Recognition of invoices from scanned documents

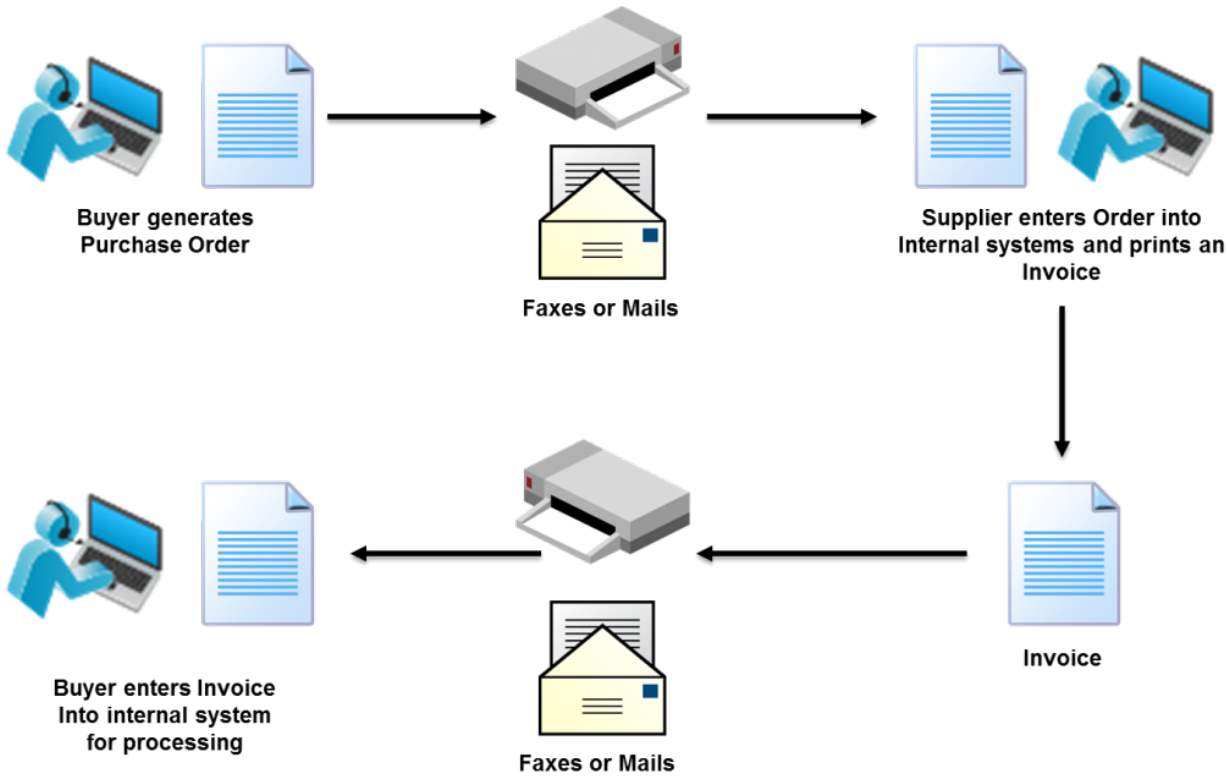
Hien Thi Ha
Natural Language Processing Center
Faculty of Informatics - Masaryk University

