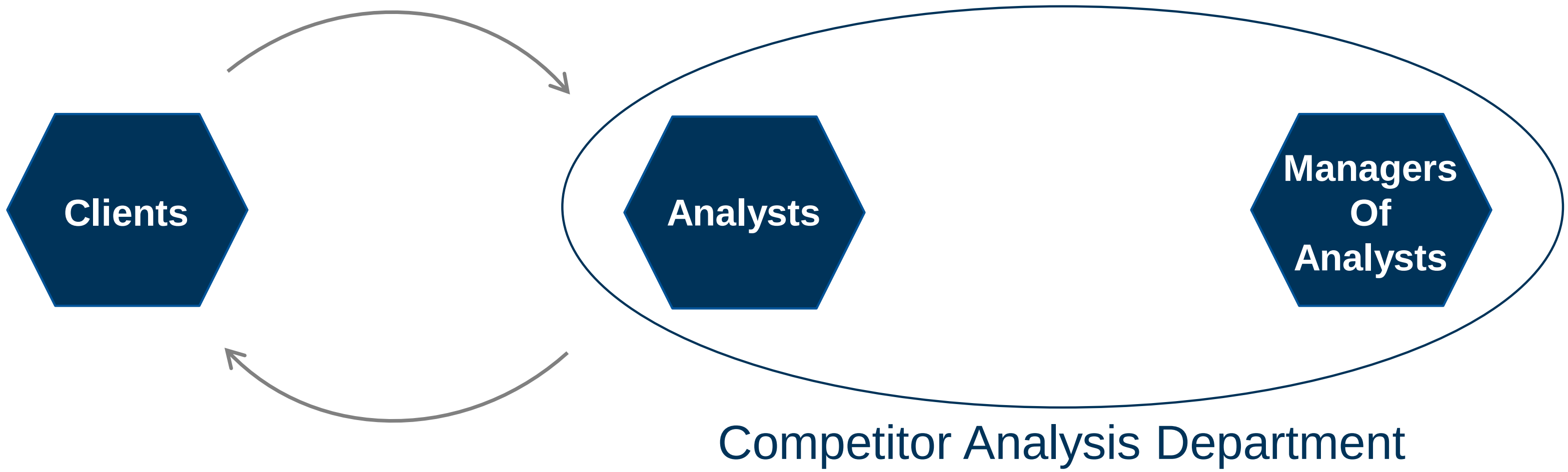
A Machine Learning based approach for the Competitor Information Analysis in the Automotive Industry

Roman Pass, 19.12.2016, Munich

Chair of Software Engineering for Business Information Systems (sebis)
Faculty of Informatics
Technische Universität München
wwwmatthes.in.tum.de

Introduction

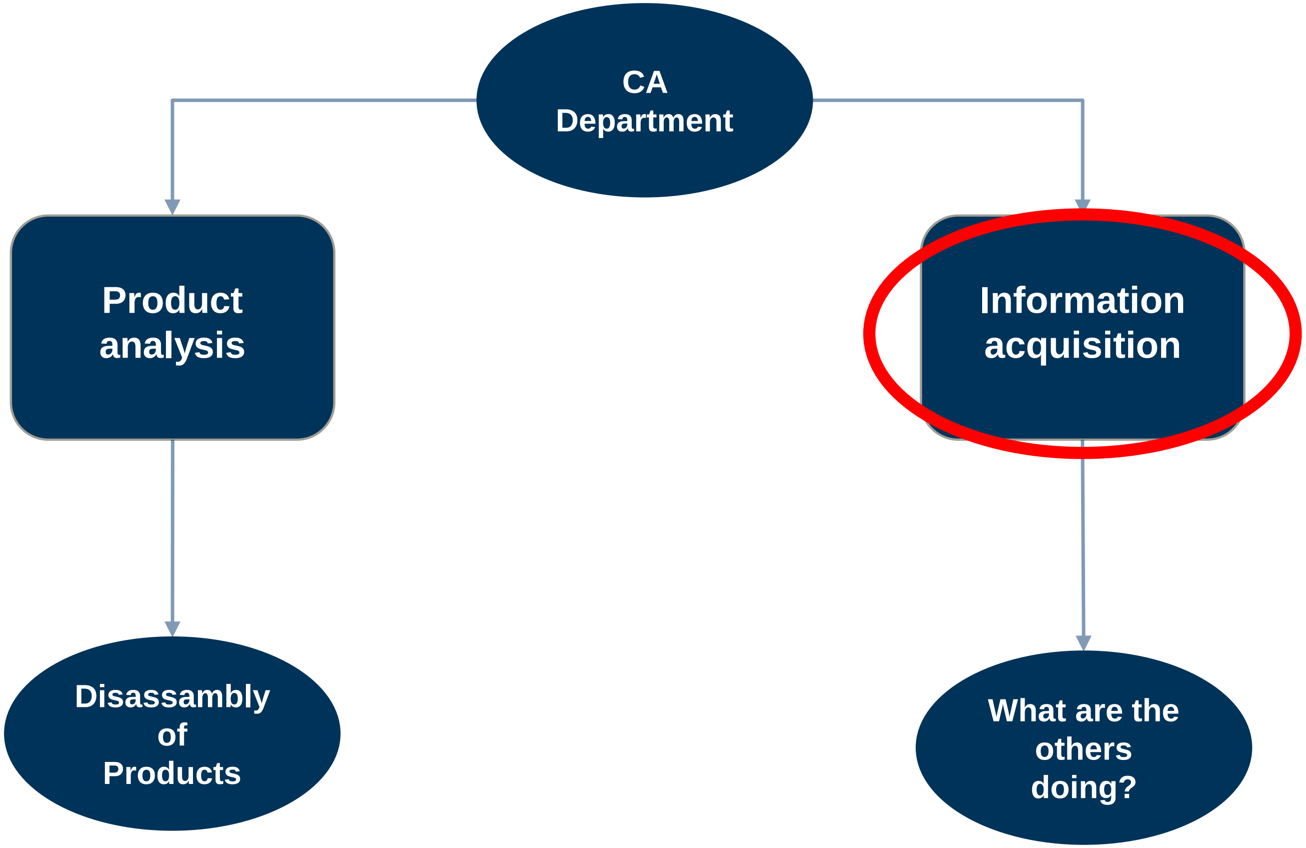
Competitor Analysis (CA) - Roles



Source: Organizing Competitor Analysis Systems, S.Ghoshal, E. Westney

Introduction

Competitor Analysis (CA) - Tasks



Introduction

Motivation: Information acquisition

What kind of information?
- Not only about cars
- New technologies, interior design (furniture), materials, IoT, Smartphones,…

Employees search for information on the **internet**
- Newsfeeds
- Social Networks
- Patent databases
- Search engines

