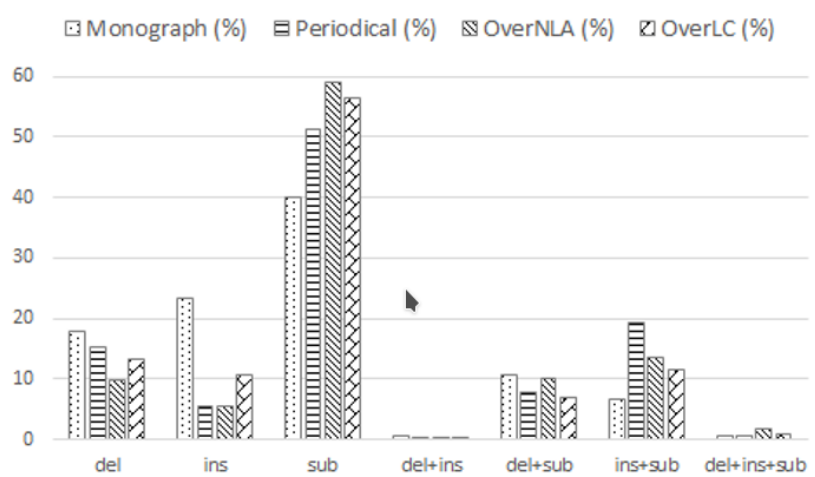
Approximate string matching for detecting keywords in scanned business documents

Thi Hien Ha

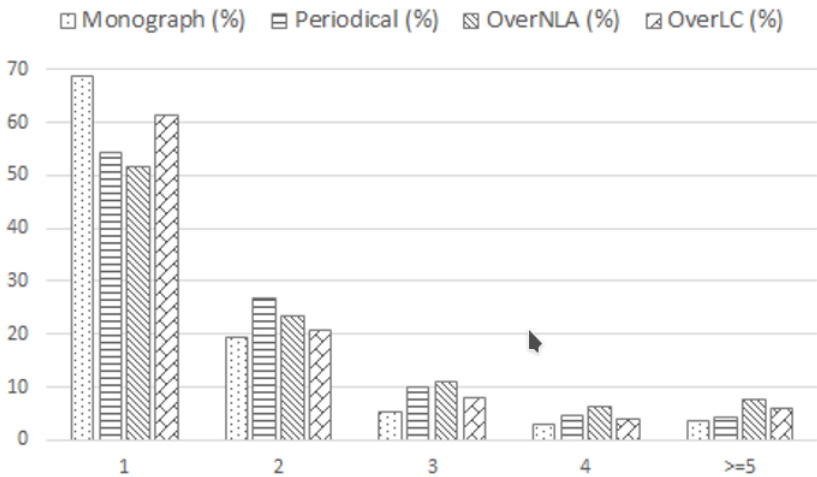
Natural Language Processing Centre - FI - MU

December 6^{th}, 2019

Statistical analysis of OCR errors[1]



Error rates based on operation types



based on edit distances

[1] Nguyen at al.: Deep statistical analysis of ocr errors for effective post-ocr processing. In: 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL). (June 2019) 29–38

