	zbog teške krađe.Potraga za automobilom objavljena je preko Interpolakojem je i prijavljen ne
19	892619	poštovanje zakonskih odredbi stajalo je vlasnika kioska u Biogradu oduzimanja 1400 komada petardi i	kaznene prijave	zbog nedopuštene trgovine. Taj Biograđanin, naime, nije imao dokumentaciju o porijeklu eksplozivni
20	963590	ina njih bavila se švercom automobila, kamatarenjem i utjerivanjem dugova. To pokazuje i dvadesetak	kaznenih prijava	protiv Brkića koje se uglavnom svode na nasilničko ponašanje.Na drugoj se strani našao Josip
21	977283	je nedugo nakon toga na kućnoj adresi te mu je izmjereno više od promila alkohola u krvi. Bit će	kazneno prijavljen	Od policijske uprave istarske jučer nismo mogli dobiti nikakve dodatne informacije pa niti podat
22	990100	ak Josip Posarić (64) iz Mača. Protiv 34-godišnjeg vozača »punta« iz Zagreba bit će podnijeta	kaznena prijava	nadležnom odvjetništvu.Prizor s mjesta nesrećeZORAN GREGUREKZoran GREG Francuzi rekli NE europsk
23	1015929	ilo vidi u činjenici što su prvi uhiceni svoje iskaze na policiji dali 29. veljače 2004., a na	kaznenoj prijavi	od 1. ožujka 2003. nije bilo Petračeva imena. Smatra a kaznena prijava protiv Petrača podignuta j
24	1016002	i 29. veljače 2004., a na kaznenoj prijavi od 1. ožujka 2003. nije bilo Petračeva imena. Smatra a	kaznena prijava	protiv Petrača podignuta je tek 16. ožujka 2004. jer su, prema njegovu mišljenju, za to postojali

CADIAL Search

- <http://cadial.hidra.hr>
- Full text search over a collection of 20,000 legal documents
- Documents are automatically indexed using Eurovoc descriptors
- HIDRA (Croatian Information, Documentation and Referral Agency) assures that additional metadata is correct:
 - Regulation status (valid / invalid)
 - Area of activity
 - EU accession chapter

CADIAL Search (2)

- Possibility to search:
 - Full text
 - Titles
 - EUROVOC thesaurus terms
 - Historical versions
 - ...
- Legislation: semi-structured documents: possibility to take the structure into account when computing the relevance of article, section etc.
- Successful participation at the INEX competition 2008 [Mijic et al. 2009]

CADIAL Search - demo

Search: ☒ Text ☒ Title ☒ Descriptors ☒ Valid regulations ☒

Show: ☒ Titles only ☐ Basic information ☐ Detailed information ☐

Show advanced options ☐

Database updated: 11.03.2010.

Results: 7 (0.2188 s)

1. Pregled činidbenih jamstava izdanih od 1. 1. 2002. do 31. 12. 2002. (NN 028/2004) -
...d.d. HRVATSKA BRODOGRADNJA TROGIR d.d....d.d. HRVATSKA
BRODOGRADNJA TROGIR d.d....d.d. HRVATSKA BRODOGRADNJA
TROGIR d.d....
NN004354.xml - 7 KB
Eurovoc descriptors: [financial intervention](#), [guarantee](#)
Comments:
• Datum stupanja na snagu: 2004-03-05
Linked documents:
• Temeljni propis: [Zakon o izvršavanju Državnog proračuna Republike
Hrvatske za 2002. godinu \(NN 116/2001\) - NE VAŽI!](#)

2. Odluka o usvajanju Jamstvenih programa Hrvatske agencije za
malo gospodarstvo (NN 056/2009) -
...društava i obrta, djelatnost **brodogradnje**, proizvodnje čelika, pomorskog...
NN032622.xml - 10 KB
Eurovoc descriptors: [entrepreneur](#), [entrepreneurship](#), [credit](#), [guarantee](#), [State](#)
Comments:
• Datum stupanja na snagu: 2008-05-07
Linked documents:
• Temeljni propis: [Zakon o Vladi Republike Hrvatske \(NN 101/1998\)](#)

3. Pregled financijskih jamstava izdanih od 1. 1. 2001. - 31. 12.
2001. (NN 014/2005) -
...Č 03.04.2001. Hrvatskoj **brodogradnji** Trogir d.o...
NN014691.xml - 11 KB