Problem
- ~ 40 Employees
- 2 hrs. per day

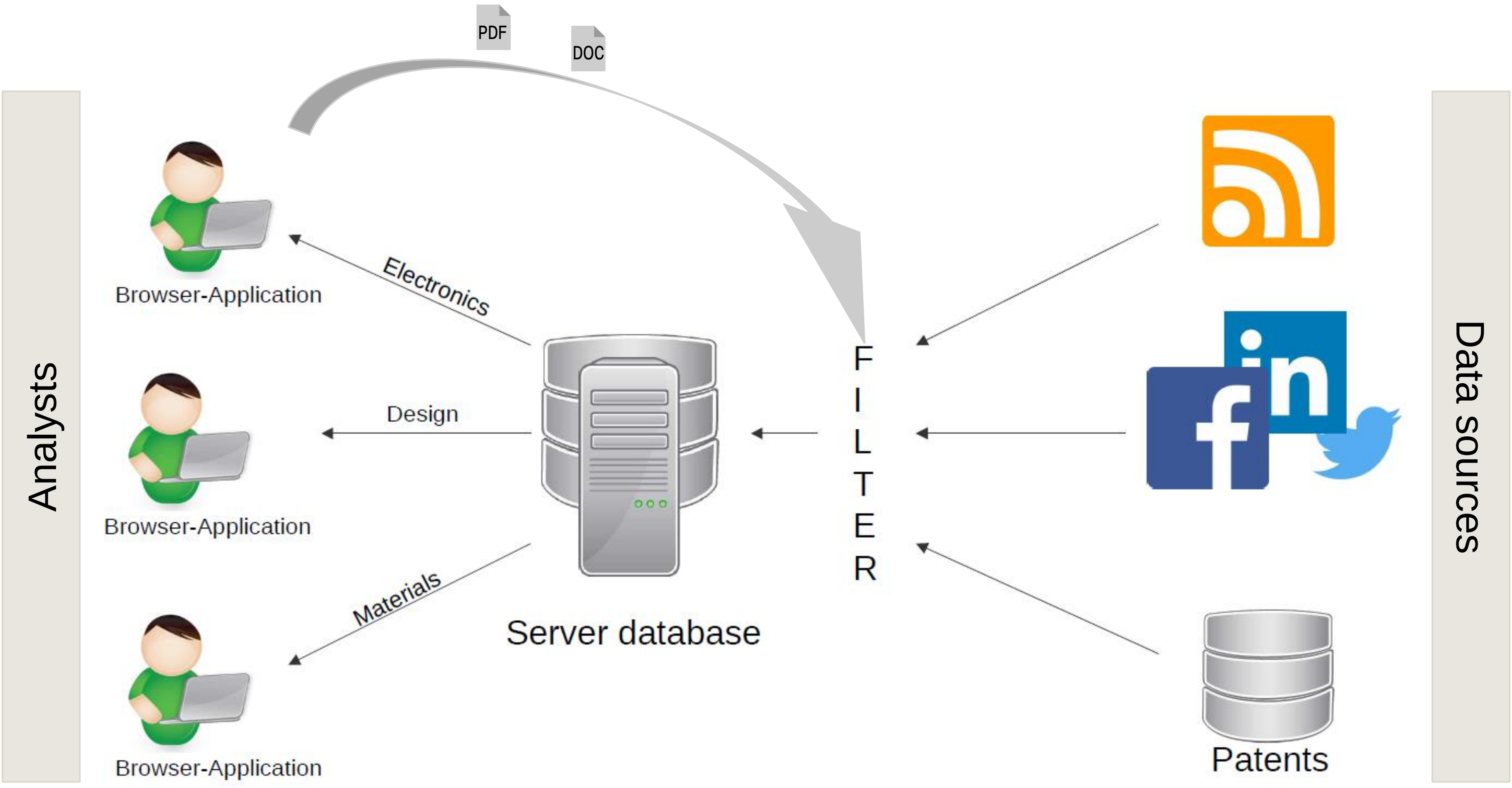
Introduction

Research Questions

- How to improve the existing competitor information analysis process in the automotive domain?
- What kind of **documents** and document **sources** are used for competitor information analysis in the automotive domain?
- Which categories/topics do these documents belong to?
- Which machine learning algorithm performs best for automatic classification of documents to support competitor information analysis in the automotive domain?

Introduction

Information acquisition



Approach

- Comparison of available competitive analysis tools

- User analysis and requirements elicitation at the competitor analysis department
- Mockup design

- Collect sample documents
- Identifying topics for classification
- Manual classification of documents
- Create pipeline for automatic classification
- Analysis of results

- Implement/use server-side **document-management system** for classifying and storing document

- Implement a client application to search, retrieve, and view documents

- Evaluate the system at BMW

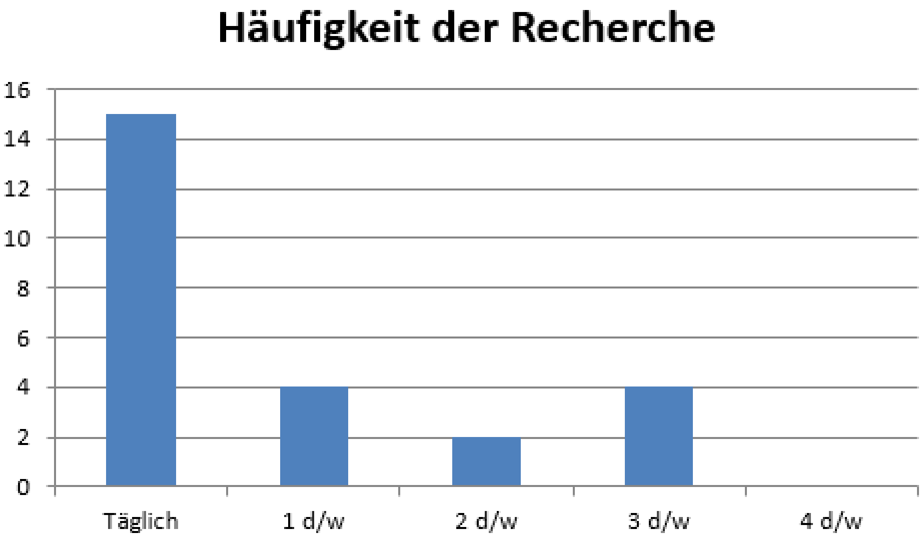
Tool Comparison

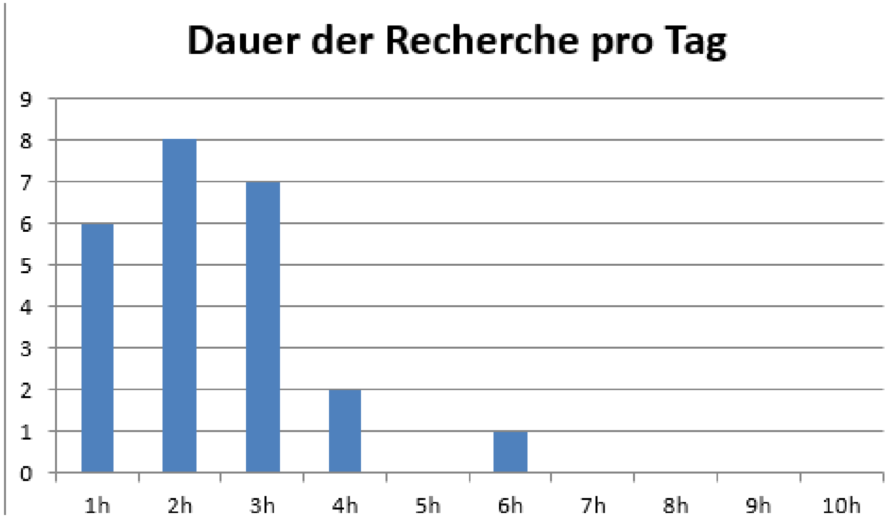
	Keyword mentions	Ad monitoring	backlinks	Social media	Company documents	Newsfeeds	Patents
Topsy	1	0	0	1	0	0	0
The Search Monitor	1	1	0	1	0	1	0
Socialmention	1	0	0	1	0	0	0
Majestic SEO	0	0	1	0	0	0	0
Infinigraph	0	0	0	1	0	0	0
Google Alerts	1	0	0	0	0	0	0

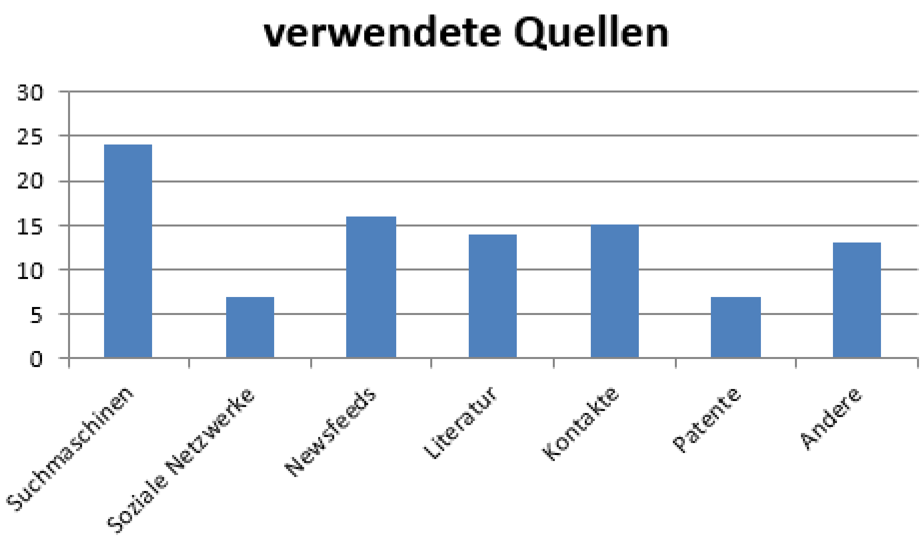
CA relevant (Keyword mentions, Social media, Company documents, Newsfeeds, Patents)