OCR errors – Standard mapping

GT Char	Monograph	Periodical	Overproof NLA	Overproof LC
a	{a: 99.5, @: 0.5}	{a: 97.5, u: 0.4, n: 0.2, e: 0.2, i: 0.2, @: 1.5}	{a: 92.7, n: 2.1, i: 1.1, u: 1.0, o: 0.3, m: 0.2, @: 2.6}	{a: 92.8, n: 2.7, u: 0.8, i: 0.5, m: 0.3, o: 0.2, @: 2.7}
b	{b: 98.7, h: 0.8, @: 0.5}	{b: 96.7, h: 1.6, @: 1.7}	{b: 96.2, h: 1.7, l: 0.5, t: 0.3, @: 1.3}	{b: 93.9, h: 1.8, l: 0.5, n: 0.4, i: 0.3, t: 0.3, o: 0.2, m: 0.2, @: 2.4}
c	{c: 97.0, o: 2.0, e: 0.6, @: 0.4}	{c: 96.2, e: 1.2, o: 1.0, @: 1.6}	{c: 93.9, e: 1.7, o: 1.5, r: 0.4, t: 0.2, i: 0.2, @: 2.1}	{c: 92.2, o: 3.1, e: 1.6, u: 0.3, s: 0.3, n: 0.2, a: 0.2, r: 0.2, t: 0.2, @: 1.7}
d	{d: 99.7, @: 0.3}	{d: 98.4, l: 0.2, i: 0.2, @: 1.2}	{d: 97.1, a: 0.4, l: 0.2, i: 0.2, @: 2.1}	{d: 96.8, l: 0.5, i: 0.3, u: 0.3, @: 2.1}
e	{e: 98.7, o: 0.2, @: 1.1}	{e: 96.9, o: 0.6, c: 0.5, a: 0.2, s: 0.2, @: 1.6}	{e: 86.1, o: 9.2, c: 1.6, i: 0.3, a: 0.2, @: 2.6}	{e: 80.8, o: 14.8, c: 0.9, u: 0.4, i: 0.3, r: 0.2, n: 0.2, @: 2.4}
f	{f: 98.2, @: 1.8}	{f: 96.2, t: 1.2, l: 0.9, i: 0.4, @: 1.3}	{f: 94.3, l: 1.5, t: 1.0, i: 0.9, @: 2.3}	{f: 94.1, l: 1.8, t: 1.4, i: 0.6, @: 2.1}
g	{g: 99.6, @: 0.4}	{g: 98.3, @: 1.7}	{g: 93.4, c: 0.4, p: 0.4, r: 0.4, e: 0.3, s: 0.3, i: 0.3, u: 0.3, t: 0.2, f: 0.2, @: 3.8}	{g: 95.2, j: 0.3, i: 0.3, c: 0.2, e: 0.2, @: 3.8}
h	{h: 99.1, b: 0.4, @: 0.5}	{h: 95.2, b: 1.7, i: 0.4, n: 0.2, @: 2.5}	{h: 95.1, b: 1.1, l: 0.8, i: 0.7, n: 0.2, @: 2.1}	{h: 95.7, l: 1.0, i: 0.6, b: 0.5, n: 0.3, @: 1.9}
i	{i: 99.1, @: 0.9}	{i: 97.6, l: 0.6, t: 0.2, @: 1.6}	{i: 90.7, l: 3.3, m: 0.4, t: 0.3, u: 0.2, n: 0.2, @: 4.9}	{i: 94.0, l: 1.6, @: 4.4}
j	{j: 99.7, @: 0.3}	{j: 97.4, i: 0.3, l: 0.3, c: 0.2, @: 1.8}	{j: 85.0, i: 1.5, l: 0.4, t: 0.4, @: 12.7}	{j: 92.7, @: 7.3}
k	{k: 99.5, @: 0.5}	{k: 98.6, t: 0.2, @: 1.2}	{k: 95.6, l: 1.0, i: 0.3, h: 0.2, t: 0.2, @: 2.7}	{k: 97.5, a: 0.2, i: 0.2, h: 0.2, @: 1.9}
l	{l: 95.6, i: 0.8, d: 0.2, @: 3.4}	{l: 96.9, i: 0.8, t: 0.2, @: 2.1}	{l: 96.2, i: 0.8, @: 3.0}	{l: 96.8, i: 0.7, @: 2.5}
m	{m: 99.1, @: 0.9}	{m: 97.4, n: 0.5, i: 0.2, @: 1.9}	{m: 94.3, n: 1.6, i: 0.8, r: 0.5, u: 0.2, @: 2.6}	{m: 93.9, n: 1.3, i: 1.1, u: 0.3, r: 0.2, t: 0.2, @: 3.0}