December 2, 2017

Outline

- Motivation
- Classification system
- Experiments

Invoice processing systems



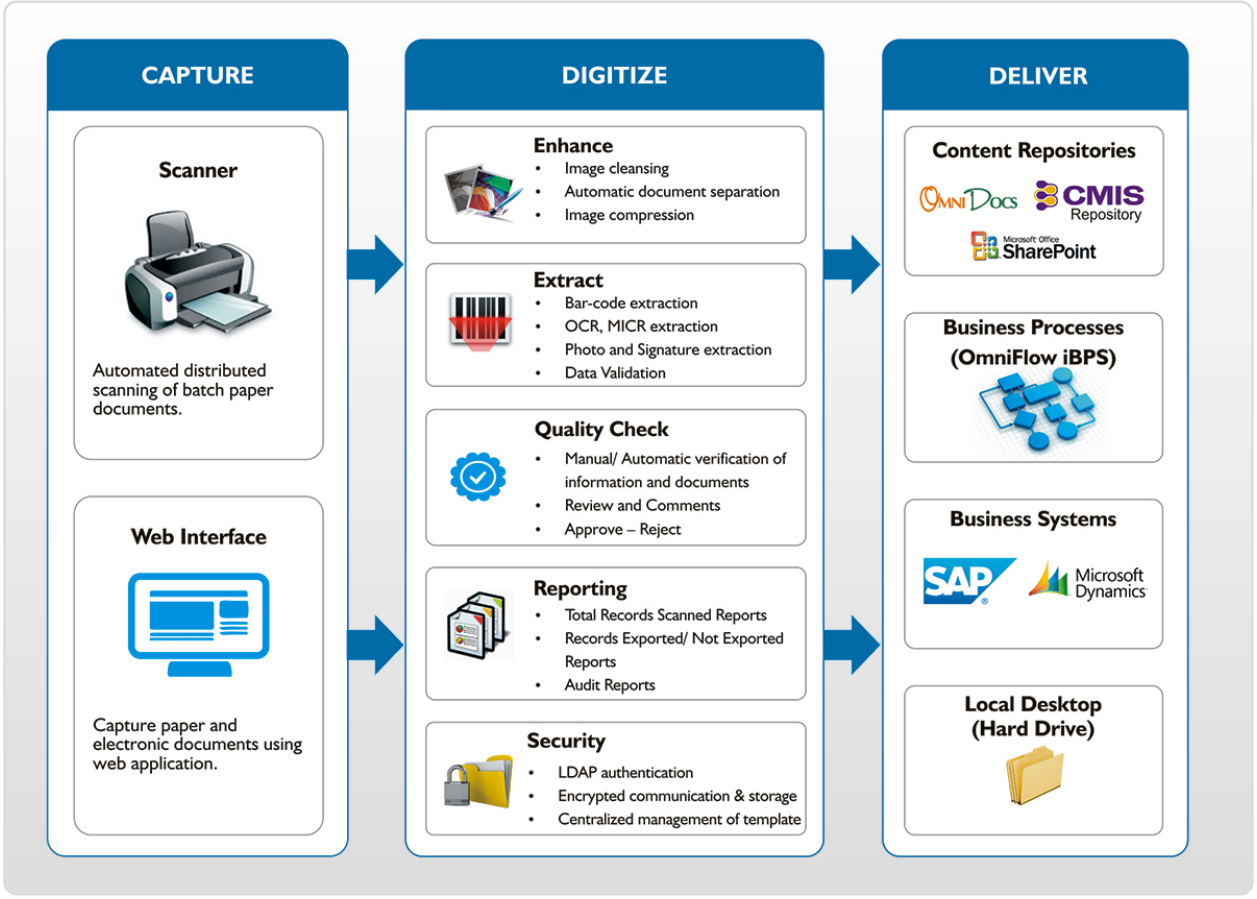
Manual process's challenges

- Large volume
- Variety of layout formats and delivery formats
- time-consuming
- expensive
- errors

Applications

- Document management systems
- Information extraction systems
- ...