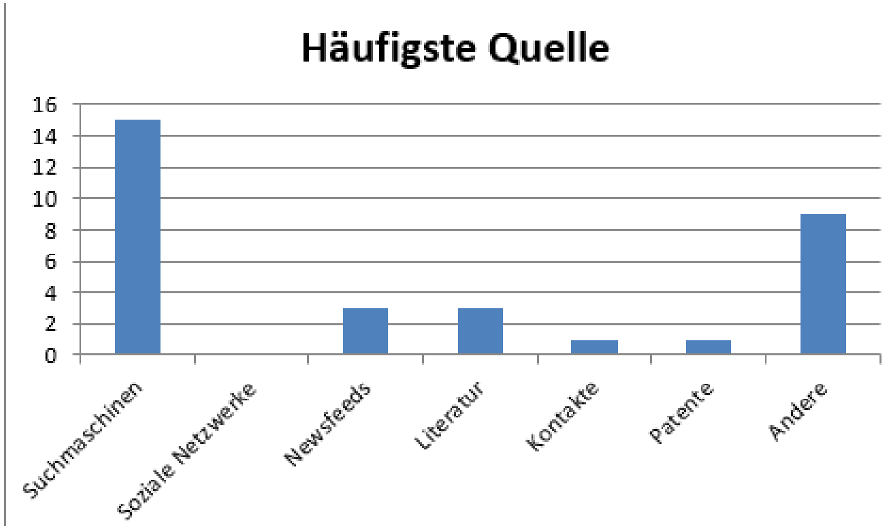
CA irrelevant (Ad monitoring, backlinks)

User Analysis (1)

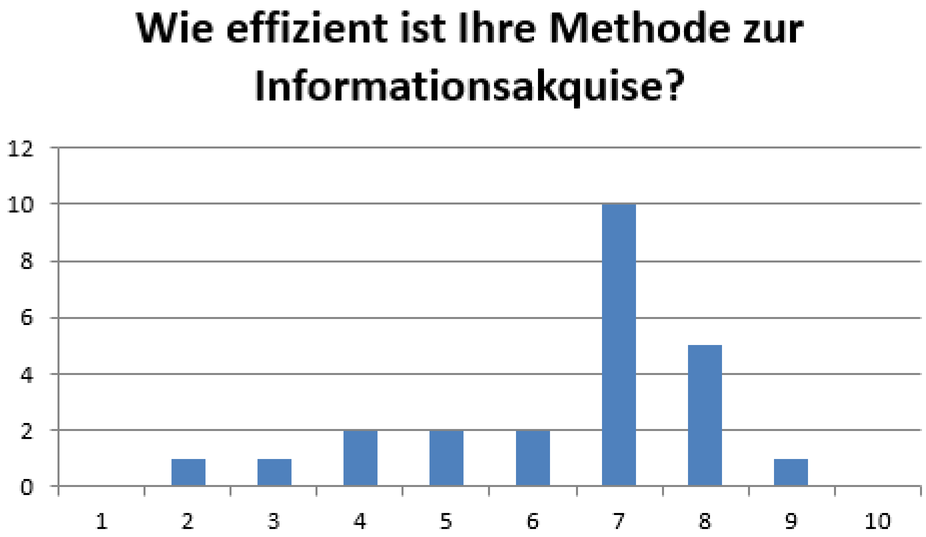


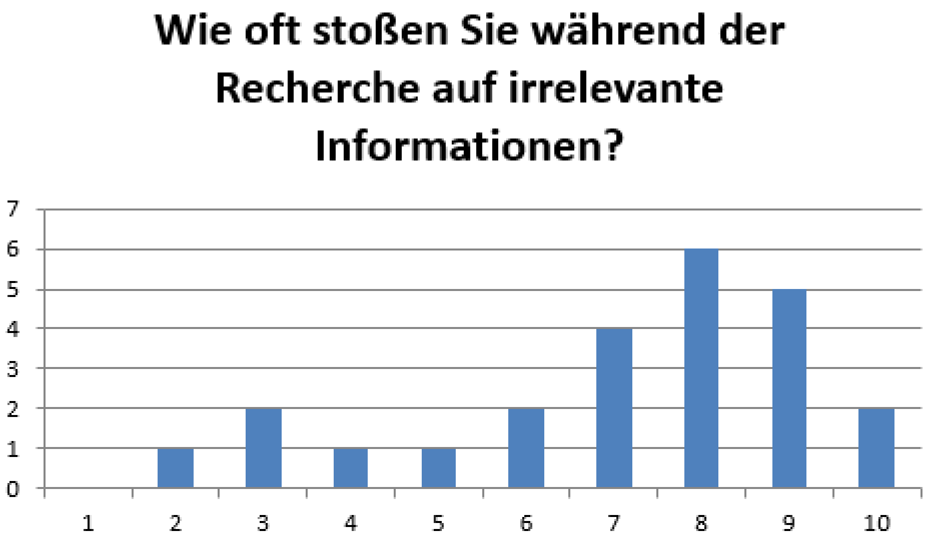




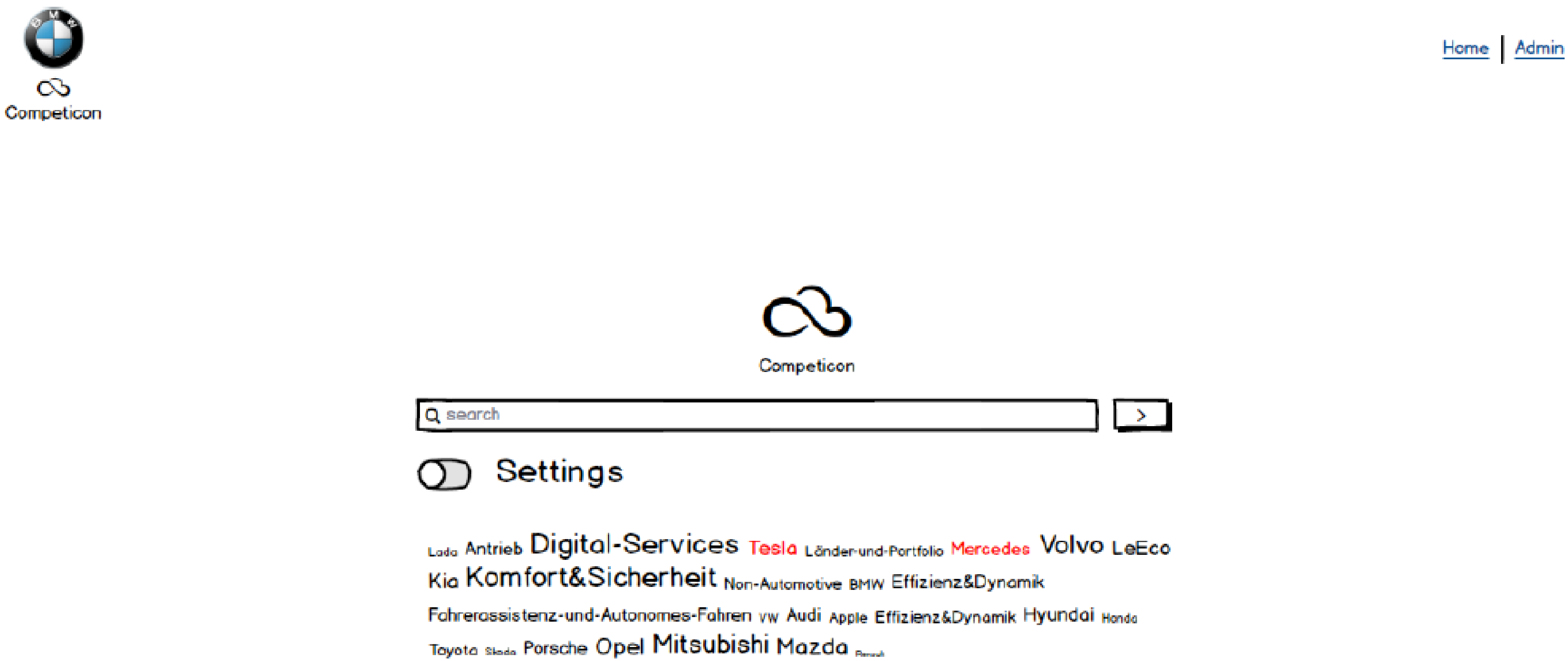


User Analysis (2)