n	{n: 99.1, u: 0.2, @: 0.7}	{n: 96.4, u: 1.2, a: 0.3, m: 0.2, o: 0.2, i: 0.2, @: 1.5}	{n: 96.2, u: 1.0, i: 0.4, m: 0.3, a: 0.2, @: 1.9}	{n: 92.6, u: 4.0, i: 0.8, m: 0.2, a: 0.2, @: 2.2}
o	{o: 99.4, @: 0.6}	{o: 97.9, e: 0.5, a: 0.2, @: 1.4}	{o: 98.0, n: 0.2, i: 0.2, @: 1.6}	{o: 97.2, n: 0.3, u: 0.3, e: 0.3, @: 1.9}
p	{p: 99.8, @: 0.2}	{p: 98.7, n: 0.2, @: 1.1}	{p: 97.9, n: 0.7, i: 0.2, r: 0.2, @: 1.0}	{p: 96.8, n: 0.5, j: 0.3, o: 0.2, i: 0.2, r: 0.2, @: 1.8}
q	{q: 99.4, @: 0.6}	{q: 97.7, o: 0.2, i: 0.2, j: 0.2, @: 1.7}	{q: 97.3, a: 1.5, o: 0.9, @: 0.3}	{q: 90.7, i: 3.3, m: 2.9, @: 3.1}
r	{r: 99.4, @: 0.6}	{r: 98.5, i: 0.3, t: 0.2, @: 1.0}	{r: 93.4, i: 3.3, l: 0.4, n: 0.3, t: 0.2, @: 2.4}	{r: 98.1, i: 0.2, t: 0.2, @: 1.5}
s	{s: 98.8, a: 0.5, f: 0.3, @: 0.4}	{s: 94.2, a: 0.8, e: 0.7, t: 0.3, i: 0.3, @: 3.7}	{s: 91.7, a: 1.2, i: 0.5, e: 0.3, n: 0.2, b: 0.2, t: 0.2, @: 5.7}	{s: 90.8, t: 0.6, i: 0.5, e: 0.5, a: 0.4, n: 0.3, f: 0.3, u: 0.2, l: 0.2, o: 0.2, h: 0.2, @: 5.8}
t	{t: 99.7, @: 0.3}	{t: 98.7, i: 0.2, l: 0.2, @: 0.9}	{t: 97.7, l: 0.7, i: 0.2, @: 1.4}	{t: 98.0, l: 0.6, i: 0.2, @: 1.2}
u	{u: 99.2, n: 0.2, @: 0.6}	{u: 96.6, n: 1.1, a: 0.7, o: 0.3, i: 0.2, @: 1.1}	{u: 96.1, n: 1.0, i: 0.6, a: 0.3, m: 0.2, @: 1.8}	{u: 96.1, i: 0.7, a: 0.5, o: 0.2, n: 0.2, j: 0.2, @: 2.1}
v	{v: 99.6, @: 0.4}	{v: 97.9, r: 0.7, y: 0.2, @: 1.2}	{v: 92.2, i: 0.8, r: 0.5, y: 0.3, n: 0.3, t: 0.2, @: 5.7}	{v: 97.7, i: 0.3, r: 0.3, m: 0.3, @: 1.4}
w	{w: 99.6, @: 0.4}	{w: 98.7, @: 1.3}	{w: 92.8, v: 1.1, n: 0.5, y: 0.3, m: 0.2, i: 0.2, @: 4.9}	{w: 98.1, v: 0.2, o: 0.2, @: 1.5}
x	{x: 99.0, @: 1.0}	{x: 97.4, i: 0.8, s: 0.2, r: 0.2, t: 0.2, @: 1.2}	{x: 94.6, v: 0.9, i: 0.7, t: 0.6, o: 0.4, n: 0.3, s: 0.2, @: 2.3}	{x: 97.1, g: 1.2, t: 0.6, @: 1.1}
y	{y: 99.5, @: 0.5}	{y: 98.0, v: 1.1, @: 0.9}	{y: 87.9, j: 3.4, v: 3.1, i: 0.4, r: 0.3, s: 0.2, @: 4.7}	{y: 96.9, v: 1.3, j: 0.3, f: 0.2, @: 1.3}
z	{z: 99.2, s: 0.5, @: 0.3}	{z: 86.0, s: 2.5, x: 1.6, r: 1.2, i: 1.1, a: 0.9, g: 0.3, t: 0.3, v: 0.3, c: 0.2, b: 0.2, e: 0.2, k: 0.2, l: 0.2, o: 0.2, n: 0.2, u: 0.2, @: 4.2}	{z: 68.7, r: 6.2, s: 1.9, b: 1.6, n: 1.6, m: 1.5, y: 1.5, i: 0.8, u: 0.7, l: 0.5, @: 15.0}	{z: 98.1, @: 1.9}