Eurovoc thesaurus

Eurovoc / 375

politics / 43

international relations / 22

european communities / 5

law / 6

economics / 30

trade / 2

finance / 10

monetary relations / 1

financial institutions and credit / 2

free movement of capital / 1

public finance and budget policy / 3

budget / 2

taxation / 1

social questions / 10

education and communications / 52

science / 3

business and competition / 16

employment and working conditions / 63

transport / 15

environment / 27

agriculture, forestry and fisheries / 4

agri-foodstuffs / 1

production, technology and research / 7

energy / 1

industry / 37

state institutions of the republic of croatia / 19

HR geography / 2

Zakon o knjižnicama (NN 105/1997)

Regulation type: [law regulation](#)

Regulation status:

- VALID REGULATION

Eurovoc descriptors (6):

(hide list...)

- [information profession](#)
- [library](#)
- [national library](#)
- [university library](#)
- [legal deposit](#)
- [library stock conservation](#)

Area of activity:

- [information media documentation statistic](#)

Comments (2):

- Datum stupanja na snagu: 1997-10-17
- Stupanjem na snagu Zakona o odgoju i obrazovanju u osnovnoj i srednjoj školi (NN 87/08) prestaje se primjenjivati odredba članka 28. ovog Zakona na školske knjižnice

Linked documents (4):

(hide list...)

- Izmjene/dopune: [Ispravak Zakona o knjižnicama \(NN 005/1998\)](#)
- Izmjene/dopune: [Zakon o izmjenama i dopunama Zakona o knjižnicama \(NN 104/2000\)](#)
- Izmjene/dopune (EU usklađenje): [Zakon o izmjenama i dopunama Zakona o knjižnicama \(NN 069/2009\)](#)
- Vidi i: [Zakon o odgoju i obrazovanju u osnovnoj i srednjoj školi \(NN 087/2008\)](#)

Bylaws (8):

(hide list...)

- [Pravilnik o upisniku knjižnica i knjižnica u sastavu \(NN 139/1998\)](#)
- [Pravilnik o polaganju stručnih ispita u knjižničarskoj struci \(NN 021/1999\)](#)
- [Standardi za narodne knjižnice u Republici Hrvatskoj \(NN 058/1999\)](#)
- [Uredba o osnivanju Hrvatske knjižnice za slijepe \(NN 115/1999\)](#)
- [Standard za školske knjižnice \(NN 034/2000\)](#)
- [Pravilnik o matičnoj djelatnosti knjižnica u Republici Hrvatskoj \(NN 043/2001\)](#)
- [Pravilnik o reviziji i otpisu knjižnične građe \(NN 021/2002\)](#)
- [Pravilnik o zaštiti knjižnične građe \(NN 052/2005\)](#)

Original document: [Zakon o knjižnicama \(NN 105/1997\)](#)

ZASTUPNIČKI DOM SABORA REPUBLIKE HRVATSKE

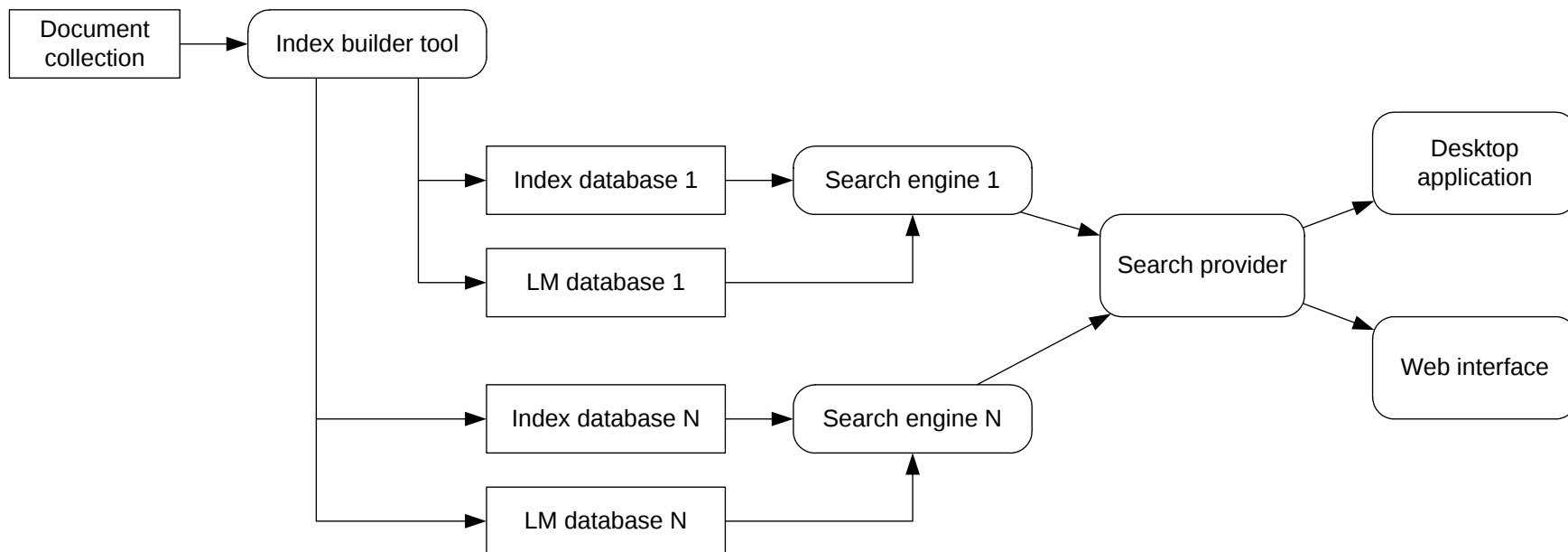
Na osnovi članka 89. Ustava Republike Hrvatske, donosim