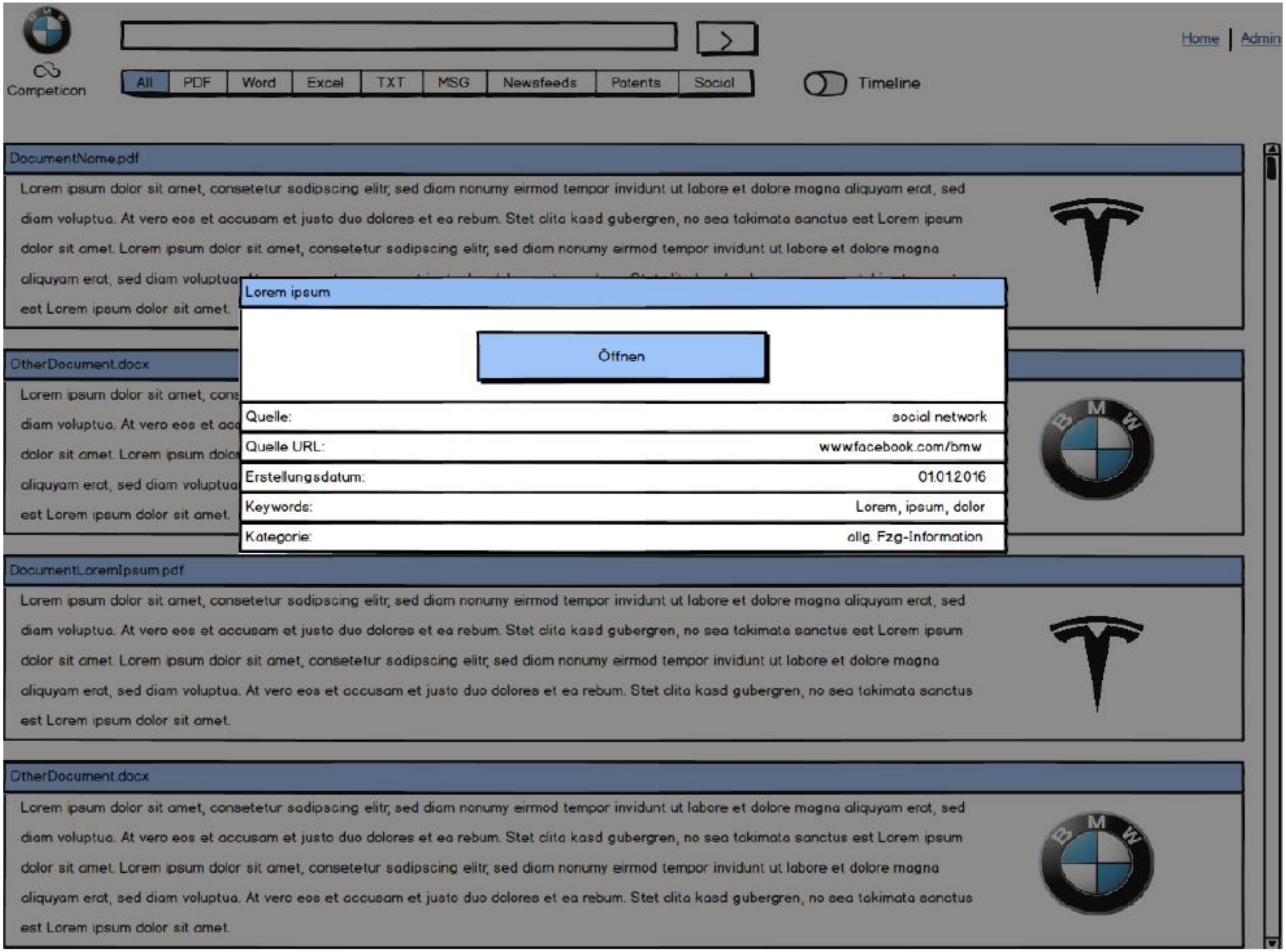


Mockups (1)



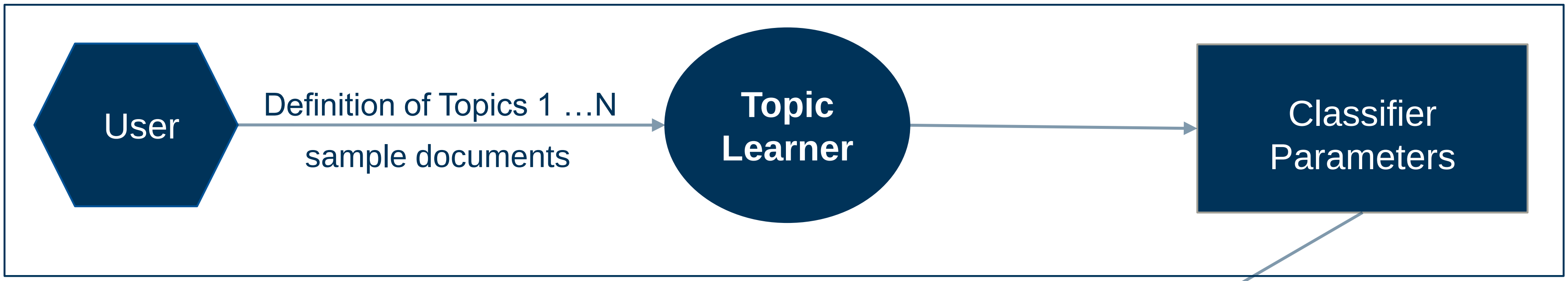
Competicon
Home | Admin
Competicon
search
Settings
Lada Antrieb Digital-Services Tesla Länder-und-Portfolio Mercedes Volvo LeEco
Kia Komfort&Sicherheit Non-Automotive BMW Effizienz&Dynamik
Fahrerassistenz-und-Autonomes-Fahren VW Audi Apple Effizienz&Dynamik Hyundai Honda
Toyota Skoda Porsche Opel Mitsubishi Mazda

Mockups (2)

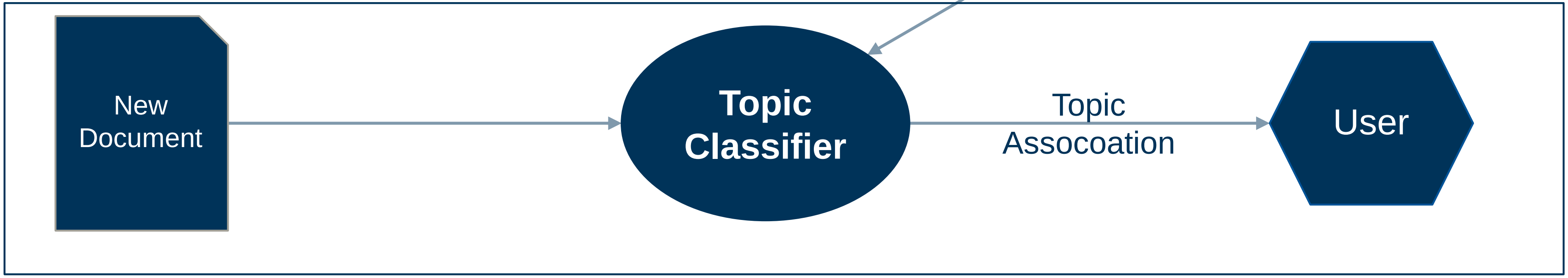

Competicon
All
PDF
Word
Excel
TXT
MSG
Newsfeeds
Patents
Social
Timeline
Home | Admin
DocumentName.pdf
OtherDocument.docx
Lorem ipsum
Öffnen
Quelle: social network
Quelle URL: www.facebook.com/bmw
Erstellungsdatum: 01.01.2016
Keywords: Lorem, ipsum, dolor
Kategorie: allg. Fzg-Information
DocumentLoremIpsum.pdf
Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.
OtherDocument.docx
Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