Automated distributed scanning of batch paper documents



Invoice definition

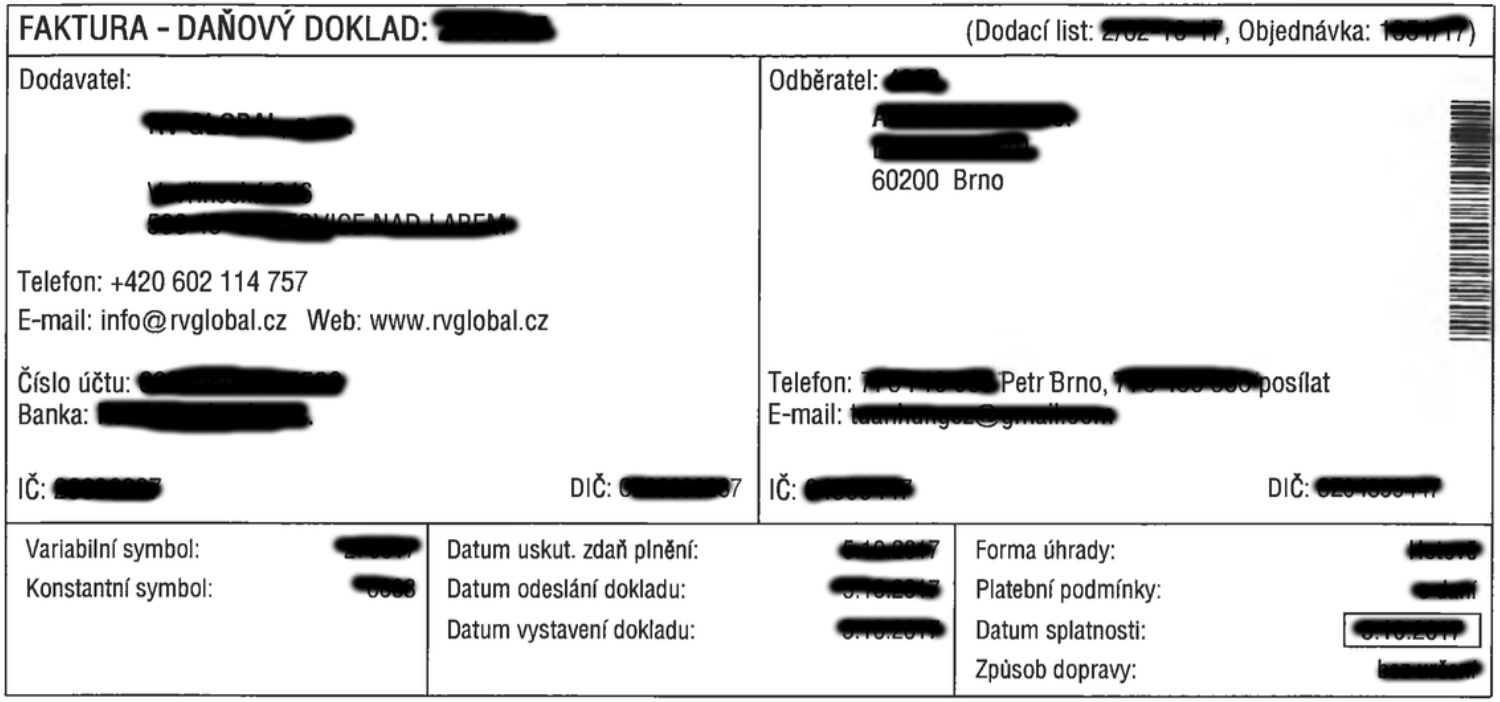
"A nonnegotiable commercial instrument issued by a seller to a buyer. It identifies both the trading parties and lists, describes, and quantifies the items sold, shows the date of shipment and mode of transport, prices and discounts (if any), and delivery and payment terms"

(http://www.businessdictionary.com/definition/invoice.html)

Types of invoice:

- commercial invoice
- consular invoice
- custom invoice
- pro-forma invoice

Examples of invoice



FAKTURA - DAŇOVÝ DOKLAD: [redacted] (Dodací list: [redacted], Objednávka: [redacted])

Dodavatel:

[redacted]

Telefon: +420 602 114 757
E-mail: info@rvglobal.cz Web: www.rvglobal.cz

Číslo účtu: [redacted]
Banka: [redacted]

IČ: [redacted] DIČ: [redacted]

Odběratel: [redacted]