ODLUKU

O PROGLAŠENJU ZAKONA O KNJIŽNICAMA

Prolašavam Zakon o **knjižnicama**, koji je donio Zastupnički dom Sabora Republike Hrvatske na sjednici 19. rujna 1997.

CADIAL Search - architecture



CADIAL Search

- Was it worth the effort?
 - The CADIAL project has received the 2009 Prime Minister Award for special achievements in the field of e-Government in Croatia and the 2009 "Golden Tesla's Egg" Award of the VIDI publishing house for the best innovative solution in ICT for the category Academic Institutions.
 - Search engine was easily embedded in SAT project.

Modeling parallelization

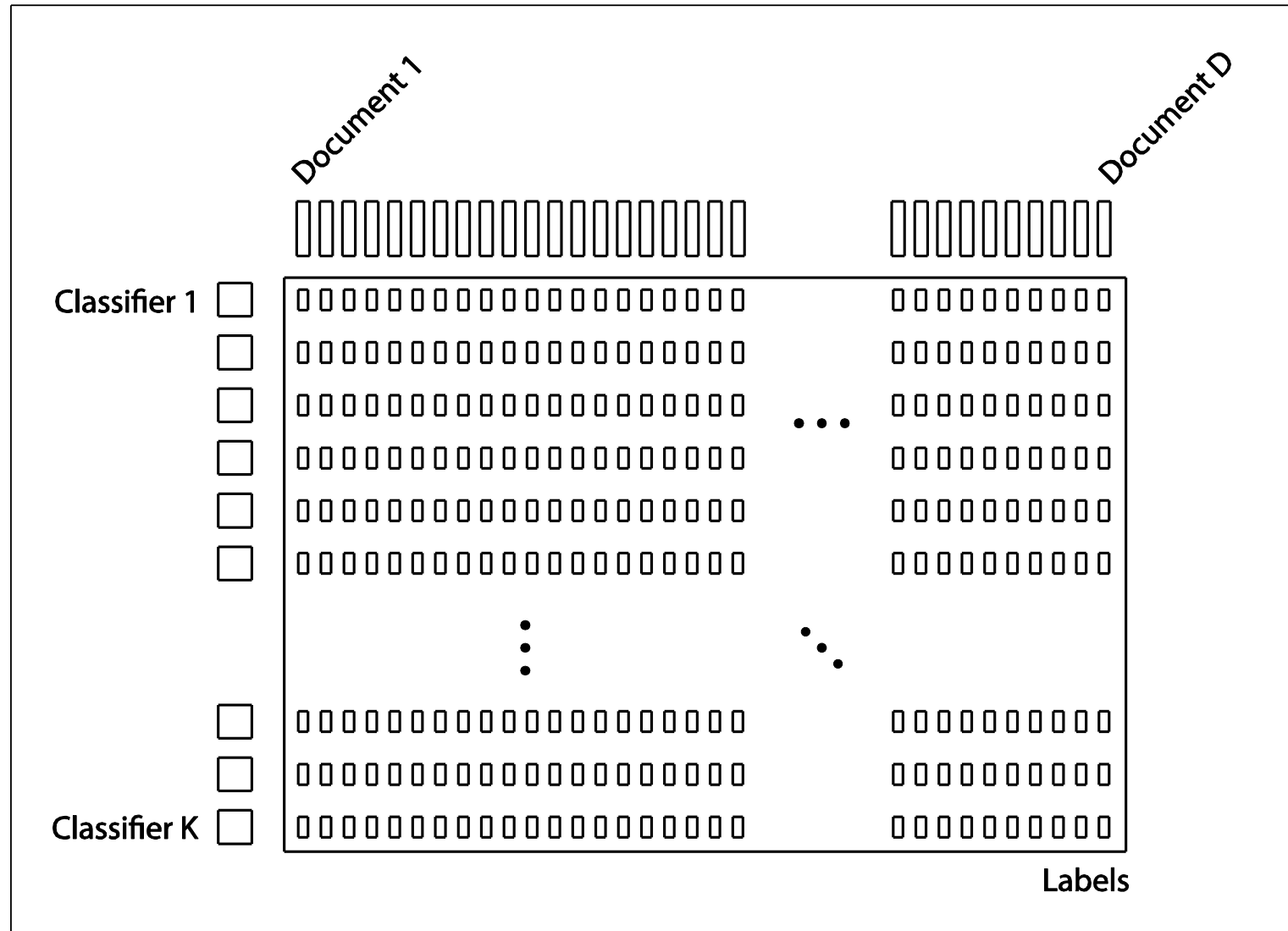
Modeling parallelization

- A simple model of classification calculation speedup due to parallelization
- Possible choices:
 - using one or more computers
 - using computers with less or more memory
 - using processors with one or more cores

Modeling parallelization (2)

- I will show how system architects need to be aware of seemingly trivial relationships
 - e.g. ratio of time to load a classifier from disk and time to classify a document collection
- It is a must to know the intended calculation size in advance
- We will see how these analyses can help decide hardware choice and software development direction

Task description



Task description

- Calculation of each category-document value is independent
 - these problems are known as ***embarrassingly parallel problems***
- Nevertheless, considerable performance variations are possible due to ***housekeeping tasks***
 - e.g. memory allocation, fetching an object from memory, disk, or database