Automatic Document Classification

Learning Phase



Classification Phase

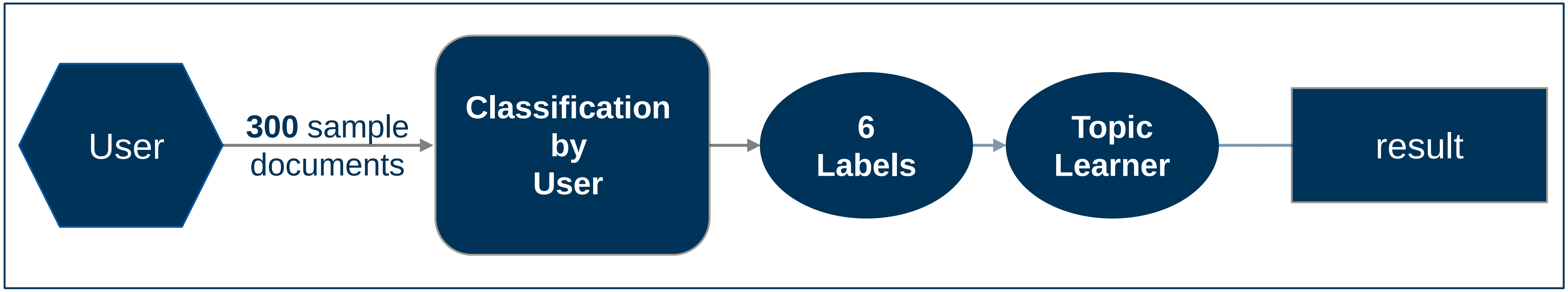


Source: Automatic Document Classification: A thorough Evaluation of various Methods, C. Goller[1], J. Löning[2], T. Will[1], W. Wolff[2]

Automatic Document Classification

Learning phase

- Sample documents sorted by document patterns
- One document belongs to one label



Automatic Document Classification

Initial results of automatic classification

Classification	accuracy	Ø precision	Ø recall
kNN	91,28 %	88,12 %	85 %
Naive Bayes	92,77 %	89,46 %	88,02 %
1-vs-All	47,66 %	7,94 %	16,67 %

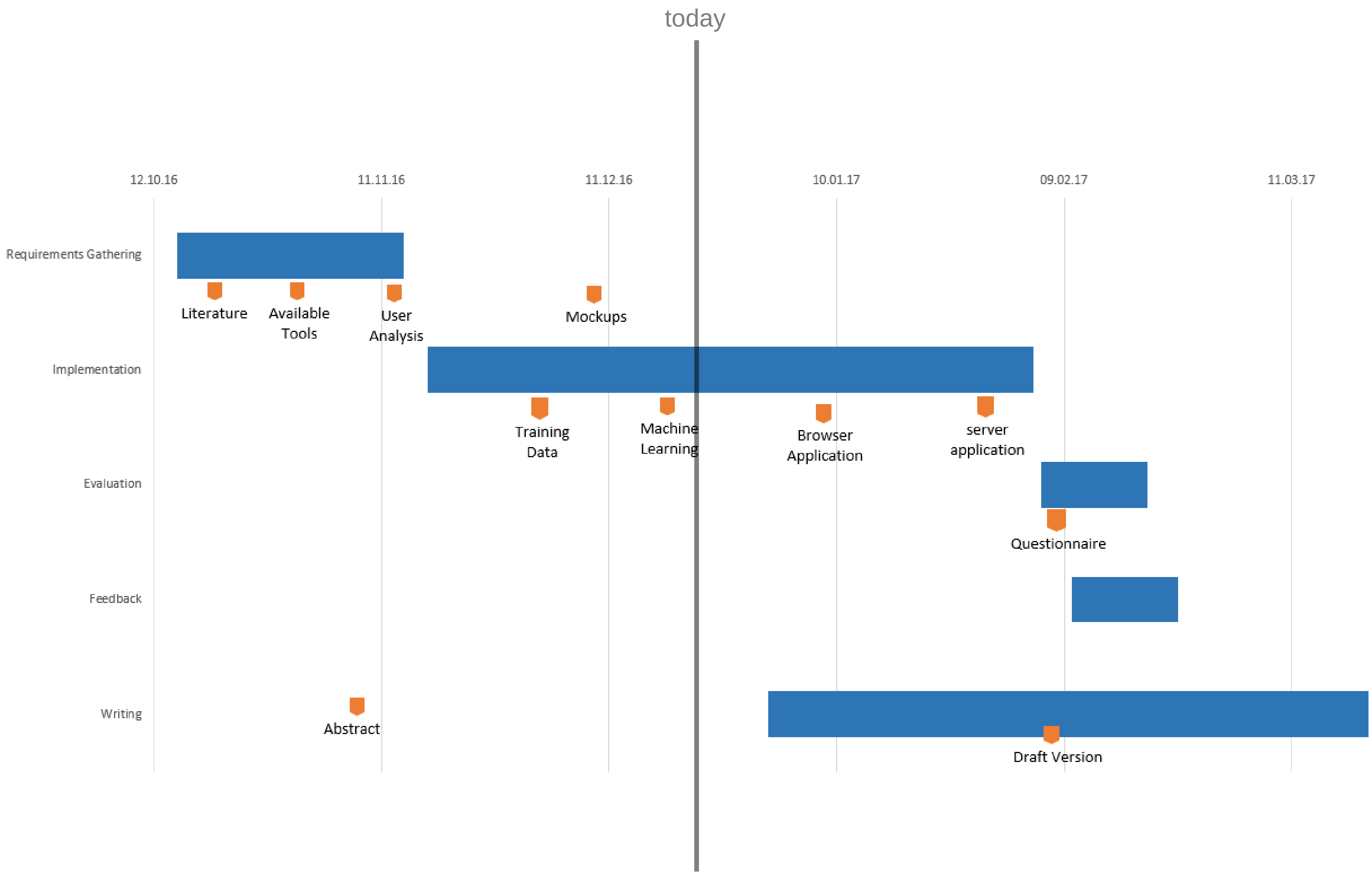
Tool support for competitor information analysis

Technology stack

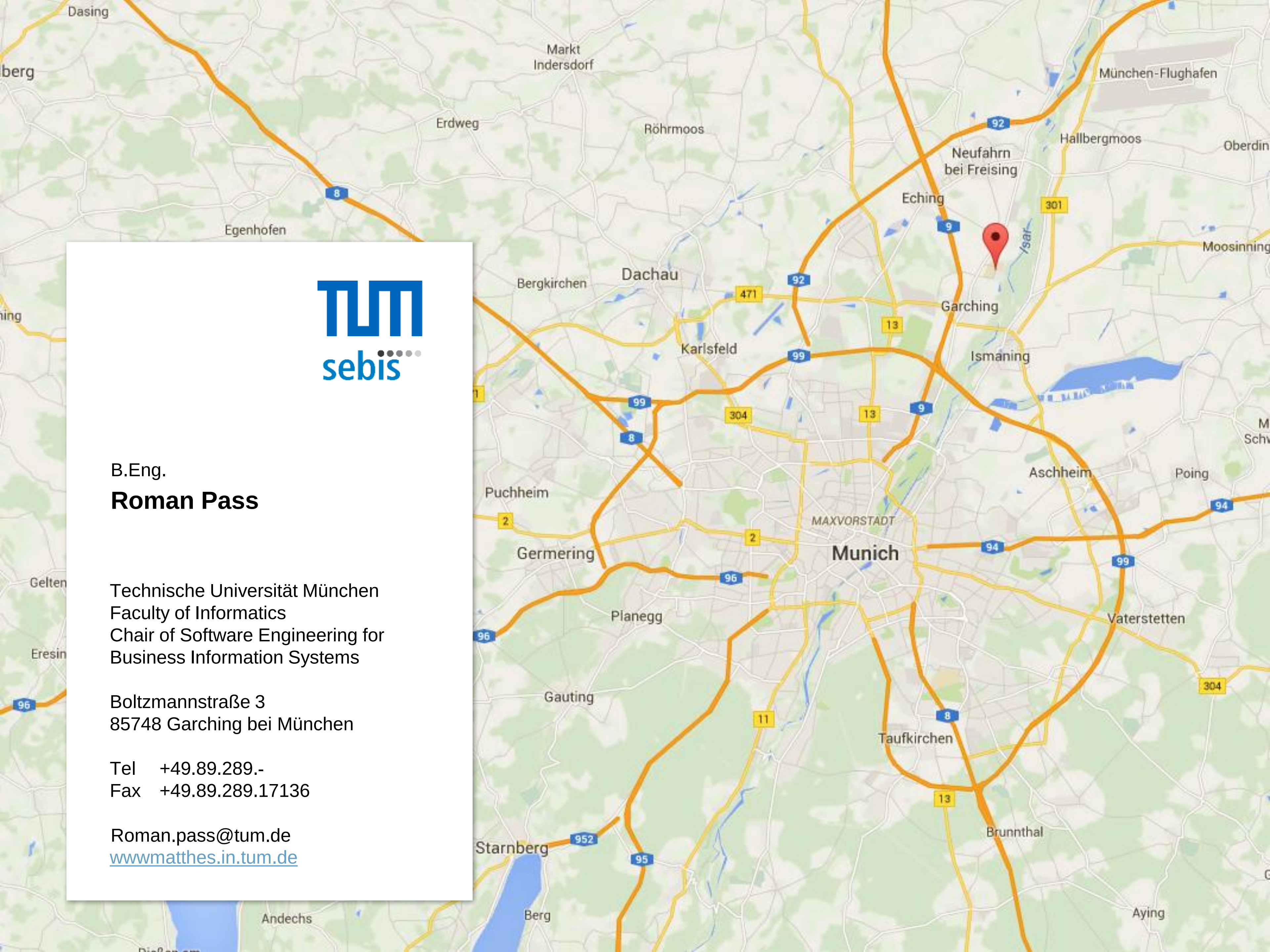
Client

Server

Road Plan



today
12.10.16
11.11.16
11.12.16
10.01.17
09.02.17
11.03.17
Requirements Gathering
Literature
Available Tools
User Analysis
Mockups
Implementation
Training Data
Machine Learning
Browser Application
server application
Evaluation
Questionnaire
Feedback
Writing
Abstract
Draft Version