OCR errors – Non standard mapping 1:n

GT Char	Monograph	Periodical	Overproof NLA	Overproof NC
a			{ii: 0.05, in: 0.03, -i: 0.02, .i: 0.02}	{ii: 0.21, it: 0.05, in: 0.05, .i: 0.05, iu: 0.03}
b		{li: 0.19, ti: 0.02, th: 0.02, l.: 0.02}		{'h: 0.11, ili: 0.04}
c		{See: 0.03, foe: 0.02}	{t-: 0.05, e-: 0.04, le: 0.03, i': 0.02, .e: 0.02}	{Hle: 0.07, 'C: 0.02, iriw: 0.02}
d			{il: 0.15, tl: 0.05, cl: 0.03, ri: 0.03, t4: 0.02}	{il: 0.15, cl: 0.07, rt: 0.06, tl: 0.05, nl: 0.04}
e			{io: 0.04, lc: 0.02, ic: 0.02}	{io: 0.14, iu: 0.03, no: 0.02, oo: 0.02, n;: 0.02}
f			{'l: 0.03, l': 0.02}	{l': 0.1, l": 0.05, he: 0.02}
g			{iR: 0.09, a-: 0.08, tr: 0.08, fr: 0.07, er: 0.06}	{i": 0.33, e:: 0.21, uu: 0.14, (;: 0.14, ..:.-: 0.13}
h	{li: 0.07}	{li: 0.78, ii: 0.23, il: 0.07, ri: 0.05, ir: 0.04}	{li: 0.3, il: 0.06, ll: 0.05, ji: 0.02, i(.li: 0.02}	{li: 0.21, di: 0.04, Ii: 0.04, 'li: 0.04, ti: 0.03}
i			{vl: 0.03, ll: 0.02}	{ll: 0.04, l': 0.02, '.: 0.02}
j		{.t: 0.08, i.: 0.08}		
k		{lc: 0.06, fc: 0.03}	{lr: 0.12, l,: 0.12, lt: 0.08, fc: 0.06, ',: 0.04}	
l			{ii: 0.02, uit: 0.02, ->: 0.02}	{'.: 0.05}
m	{rn: 0.36, ni: 0.04, in: 0.03}	{in: 0.17, ra: 0.12, rn: 0.09, ni: 0.08, tn: 0.06}	{in: 0.37, rn: 0.29, ni: 0.13, ra: 0.09, tn: 0.08}	{in: 0.65, ni: 0.48, ro: 0.16, rn: 0.15, tn: 0.11}
n		{r.: 0.07, ri: 0.03, ii: 0.03}	{ii: 0.11, ti: 0.03}	{ii: 0.12, ti: 0.11, ri: 0.08, t.: 0.06, iti: 0.03}
o				{in: 0.03, .i: 0.02, i.: 0.02}
p		{ji: 0.03}	{ii: 0.05, iv: 0.03, .i: 0.02}	{fi: 0.1, iiiti: 0.07, ii: 0.03}
q	{cp: 0.03}	{tj: 0.1, .l: 0.05, ri: 0.05, -'t: 0.05}	{.v: 0.03}	
r			{ii: 0.02, i-: 0.02, li: 0.02, i': 0.02}	{ii: 0.04, t': 0.02}
s			{la: 0.03, t,: 0.02, iB: 0.02}	{.-: 0.04, c-: 0.04, nl': 0.04, i': 0.04, ": 0.03}
t				{ln: 0.03, Uo: 0.02}
u		{ti: 0.04, ii: 0.02, tt: 0.02, it: 0.02}	{ii: 0.19, ti: 0.08, li: 0.04, tii: 0.03, i.: 0.02}	{ti: 0.11, ii: 0.1, tl: 0.06, ri: 0.05, i': 0.04}
v			{Ham: 0.09, %': 0.05, s': 0.04, «.: 0.02}	{*: 0.24}
w		{vv: 0.03, vr: 0.02, sr: 0.02}	{vv: 0.44, tv: 0.15, ir: 0.07, *v: 0.05, v»: 0.05}	{st: 0.11, fiH: 0.07}
x	{'~: 0.02}	{ts: 0.03}	{.i: 0.39}	
y			{nj: 0.07, i,: 0.05, ij: 0.05,)*: 0.05, 'j: 0.04}	{tv: 0.04, iiv: 0.04, ino: 0.04, IV: 0.04}
z		{sa: 0.16, .i: 0.16, r.: 0.16, id: 0.16, ti: 0.16}		