60200 Brno

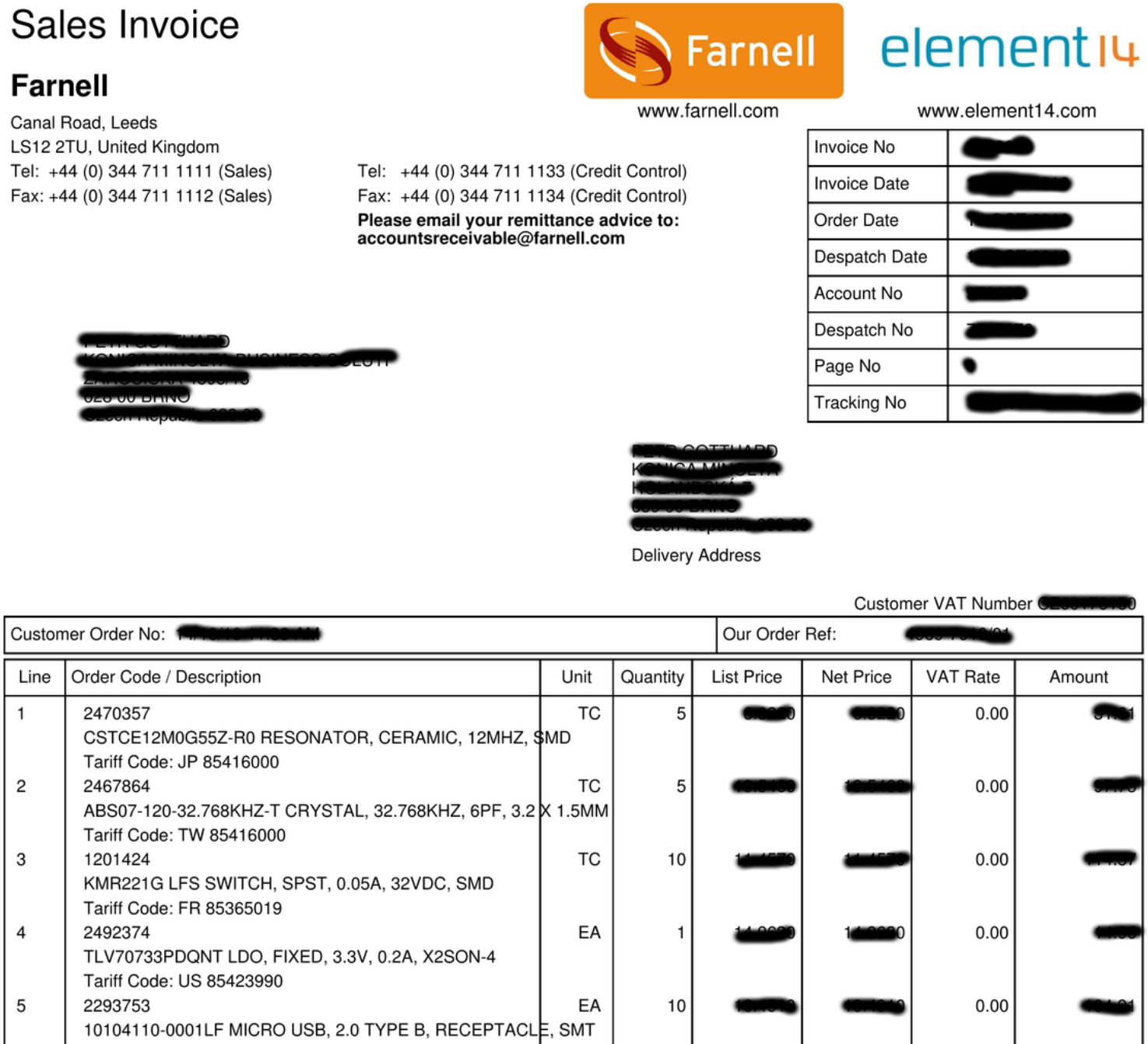
Telefon: [redacted] Petr Brno, [redacted] posílat
E-mail: [redacted]

IČ: [redacted] DIČ: [redacted]

Variabilní symbol: [redacted]	Datum uskut. zdaň plnění: [redacted]	Forma úhrady: [redacted]
Konstantní symbol: [redacted]	Datum odeslání dokladu: [redacted]	Platební podmínky: [redacted]
	Datum vystavení dokladu: [redacted]	Datum splatnosti: [redacted]
		Způsob dopravy: [redacted]

Číslo zboží	Název zboží	Množství	Jedn.	PC bez DPH	PC s DPH	Sazba DPH	DPH Celkem	Celkem s DPH
13240	:D EMOJI sprchový gel s pump. 500ml 3 druhy	48	ks	[redacted]	[redacted]	21%	[redacted]	[redacted]
13242	:D EMOJI šampon+sprchový gel 2v1 300ml 2 druhy	24	ks	[redacted]	[redacted]	21%	[redacted]	[redacted]
13241	:D EMOJI tekuté mýdlo 350ml 3 druhy	24	ks	[redacted]	[redacted]	21%	[redacted]	[redacted]
12761	:D MAZLÍČCI vlhčené ubrousky 63ks s klipem	75	ks	[redacted]	[redacted]	21%	[redacted]	[redacted]
13244	:D MIMONI sprchový gel+šampon 2v1 236ml	120	ks	[redacted]	[redacted]	21%	[redacted]	[redacted]
13105	:D PRINCESS Kráska a zvíře sprchový gel 400ml	36	ks	[redacted]	[redacted]	21%	[redacted]	[redacted]
13160	:D ŠMOULOVÉ tekuté mýdlo 350ml 3 druhy	42	ks	[redacted]	[redacted]	21%	[redacted]	[redacted]
13159	:D ŠMOULOVÉ vlhčené ubrousky 72ks	22	ks	[redacted]	[redacted]	21%	[redacted]	[redacted]
40713	AIRALL Solid osvěžovač 170g APPLE	128	ks	[redacted]	[redacted]	21%	[redacted]	[redacted]
40712	AIRALL Solid osvěžovač 170g BLOSSOM	137	ks	[redacted]	[redacted]	21%	[redacted]	[redacted]
40710	AIRALL Solid osvěžovač 170g BREEZE	133	ks	[redacted]	[redacted]	21%	[redacted]	[redacted]