B.Eng.
Roman Pass

Technische Universität München
Faculty of Informatics
Chair of Software Engineering for
Business Information Systems

Boltzmannstraße 3
85748 Garching bei München

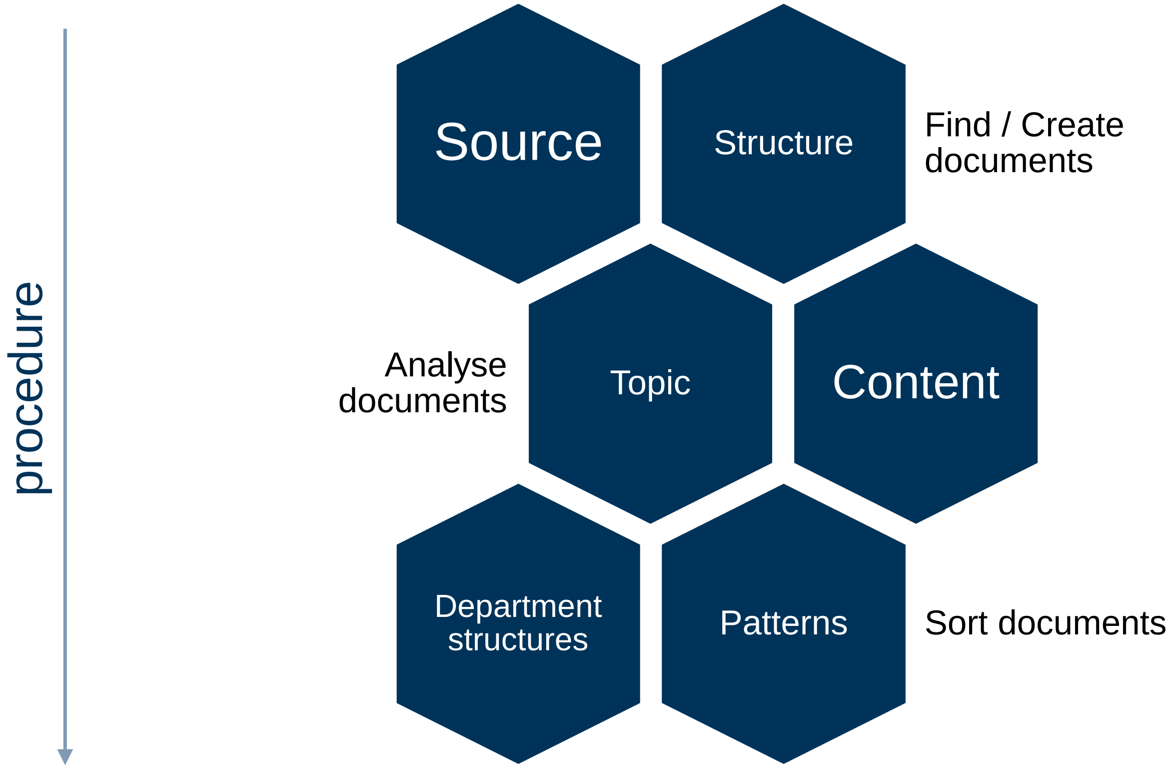
Tel +49.89.289.-
Fax +49.89.289.17136

Roman.pass@tum.de
wwwmatthes.in.tum.de

Backup-Slides

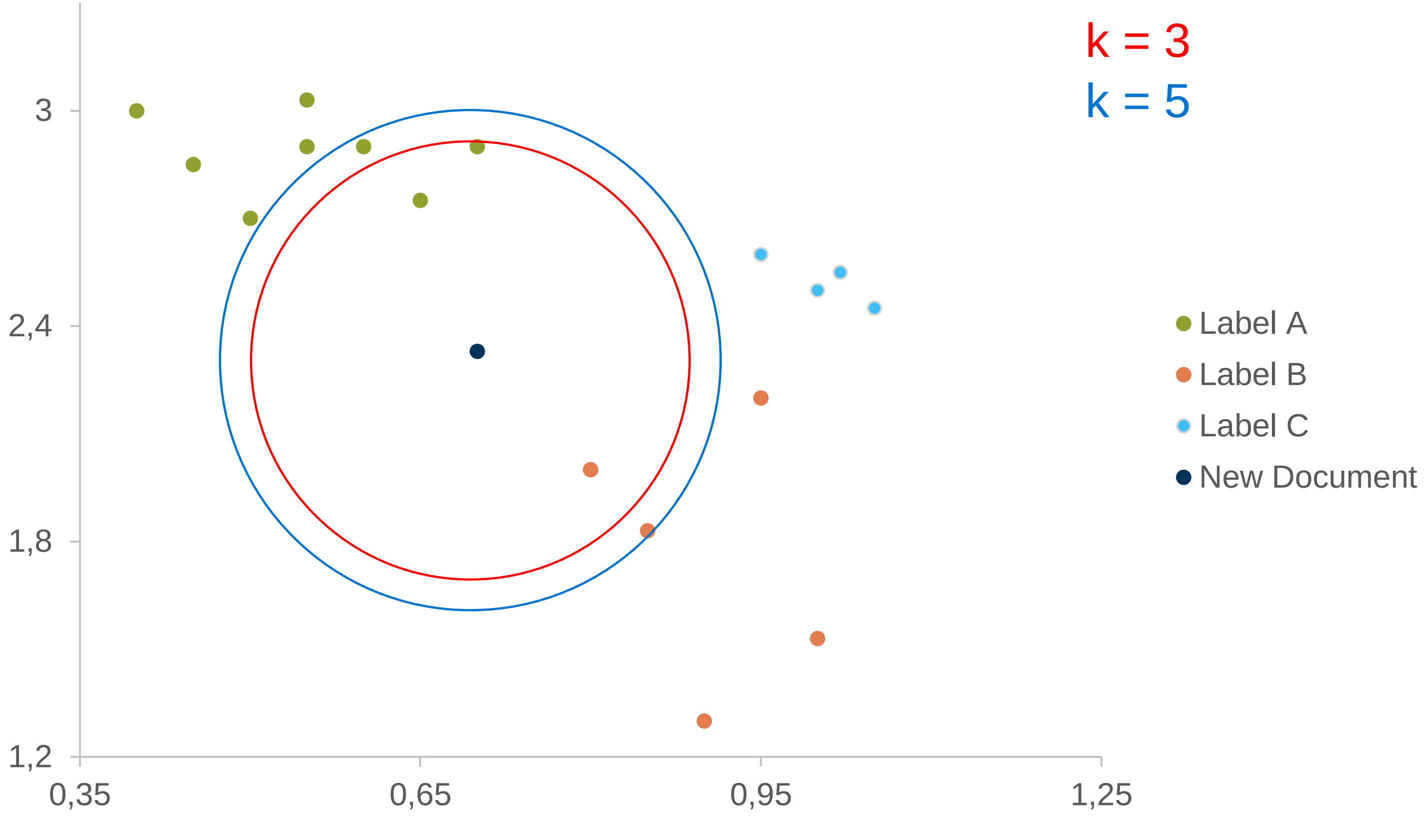
Automatic Document Classification

Document Preparation



Automatic Document Classification

K-Nearest Neighbor



Automatic Document Classification

Naive Bayes

Sunny	Temperature	Exams	Swim
Yes	Hot	No	True
Yes	Warm	Yes	False
No	Yes	No	True
Yes	Cold	Yes	False
Yes	Cold	No	False