OCR errors – Non standard mapping n: 1

OCR Char	Monograph	Periodical	Overproof NLA	Overproof NC
a	{whe: 0.03, we: 0.02, The: 0.02}	{ste: 0.07, ur: 0.05, pe: 0.05, our: 0.04, es: 0.04, co: 0.02, nc: 0.02, us: 0.02, ec: 0.02, et: 0.02, ly: 0.02}	{s.: 1.69, s,: 0.36, ce: 0.09, ut: 0.07, en: 0.03, nd: 0.03, er: 0.03}	{si: 0.55, he: 0.07}
b	{li: 0.03}	{li: 0.08, ch: 0.02, el: 0.02, le: 0.02, th: 0.02}	{hi: 0.07, li: 0.06, is: 0.05}	{si: 0.21}
c	{pe: 0.05}	{el: 0.04, ea: 0.03, pe: 0.02, es: 0.02, rs: 0.02}	{e,: 0.47, le: 0.13, ee: 0.12, ne: 0.12, er: 0.03, ng: 0.02}	{ess: 0.36, es: 0.25, ee: 0.18, se: 0.1, le: 0.02}
d	{il: 2.62, ll: 0.45, ill: 0.02}	{il: 0.1, el: 0.06, ll: 0.04, ct: 0.03, rt: 0.02, al: 0.02, ti: 0.02}	{il: 0.08, ol: 0.03}	{si: 0.24, on: 0.08}
e	{ho: 0.04, oun: 0.02}	{io: 0.02}	{s.: 0.12, ic: 0.07, ol: 0.03}	{can: 1.31, ic: 0.57, ac: 0.43, his: 0.27, ct: 0.02}
f		{wa: 0.02}	{ta: 0.13}	
h	{li: 0.28, la: 0.02}	{li: 0.08, la: 0.06, is: 0.04, le: 0.04, si: 0.02}	{ly: 0.38, li: 0.26, s.: 0.04}	{ld: 0.12}
i	{wa: 0.02}	{ac: 0.03, ll: 0.03, ea: 0.03, ec: 0.02, ra: 0.02, pe: 0.02, ho: 0.02}	{r.: 3.46, s.: 0.39, nce: 0.38, al: 0.05, as: 0.05, st: 0.04, ha: 0.04, er: 0.04, nd: 0.03, at: 0.02}	{ta: 0.26, on: 0.05}
j		{ie: 0.03, ee: 0.02, la: 0.02}	{or: 0.05}	
k		{le: 0.04, ic: 0.03, io: 0.03, is: 0.02}	{ly: 0.22}	
l		{ir: 0.03, ie: 0.03, is: 0.02}	{ni: 0.26, ri: 0.19, si: 0.18, di: 0.14, r.: 0.07, st: 0.02}	{ir: 1.02, ai: 0.6, ot: 0.21, in: 0.12, re: 0.06}
m	{in: 0.07, ste: 0.03, ra: 0.02}	{us: 0.15, ns: 0.11, un: 0.1, in: 0.1, res: 0.08, nt: 0.08, ur: 0.05, ss: 0.05, ver: 0.04, ee: 0.04, ra: 0.04, ne: 0.03, rs: 0.02, an: 0.02, ar: 0.02, re: 0.02, si: 0.02, io: 0.02, co: 0.02}	{n,: 0.43, ur: 0.3, ni: 0.29, in: 0.29, ia: 0.25, ns: 0.16, ai: 0.15, ree: 0.12, as: 0.07, rs: 0.05, or: 0.04, ou: 0.03, an: 0.03, on: 0.02, ra: 0.02, he: 0.02}	{ld: 0.55, ns: 0.42, ll: 0.21, nt: 0.11, es: 0.09, ee: 0.08, on: 0.05}