Examples of invoice



Sales Invoice

Farnell

Canal Road, Leeds
LS12 2TU, United Kingdom
Tel: +44 (0) 344 711 1111 (Sales)
Fax: +44 (0) 344 711 1112 (Sales)

Tel: +44 (0) 344 711 1133 (Credit Control)
Fax: +44 (0) 344 711 1134 (Credit Control)
Please email your remittance advice to:
accountsreceivable@farnell.com

www.farnell.com

www.element14.com

Invoice No	[redacted]
Invoice Date	[redacted]
Order Date	[redacted]
Despatch Date	[redacted]
Account No	[redacted]
Despatch No	[redacted]
Page No	[redacted]
Tracking No	[redacted]

[redacted]

[redacted]

Delivery Address

Customer VAT Number [redacted]

Customer Order No: [redacted]	Our Order Ref: [redacted]

Line	Order Code / Description	Unit	Quantity	List Price	Net Price	VAT Rate	Amount
1	2470357 CSTCE12M0G55Z-R0 RESONATOR, CERAMIC, 12MHZ, SMD Tariff Code: JP 85416000	TC	5	[redacted]	[redacted]	0.00	[redacted]
2	2467864 ABS07-120-32.768KHZ-T CRYSTAL, 32.768KHZ, 6PF, 3.2 X 1.5MM Tariff Code: TW 85416000	TC	5	[redacted]	[redacted]	0.00	[redacted]
3	1201424 KMR221G LFS SWITCH, SPST, 0.05A, 32VDC, SMD Tariff Code: FR 85365019	TC	10	[redacted]	[redacted]	0.00	[redacted]
4	2492374 TLV70733PDQNT LDO, FIXED, 3.3V, 0.2A, X2SON-4 Tariff Code: US 85423990	EA	1	[redacted]	[redacted]	0.00	[redacted]
5	2293753 10104110-0001LF MICRO USB, 2.0 TYPE B, RECEPTACLE, SMT	EA	10	[redacted]	[redacted]	0.00	[redacted]

Examples of invoice

Sales Invoice

Farnell element14

box

INVOICE

[redacted]

PO Number:

Sold To:
[redacted]

Brno
Czech Republic
62800
Czech Republic

Account Number: [redacted]
Invoice Number: [redacted]
PO Number:

Invoice Date: [redacted]
Payment Due By: [redacted]
Payment Terms: [redacted]
Payment Method: [redacted]
Cust VAT No.: [redacted]
Box VAT No.: [redacted]

Bill To:
[redacted]

Brno
Czech Republic
62800
Czech Republic

Service	Unit Price	Quantity	Subtotal	Tax	TOTAL
Personal Pro *Service Period: 04/29/2017-05/28/2017*	$[redacted]	1	$[redacted]	$0.00	$[redacted]

Sterling Equivalent: *Subtotal:* £$ConvertedAmountWithoutTax__c *VAT:* £$ConvertedAmount__c *Conversion Rate: $ExchangeRate__c* ***Total:*** *£$ConvertedTotalAmount__c*

Invoice Subtotal:	$10.00
VAT:	$0.00
Total:	**$[redacted]**
Balance Due:	**$0.00**
Currency:	**USD**

REVERSE CHARGE MECHANISM APPLIES
Under Article 196 of Council Directive 2006/112/EC you may be required to account for any VAT on the services covered by this invoice.

Image converter

- Portable Document Format to Portable Pixmap converter (pdftoppm)
  pdftoppm [options] [PDF-file [PPM-file-prefix]]
- Example:
  pdftoppm -rx300 -ry300 inputfile.pdf outputfile
- Multiple pages document: ...-1.png, ...-2.png,...

OCR tool

Tesseract Open Source OCR Engine (tesseract-ocr)
(See `https://github.com/tesseract-ocr/tesseract/wiki/Documentation`)
tesseract imagename outputbase [-l lang] [-psm pagesegmode] [config...]