Notation

P – number of processes

J – number of cores available to a process

K – number of categories

k – number of classifiers in a process

D – number of documents

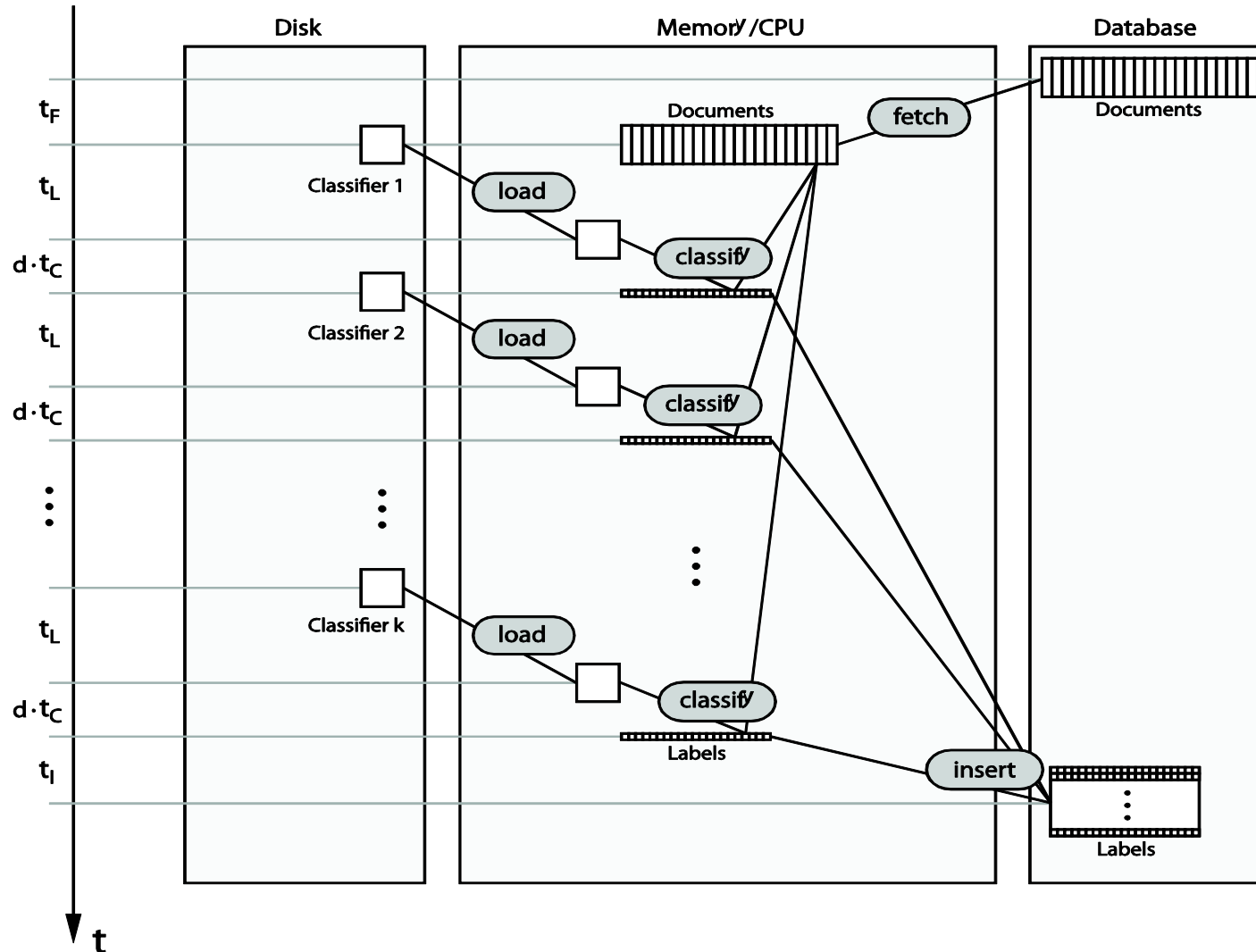
d – number of documents in a process

Some preliminary simplifications

- Waiting time for RAM is negligible
- Document vectors always fit into RAM
 - justified because a classification task is in the range of 10.000 documents of which each is a few kilobytes
- Classifier engines do not necessarily fit into RAM
 - in real setting, they are in a range of 1-3MB, and there are thousands of categories

No concurrency (0P0M)

1) Calculation in one process on one core (0P), with little memory (0M)