Automatic Document Classification

Naive Bayes

Combination	Count
Sunny-Swim	11
Not Sunny-Swim	7
Cold-Swim	2
Warm-Swim	12
Hot-Swim	15
Exams-Swim	2
noExams-Swim	11

Combination	Count
Sunny-NoSwim	5
Not Sunny-NoSwim	6
Cold-NoSwim	11
Warm-NoSwim	4
Hot-NoSwim	3
Exams-NoSwim	10
noExams-NoSwim	2

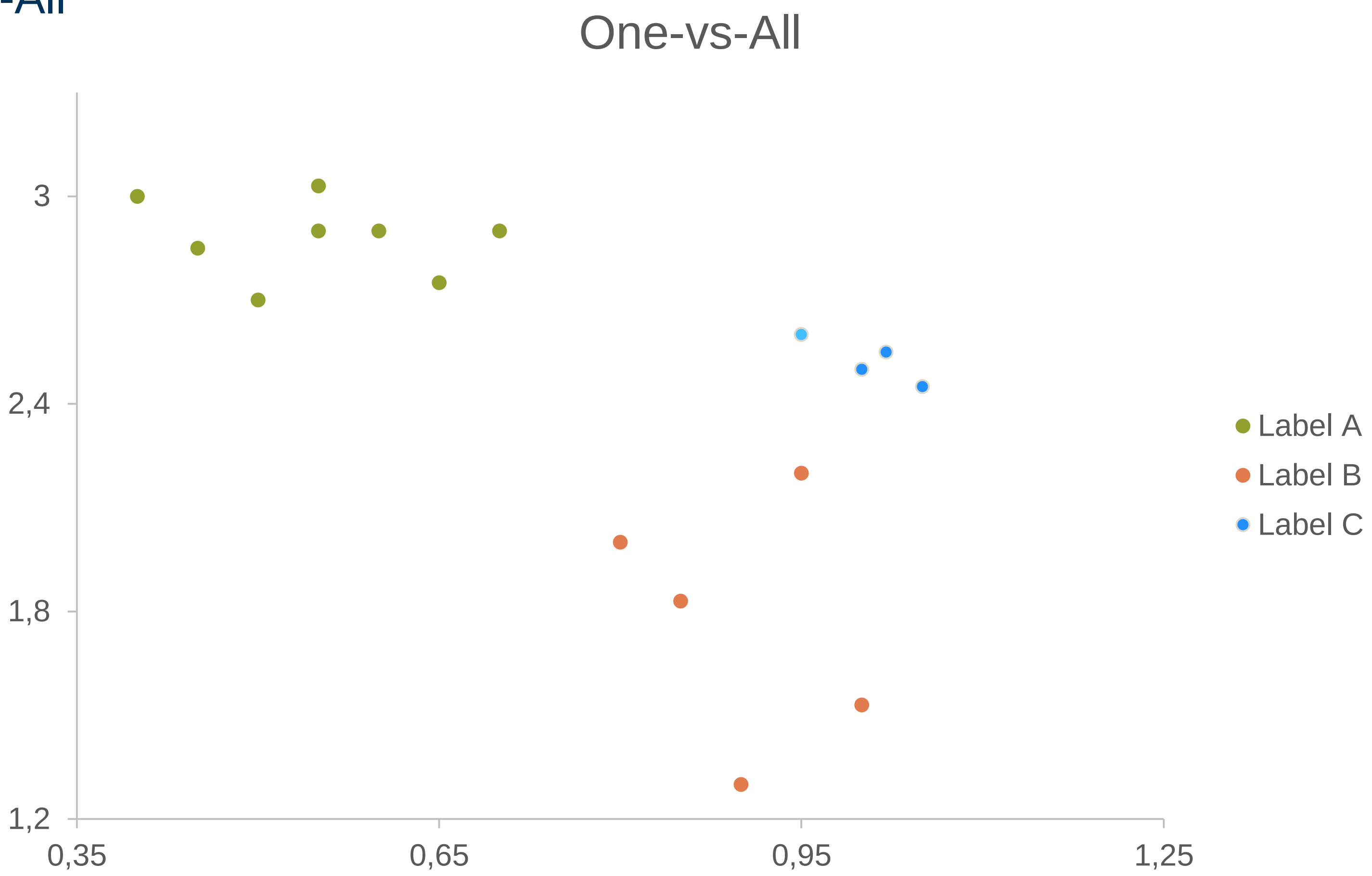
Let‘s say we have a <u>sunny</u>, <u>warm</u> day and it‘s <u>exam-time</u>
P(swim | sun, weather, exams) = P(swim|sun)P(swim|weather) P(swim|exams).
P(swim | sun, weather, exams) = 11*12*2/60 = 288/60 = **4,4**

P(noSwim | sun, weather, exams) = P(noSwim|sun)P(noSwim|weather) P(noSwim|exams).
P(noSwim | sun, weather, exams) = 5*4*10/28 = 380/41 = **7,14**

- P(noSwim) > P(swim)