- Language setting: list of languages
- Pagesegmentation modes
- Output format: txt, searchable pdf, hocr, tvs

Example:
tesseract filename.png filename -l eng+ces hocr

Tesseract output example

```
<body>
  <div class='ocr_page' id='page_1' title='image "/home/xha1/IE_Experiment/DATA/images/10_2016_konzbul_BICSH-1.png"; bbox 0 0 2481 3509;
ppageno 0'>
   <div class='ocr_carea' id='block_1_1' title="bbox 239 297 2385 306">
    <p class='ocr_par' dir='ltr' id='par_1_1' title="bbox 239 297 2385 306">
     <span class='ocr_line' id='line_1_1' title="bbox 239 297 2385 306; baseline 0 0; x_size 4.5; x_descenders -2.25; x_ascenders
2.25"><span class='ocrx_word' id='word_1_1' title='bbox 239 297 2385 306; x_wconf 95' lang='ces' dir='ltr'> </span>
     </span>
    </p>
   </div>
   <div class='ocr_carea' id='block_1_2' title="bbox 1034 409 1594 497">
    <p class='ocr_par' dir='ltr' id='par_1_2' title="bbox 1034 409 1594 497">
     <span class='ocr_line' id='line_1_2' title="bbox 1034 409 1594 455; baseline 0 -9; x_size 46; x_descenders 9; x_ascenders 10"><span
class='ocrx_word' id='word_1_2' title='bbox 1034 410 1206 446; x_wconf 91' lang='ces' dir='ltr'><strong>Faktura</strong></span> <span
class='ocrx_word' id='word_1_3' title='bbox 1225 429 1238 436; x_wconf 99' lang='ces'>-</span> <span class='ocrx_word' id='word_1_4'
title='bbox 1254 409 1423 455; x_wconf 91' lang='ces' dir='ltr'>daňový</span> <span class='ocrx_word' id='word_1_5' title='bbox 1440 410
1594 446; x_wconf 91' lang='ces' dir='ltr'>doklad</span>
     </span>
     <span class='ocr_line' id='line_1_3' title="bbox 1105 465 1522 497; baseline 0.002 -1; x_size 39.791836; x_descenders 7.7918367;
x_ascenders 9"><span class='ocrx_word' id='word_1_6' title='bbox 1105 466 1246 496; x_wconf 90' lang='ces' dir='ltr'>number</span> <span
class='ocrx_word' id='word_1_7' title='bbox 1258 465 1270 497; x_wconf 98' lang='ces'>/</span> <span class='ocrx_word' id='word_1_8'
title='bbox 1284 465 1378 497; x_wconf 92' lang='ces' dir='ltr'>číslo:</span> <span class='ocrx_word' id='word_1_9' title='bbox 1397 465
1522 497; x_wconf 87' lang='ces'>9/2016</span>
     </span>
    </p>
   </div>
   <div class='ocr_carea' id='block_1_3' title="bbox 239 555 2385 560">
    <p class='ocr_par' dir='ltr' id='par_1_3' title="bbox 239 555 2385 560">
     <span class='ocr_line' id='line_1_4' title="bbox 239 555 2385 560; baseline 0 0; x_size 2.5; x_descenders -1.25; x_ascenders
1.25"><span class='ocrx_word' id='word_1_10' title='bbox 239 555 2385 560; x_wconf 95' lang='ces' dir='ltr'> </span>
     </span>
    </p>
   </div>
   <div class='ocr_carea' id='block_1_4' title="bbox 239 809 2385 814">
    <p class='ocr_par' dir='ltr' id='par_1_4' title="bbox 239 809 2385 814">
     <span class='ocr_line' id='line_1_5' title="bbox 239 809 2385 814; baseline 0 0; x_size 2.5; x_descenders -1.25; x_ascenders
1.25"><span class='ocrx_word' id='word_1_11' title='bbox 239 809 2385 814; x_wconf 95' lang='ces' dir='ltr'> </span>
     </span>
    </p>
   </div>
   <div class='ocr_carea' id='block_1_5' title="bbox 251 566 1494 858">
    <p class='ocr_par' dir='ltr' id='par_1_5' title="bbox 251 566 404 605">
     <span class='ocr_line' id='line_1_6' title="bbox 251 566 404 605; baseline 0 -9; x_size 39; x_descenders 9; x_ascenders 7"><span
class='ocrx_word' id='word_1_12' title='bbox 251 566 404 605; x_wconf 91' lang='ces' dir='ltr'>Vystavil:</span>
```

Language detection

Language detection methods:

- Title ("invoice", "faktura"), continuous pages (-1, -2, -3) (changing version, attachments, ocr errors)
- Keywords
- Language distribution

Result:

Total page	Correct	Error	Accuracy
1546	1505	41	0.97

(13 Polish, 3 German, 1 Italian, 17 English, 7 Czech, no Czech or English invoices)

Invoice features

- Word features (¡150 keywords)
- Title features: "invoice" /"faktura"
  - top
  - left
  - width
  - height
- page number feature:
  "(strana—Strana—str)(\s"—č—:—\.)\d{1,2}"*

Example:
{*"datum":False, "doklad":False, "faktura":True, "dič":False, "cena":False, "symbol":True, "dodavatel":False, "číslo":True, "spol":True, "variabilní":False, "ks":False, "tel":True, "množství":False, "banka":False, "dodací":False, "oddíl":True, "splatnosti":True, ..., "top":123, "width":321, "height":46, "page":1*}

Classification models

- Naive Bayes models
- Support Vector Machine models
- Logistic Regression model

Data set

Business documents: invoice, order confirmation (email), attachments,...

language	num.of.files	num.of.pages	invoice	not invoice
Czech		1105	590	515
English		390	146	257
others		10	10	0
Total	998	1505	746	772

Czech invoice data set

Table: Data set

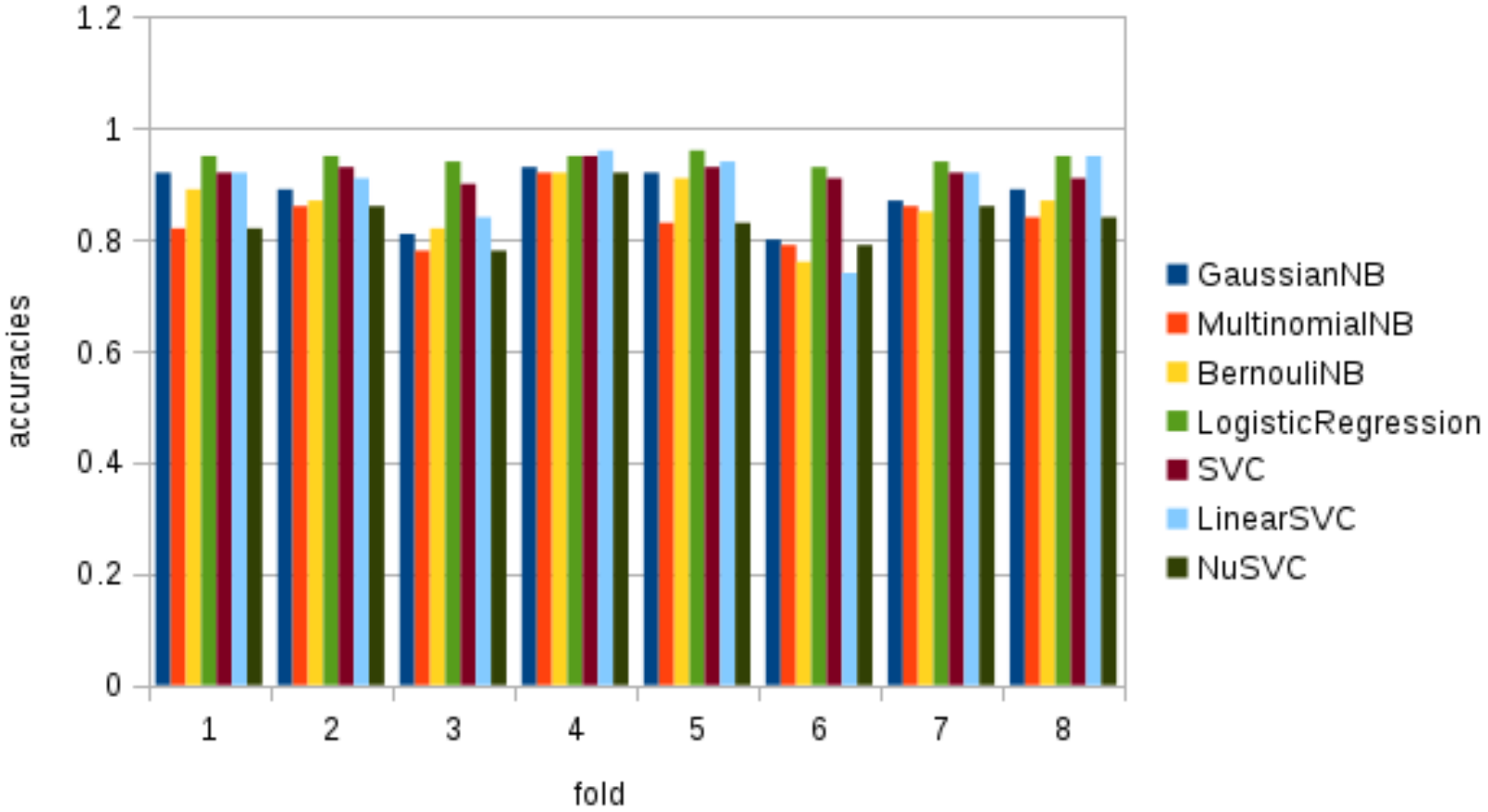
fold	invoice	not invoice	invoice	not invoice
1	528	466	62	49
2	527	467	63	48
3	527	467	63	48
4	524	470	66	45
5	529	465	61	50
6	530	465	60	50
7	534	461	56	54
8	538	457	52	58
9	542	453	48	62
10	531	464	59	51