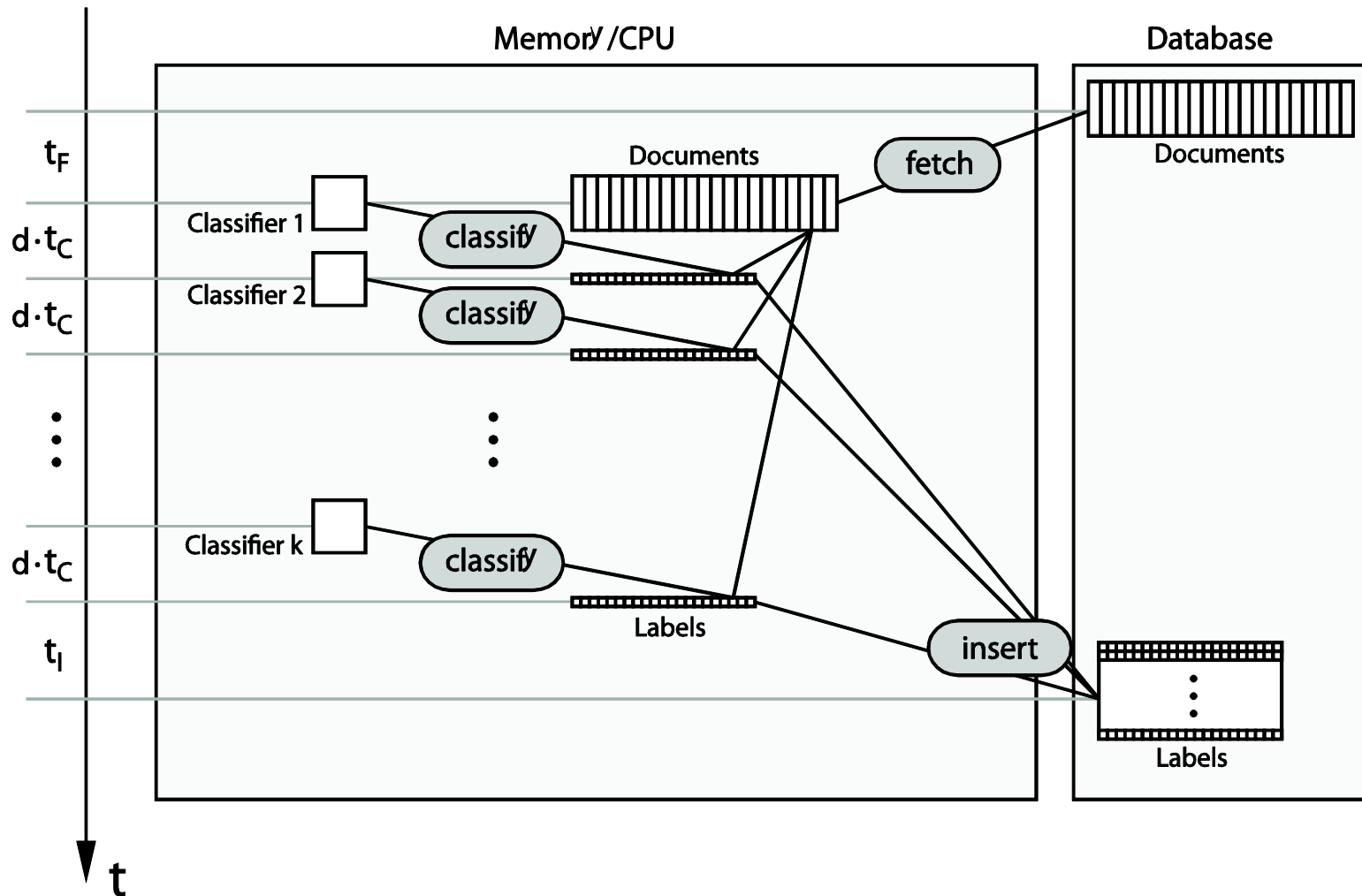
Time calculation for OPOM

- 1) fetching documents from the database (t_F)
 - we assume it is constant (regardless of #doc)
- 2) fetching a classifier (t_L) and classifying all documents ($d \cdot t_C$)
 - repeated k times
- 3) saving results to a database (t_I)
 - we assume it is constant

$$T_{OP+OM}(k, d) = t_F + k \cdot t_L + k \cdot d \cdot t_C + t_I$$

Concurrency within a process (0P1M)

2) Calculation in one process on one core (0P), with more memory (1M)