n	{ri: 0.22, ll: 0.02, ra: 0.02}	{ri: 0.24, rs: 0.14, us: 0.05, wh: 0.05, rt: 0.04, ll: 0.04, il: 0.04, Th: 0.03, ro: 0.03, ss: 0.02, ut: 0.02, re: 0.02, as: 0.02, at: 0.02, li: 0.02, ic: 0.02, is: 0.02}	{ri: 1.61, ry: 0.54, ia: 0.54, am: 0.23, ma: 0.13, ra: 0.13, s.: 0.12, s,: 0.12, st: 0.12, ll: 0.11, ti: 0.1, ay: 0.08, ar: 0.05, at: 0.05, er: 0.04, il: 0.03}	{rs: 0.99, ss: 0.47, om: 0.41, as: 0.28, ar: 0.05, es: 0.03}
o	{el: 0.02}	{el: 0.04, ay: 0.03, ee: 0.02, si: 0.02, se: 0.02}	{e,: 0.75, ic: 0.35, ie: 0.27, nc: 0.25, ne: 0.13, me: 0.12, es: 0.09, ive: 0.07, he: 0.07}	{ee: 0.32, se: 0.3, ll: 0.15, es: 0.14, en: 0.08, re: 0.03}
p			{ve: 0.12, s,: 0.12, ing: 0.1, on: 0.02}	
q			{s.: 0.27}	{ar: 0.1}
r		{ot: 0.02, ve: 0.02, la: 0.02}	{ac: 0.23, ss: 0.12, ee: 0.09, ce: 0.03, he: 0.03}	{me: 0.21, en: 0.18}
s		{ear: 0.06, ta: 0.02}	{e,: 0.14, ng: 0.07, he: 0.03}	{tor: 1.99, an: 0.08}
t		{be: 0.04, il: 0.04, ce: 0.03, ge: 0.03, ie: 0.03, nc: 0.02, si: 0.02, li: 0.02, ra: 0.02}	{ine: 0.6, e,: 0.29, one: 0.22, s.: 0.16, le: 0.13, er: 0.04, nd: 0.03}	
u	{oo: 0.02, we: 0.02, ir: 0.02, il: 0.02}	{ss: 0.12, as: 0.1, ta: 0.08, nde: 0.06, ie: 0.06, il: 0.04, ns: 0.04, is: 0.03, tr: 0.03, ne: 0.03, ri: 0.03, ai: 0.03, rt: 0.02, ec: 0.02, li: 0.02, ee: 0.02, io: 0.02, si: 0.02}	{is: 0.37, ri: 0.28, so: 0.27, ia: 0.25, ll: 0.25, rs: 0.19, as: 0.19, hi: 0.16, ti: 0.13, il: 0.12, ha: 0.11, le: 0.1, in: 0.07, ra: 0.06, st: 0.06, li: 0.06, ee: 0.05, ai: 0.04, re: 0.03, it: 0.02, he: 0.02}	{ns: 0.7, na: 0.65, fo: 0.24, st: 0.2, an: 0.15, ea: 0.14, te: 0.06, is: 0.06}
v		{ai: 0.03}	{ry: 0.04}	
w	{hav: 0.03}	{ss: 0.19, ec: 0.05, se: 0.05, ar: 0.04, tr: 0.03, ea: 0.03, co: 0.02, ve: 0.02, un: 0.02, ur: 0.02, ee: 0.02, si: 0.02, fo: 0.02, ta: 0.02}	{si: 0.11, or: 0.02}	{ear: 1.08, se: 0.29}