Automatic Document Classification

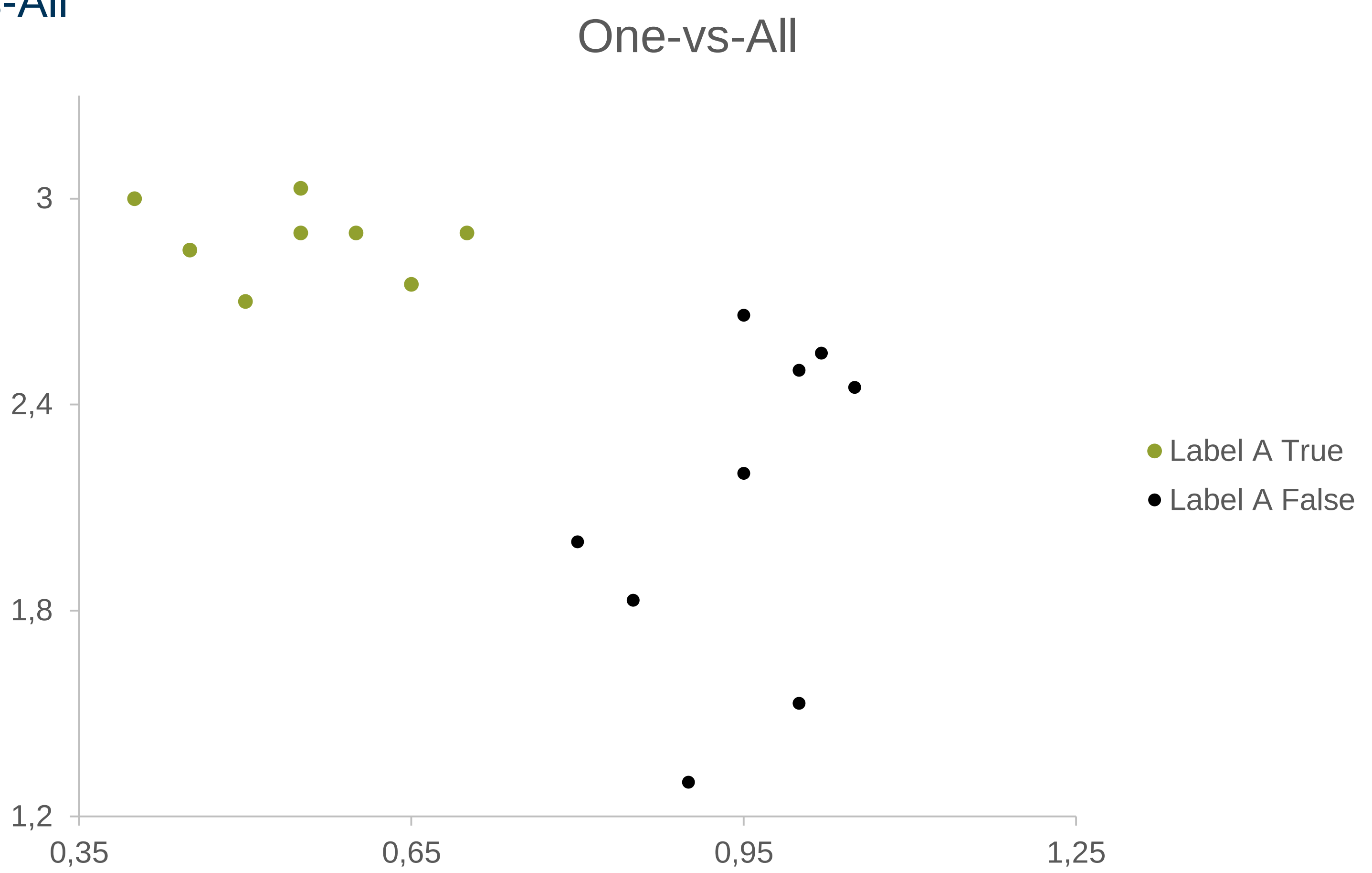
One-vs-All



-> Devide in 3 binary classifications

Automatic Document Classification

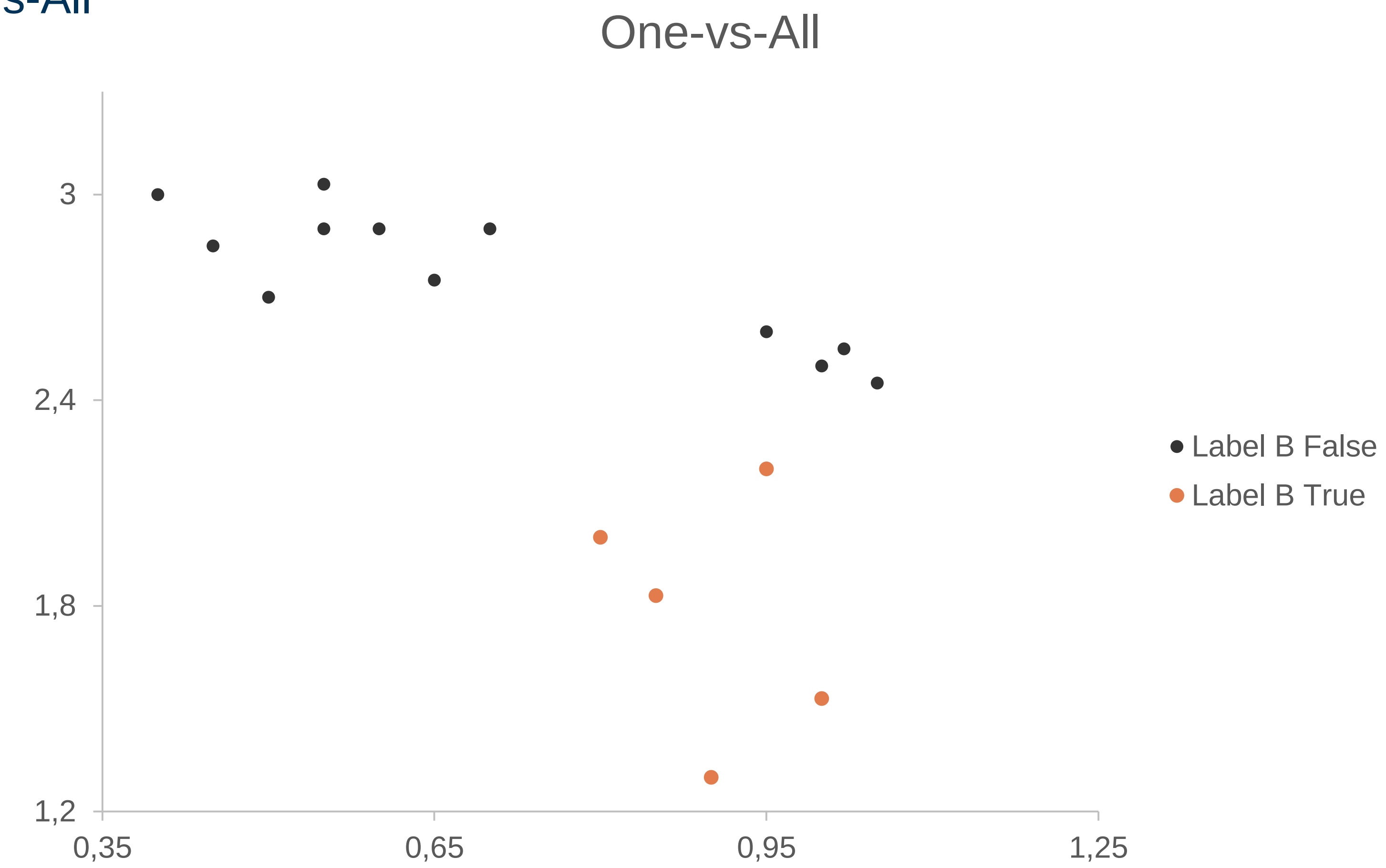
One-vs-All



-> Devide in 3 binary classifications

Automatic Document Classification

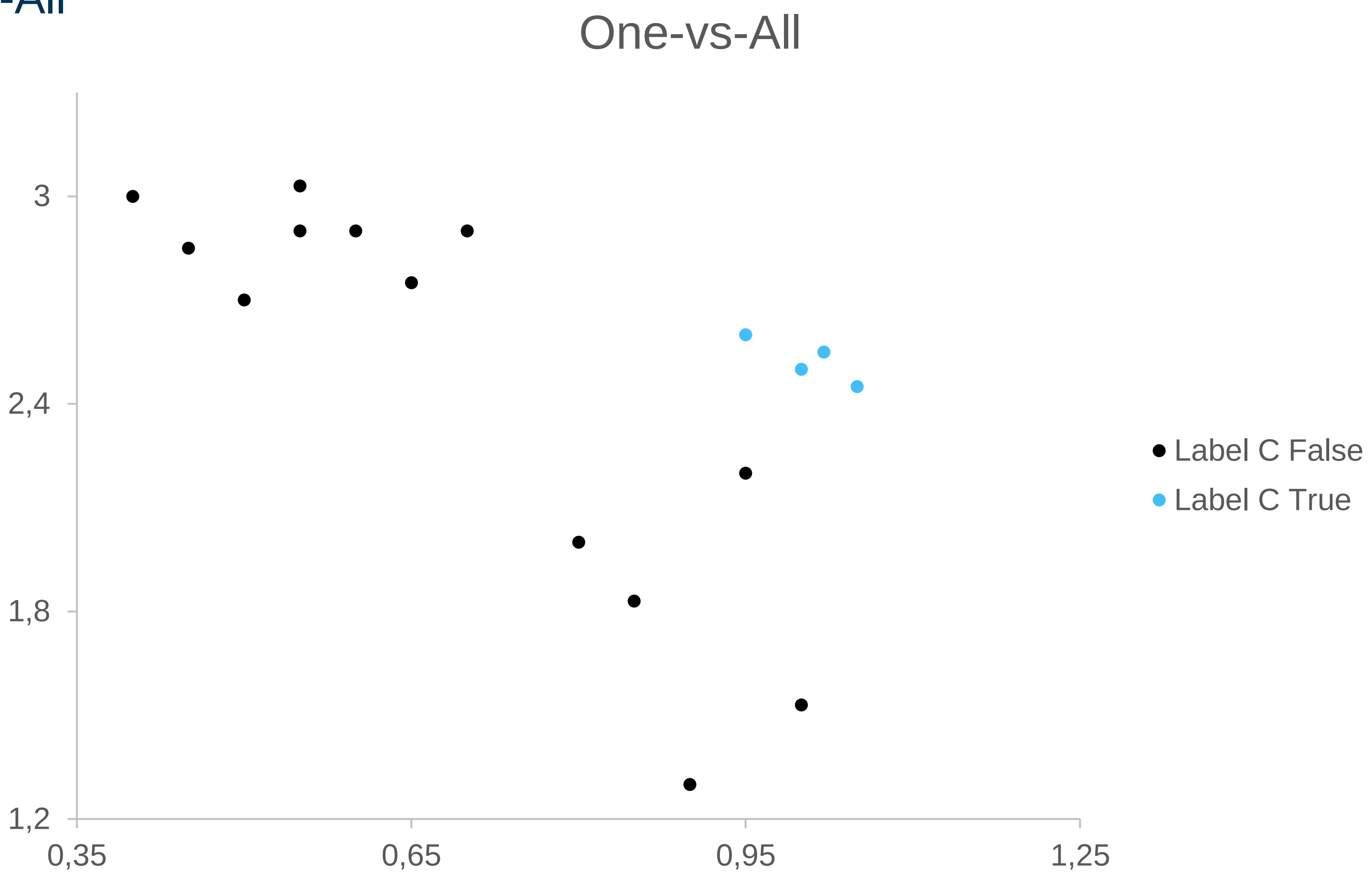
One-vs-All



-> Devide in 3 binary classifications

Automatic Document Classification

One-vs-All



-> Devide in 3 binary classifications