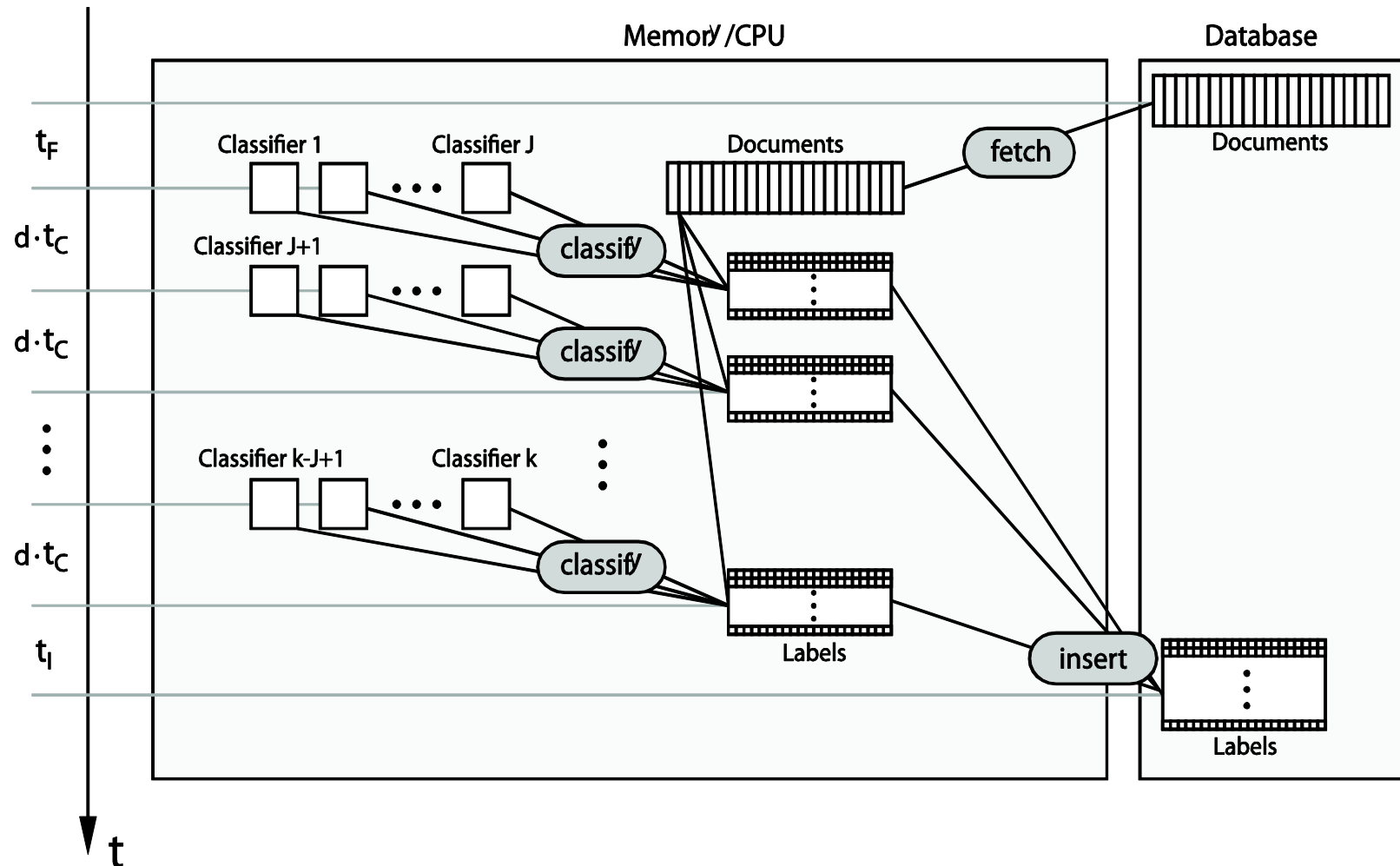
Time calculation for 0P1M

- We do not need to wait for disk anymore since all classifiers are held in memory

$$T_{0P+1M}(k, d) = t_F + k \cdot d \cdot t_C + t_I$$

Concurrency within a process (1P1M)