Potential errors in Czech scanned documents

- OCR errors:
  - Characters mapping
  - Word boundary
- Language characteristics
  - Diacritical orthographic
  - Inflectional grammar

Approximate string matching

Let Σ be a finite alphabet; $|\Sigma| = \sigma$; Σ^* is the set of all strings over Σ
Let $T \in \Sigma^*$ be a text of length n; $|T| = n$
Let $P \in \Sigma^*$ be a pattern of length m; $|P| = m$
Let d: $\Sigma^* \times \Sigma^* \rightarrow \Re$ be a distance function.
Let $k \in \Re$ be the maximum number of error allowed.
The problem is **given T, P, k and d(.), return the set of all the substrings of T:** $T_{i..j}$ **such that** $d(T_{i..j}, P) \leq$ **k.**

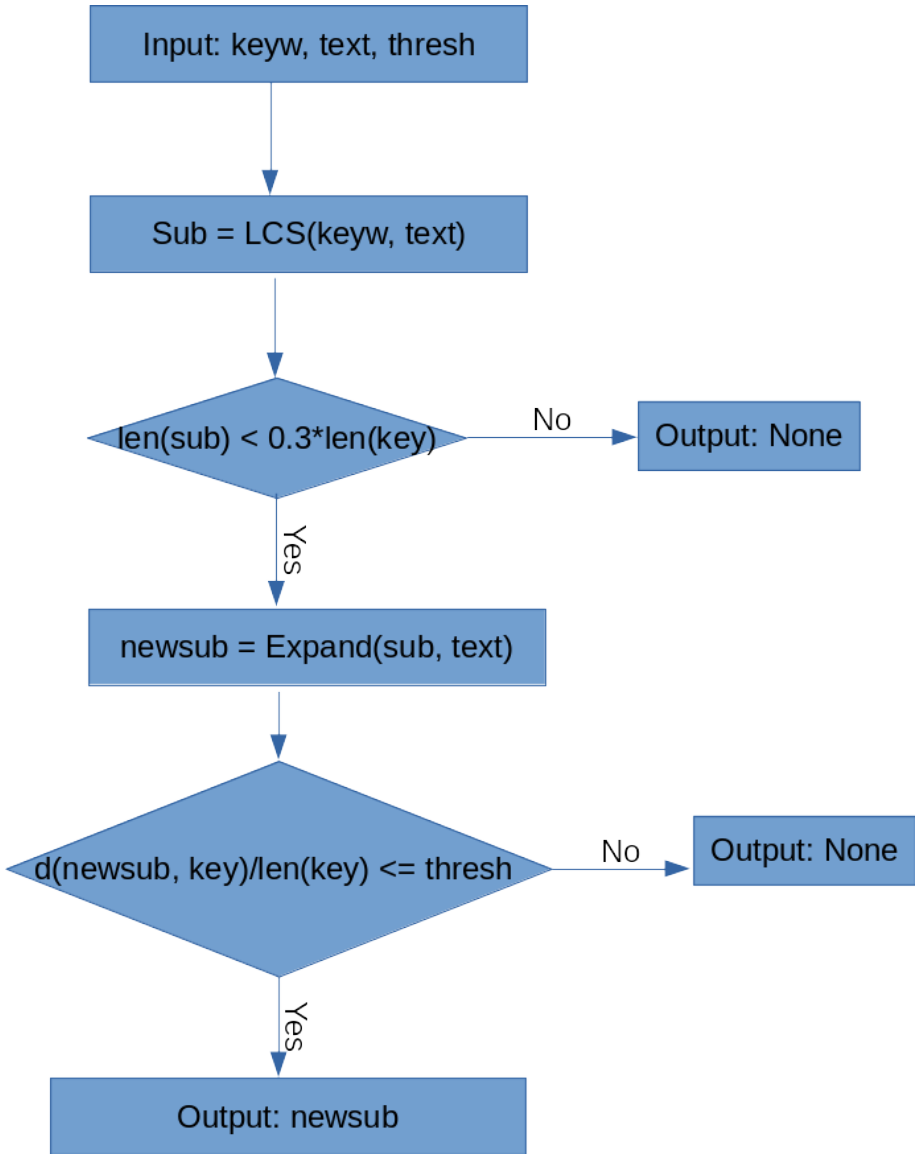
Operations

- *Insertion*: $\delta(\varepsilon, a)$: inserting the letter a
- *Deletion*: $\delta(a, \varepsilon)$: deleting the letter a
- *Substitution*: $\delta(a, b)$; $a\#b$: substituting a by b
- *Transposition*: $\delta(ab, ba)$; $a\#b$: swap the adjacent letters a and b.