Accuracies

Table: Accuracies of classifiers

classifiers	1	2	3	4	5	6	7	...	mean
GaussianNB	0.92	0.89	0.81	0.93	0.92	0.87	0.93	...	0.89
MultinomialNB	0.82	0.86	0.78	0.92	0.83	0.76	0.90	...	0.83
BernouliNB	0.89	0.87	0.82	0.92	0.91	0.85	0.92	...	0.87
Log.Regress	0.95	0.95	0.94	0.95	0.96	0.93	0.95	...	0.95
SVC	0.92	0.93	0.90	0.95	0.93	0.89	0.92	...	0.92
LinearSVC	0.92	0.91	0.84	0.96	0.94	0.85	0.93	...	0.92
NuSVC	0.82	0.86	0.78	0.92	0.83	0.76	0.90	...	0.83

result visualization



Logistic Regression results

Table: Precision, recall and F-score of Logistic Regression model

fold	TP	FP	TN	FN	precision	recall	F-score
1	60	3	46	2	0.95	0.97	0.96
2	60	3	45	3	0.95	0.95	0.95
3	59	3	45	4	0.95	0.94	0.94
4	63	2	43	3	0.97	0.95	0.96
5	60	3	47	1	0.95	0.98	0.97
6	54	2	48	6	0.96	0.90	0.93
7	54	3	51	2	0.95	0.96	0.96
8	50	6	52	2	0.89	0.96	0.93
9	43	2	60	5	0.96	0.90	0.92
10	58	4	47	1	0.94	0.98	0.96
average					0.95	0.96	0.95

Modifing features

classifiers	accuracy	precision	recall	F-score
GaussNB				
all features	0.89	0.89	0.90	0.90
keywords only	0.90	0.88	0.94	0.92
keywords+page	0.91	0.89	0.94	0.92
title+page	0.85	0.93	0.78	0.85
Logistic Regression				
all feature	0.95	0.95	0.96	0.95
keywords only	0.94	0.94	0.95	0.94
keywords+page	0.94	0.94	0.95	0.95
title+page	0.84	0.91	0.79	0.84
NuSVC				
all features	0.83	0.89	0.79	0.83
keywords only	0.93	0.91	0.96	0.94
keywords+page	0.93	0.91	0.95	0.94
title+page	0.83	0.89	0.79	0.84

missed classification e.g: FAKTURA - FAgTURA

Faktura daňový doklad **č. 2016021**

Dodavatel
Tomáš Svoboda
Běly Pažoutové 680/4
624 00 Brno
IČ 757 88 039
DIČ CZ7501103896
Nejsem plátce DPH.

Bankovní spojení
205 975 944/0300

Odběratel
Matej Dusik (BICRD)
Konica Minolta Business Solutions Czech, spol. s r.o.
Žarošická 13, 628 00 Brno,

IČ: 001 76 150

Den splatnosti	**29. 6. 2016**
Forma úhrady	převodem
Datum vystavení faktury	22. 6. 2016
Datum uskuteč. zdan. plnění	22. 6. 2016

Při opožděné úhradě účtujeme za každý den úrok z prodlení 0,2 %.

Fakturujeme vám za provedené práce

Grafické úpravy powerpointové prezentace GSC	2 500,00 Kč
Fotografie pro prezentaci 8 ks (350 Kč/ks)	2 800,00 Kč

Conclusion and future work

- Building a classification system for recognition of invoice's first pages from scanned business documents
- Enhance OCR results
- Extract higher layout structures (blocks)