3) Calculation in one process on many cores (1P), with more memory (1M)



Time calculation for 1P1M

$$T_{1P+1M}(k, d, j) = t_F + \frac{k \cdot d}{j} \cdot t_C + t_I$$

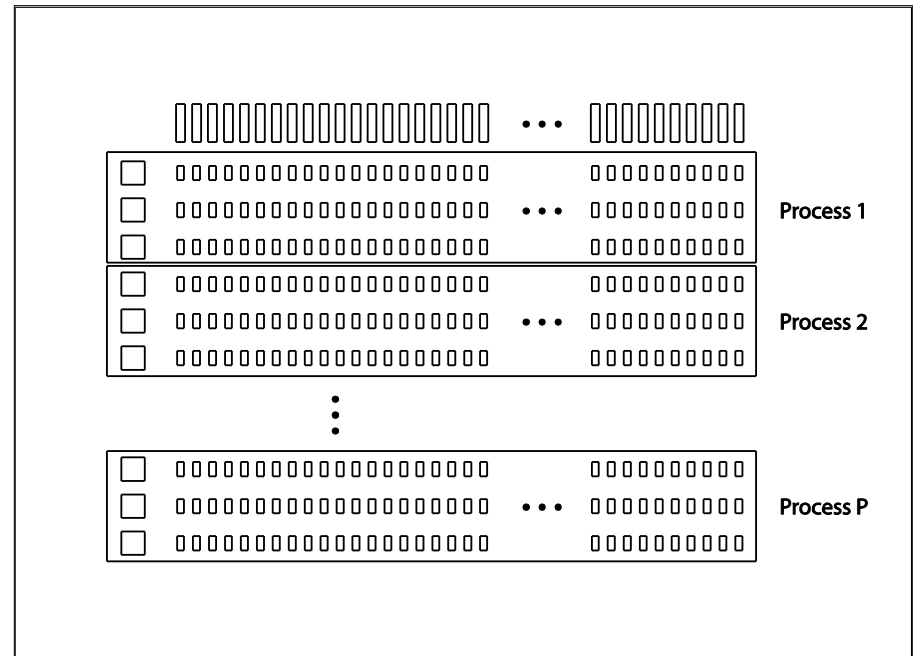
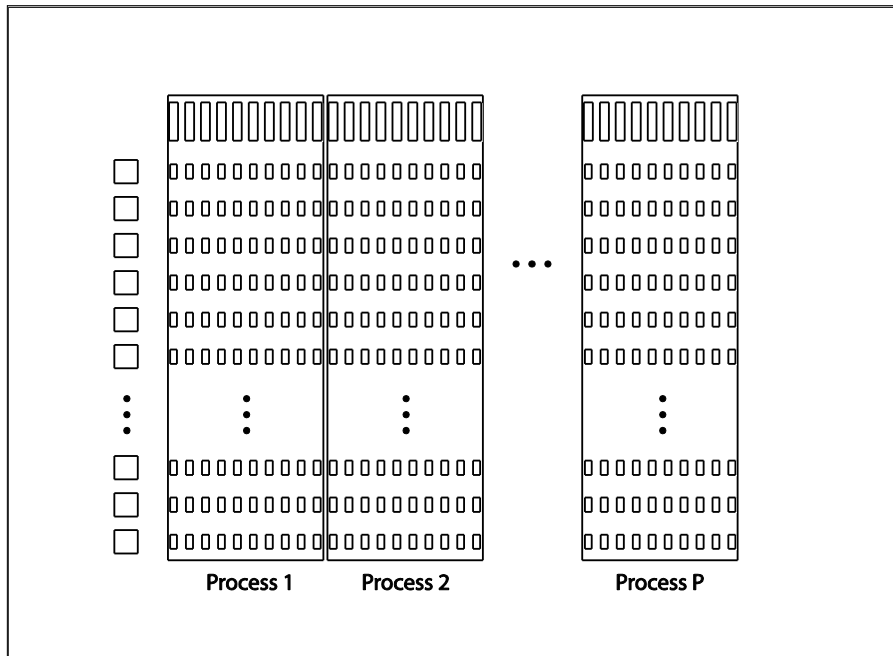
Concurrency within a process (1P0M)