Weighted edit distance

$$\delta(a,b) = \begin{cases} 0 & \text{if } a = b \\ 0.1 & \text{if } \left[\begin{array}{l} (a = \varepsilon \text{ and b is punctuation}) \\ \text{or } (b = \varepsilon \text{ and a is punctuation}) \\ \text{or } (a,b) \text{ is a common pair of OCR errors} \end{array}\right. \\ 1 & \text{otherwise} \end{cases}$$

Algorithm


Input: keyw, text, thresh
Sub = LCS(keyw, text)
len(sub) < 0.3*len(key)
No
Output: None
Yes
newsub = Expand(sub, text)
d(newsub, key)/len(key) <= thresh
No
Output: None
Yes
Output: newsub

Experiment

- Dataset: 50 Czech invoices
- Threshold: 0.15
- Common OCR errors:

Char 1	Char 2	Char 1	Char 2	Char 1	Char 2
b	h	n	r	a	á
c	o	o	0	i	í
c	(	r	i	á	í
f	l	r	t	z	ž
f	t	s	5	c	č
i	|	v	y	e	ě
i	1	v	u	e	é
i	l	z	2	s	š
l	|	l	í	u	u
l	1	y	g	u	ú
l	ť	m	n	n	ň

Result

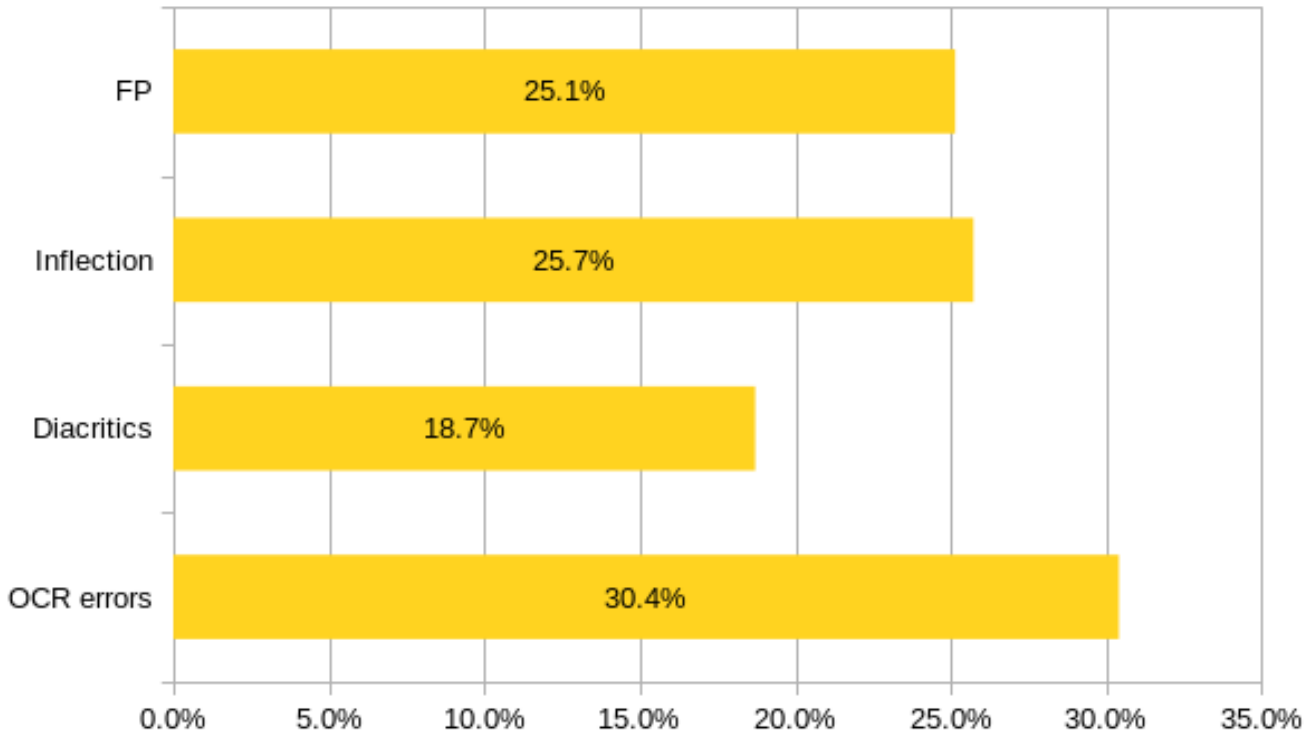


Figure 1: Analysis of keywords detected by approximate string matching but missing by regex

Result

There are 20 keywords are still missing:

- standard mapping (9/20): e.g "o"-"g", "l"-"", "t" - "|"
- non standard mapping (11/20): e.g "čt"-"d", "j"-"f""

Conclusion and future work

- The method adapts pretty well to erroneous text and inflectional language such as Czech
- Implementation based on finite state automation