- In our case of low document number it does not make much sense because in practice $t_L \gg t_C$
 - we do not need to go for multithreading if we do not have enough memory
 - if usual document count would be higher, 1P0M would come into consideration

$$T_{1P+0M}(k, d, j) = t_F + k \cdot t_L + \frac{k \cdot d}{j} \cdot t_C + t_I$$

Concurrency among processes

- Each process has dedicated cores
 - e.g. each process is a service on a different computer



Vertical distribution (RV)

- Number of documents per process is divided by P

$$T_{0P+0M+RV}(k, d) = T_{0P+0M} \left(k, \frac{d}{P} \right)$$

$$T_{0P+1M+RV}(k, d) = T_{0P+1M} \left(k, \frac{d}{P} \right)$$

$$T_{1P+1M+RV}(k, d) = T_{1P+1M} \left(k, \frac{d}{P} \right)$$

$$T_{1P+0M+RV}(k, d) = T_{1P+0M} \left(k, \frac{d}{P} \right)$$

Note: Effects of concurrent database insertions are not taken into account

Horizontal distribution (RH)

- Number of classifiers per process is smaller
- In a large memory setting (1M), formulas are alike:

$$T_{0P+1M+RH}(k, d) = T_{0P+1M} \left(\frac{k}{P}, d \right)$$

$$T_{1P+1M+RH}(k, d) = T_{1P+1M} \left(\frac{k}{P}, d \right)$$

Horizontal distribution (RH)

- In small memory (0M), the formula depends on whether the classifiers fit in memory
 - if the number of classifiers is low enough, they might fit into our low memory as well

$$T_{0P+0M+RH}(k, d) = \begin{cases} T_{0P+0M}\left(\frac{k}{P}, d\right), & \frac{k}{P} \cdot m_k > M_k \\ T_{0P+1M}\left(\frac{k}{P}, d\right), & \frac{k}{P} \cdot m_k \leq M_k \end{cases}$$

$$T_{1P+0M+RH}(k, d) = \begin{cases} T_{1P+0M}\left(\frac{k}{P}, d\right), & \frac{k}{P} \cdot m_k > M_k \\ T_{1P+1M}\left(\frac{k}{P}, d\right), & \frac{k}{P} \cdot m_k \leq M_k \end{cases}$$

- With RH, we can increase the memory space and considerably speedup the system by avoiding disk reading for classifiers

Ok, but where are the numbers?

- The numbers depend on the situation 😊
- This kind of analysis needs to be done before product development in order **to adequately choose software architecture** – why?
- Also, it can help **to choose the cheapest hardware** according to market price – how?
- In SAT, we implemented the horizontal distribution with enough memory assumption (0P+1M+RH)

CLARA

Thank you for your attention.

CLARA is an Initial Training Network in the Marie Curie Actions financed by the EC under FP7

