Západočeská univerzita v Plzni
Fakulta aplikovaných věd
Katedra informatiky a výpočetní techniky

# Bakalářská práce

# Rozpoznávání pojmenovaných entit v právních textech

Plzeň 2013 David Steinberger

# Prohlášení

Prohlašuji, že jsem bakalářskou práci vypracoval samostatně a výhradně s použitím citovaných pramenů.

V Plzni dne 10. května 2013

David Steinberger

# Poděkování

Děkuji především vedoucímu bakalářské práce Ing. Miloslavu Konopíkovi, Ph.D. za výborné vedení práce, Ing. Michalu Konkolovi za konzultace a poskytnutí knihovny strojového učení a také zadavateli JUDr. Jakubu Svobodovi, Ph.D. za konzultace v oblasti práv.

# Abstract

Tato práce se zabývá rozpoznáváním pojmenovaných entit v právních textech pomocí pravidlových i statistických metod. Pravidlové metody jsou použity k rozpoznávání označení zákonných norem a judikátů a dosahují průměrně 87% úspěšnosti. Statistické metody, které jsou založeny na strojovém učení, jsou použity k rozpoznávání předělu oddělující vyrozumění daného soudu od rekapitulace. Bylo dosaženo 45% úspěšnosti přesného určení předělové věty, avšak průměrná poměrná odchylka není větší než 8.32% celého dokumentu. Integrace výsledků této práce do vyhledávacího stroje umožňuje právníkům číst pouze relevantní rozhodnutí nebo jejich důležité části.

The thesis deals with named entity recognition in legal documents based on rule-based and statistical methods. Rule-based methods are used for the recognition of references to statutes and court judgments with an average success rate of 87%. Statistical methods based on machine learning are used for the recognition of divisions between recapitulations and most recent court judgments. Results achieve a success rate of up to 45% of the exact match of the dividing sentence with the average error below 8.32% of the entire document. The integration of the results into a search engine enables lawyers to focus on relevant decisions or their important parts only.

# Obsah

# 1 Úvod

Pojem Pojmenovaná Entita (NE[1]) je široce používán při zpracovávání přirozeného jazyka. Poprvé byl definován na konferenci *Message Understanding Conference (MUC) 6*[6] v roce 1995. Na konferenci MUC-6 se NE rozpoznávali do základních 7 kategorií (např. jména, organizace, data). Tyto výrazy jsou velmi důležité, protože většinou hrají v textu klíčovou roli při dalších analýzách. Jestliže bychom měli v textu označeny všechny NE, mnoho úkolů spojených s analýzou textu by bylo snadněji řešitelných a dosahovaly by lepších výsledků.

V následujícím úryvku je uveden příklad několika NE (tučně zvýrazněné). Je zřejmé, že NE označují ve větě prvky nesoucí největší množství informace.

> Téma této práce bylo zadáno firmou **e-Lectum s.r.o.** dne **12.11.2012** v **Plzni**.

Proces rozpoznávání NE (NER[2]) je ovlivňován mnoha faktory, kvůli kterým je nutné NER nástroje přizpůsobovat. Mezi hlavní patří rozdílnosti mezi jazyky, typ obsahu rozpoznávaného textu (prostě sdělovací, vědecký, žurnalistický atd.) nebo oblasti (sport, ekonomika, informatika atd.)[9].

Téma této práce bylo zadáno jako smluvní výzkum Západočeské Univerzitě firmou e-Lectum s.r.o., která se zabývá vývojem aplikací pro usnadnění práce právníků. Jejich cílem je vytvořit vyhledávač judikatury[3], který budou právníci používat při hledání případů podobných těm, které právě řeší. Na základě výsledku jiného rozhodnutí mohou lépe vézt svůj případ.

Tato práce se bude zabývat rozpoznáváním NE v těchto právních textech. Přesněji řečeno, budou se rozpoznávat vyrozumění soudů k jednotlivým případům. V těchto spisech se nachází velké množství odkazovaných zákonů a odkazů na jiné rozhodnutí, které je nutné označit. Díky tomu můžou právníci snáze poznat, o čem dané rozhodnutí vypovídá a snáze najít souvislosti mezi jednotlivými případy.

Tyto dokumenty mají specifickou strukturu, která se vždy opakuje. V první části se nachází informace o případu, dále se pokračuje rekapitulací z nižších soudů a nakonec rozhodnutí daného soudu. Toto rozhodnutí má pro právníky velkou hodnotu. Informace, které se nachází před ním, je možné dohledat v jiných rozhodnutích. Proto je nutné je oddělit od zbytku do-

[1] Z anglického Named Entity.
[2] Z anglického Named Entity Recognition.
[3] Publikované rozhodnutí soudu.

kumentu. V případě, že by se úspěšně povedlo tuto část oddělit, množství procházených dat by se zmenšilo často až na jednu čtvrtinu. Nejdůležitější rozhodnutí pochází od nejvyšších soudů, kam se musí dostat přes ty nižší a právě zde vzniká velké množství dat, která jsou nepotřebná.

# 2 Teorie

## 2.1 Pravidlové NER

Ručně vyrobená pravidla se používají k rozpoznávání entit, u kterých předem víme, jaký mohou mít tvar. Tvarem je myšlena specifická délka nebo kombinace písmen, čísel a interpunkčních znamének. Tyto tvary se dají většinou snadno popsat regulárním výrazem.

K vyhledání jednoduchých entit často tato pravidla stačí. Problém nastává v případě, že potřebujeme poznat entitu podle kontextu nebo je tvar příliš obecný. Pravidlové algoritmy jsou však stále velice rozšířeny. Díky jejich jednoduchosti je možné sledovat průběh rozpoznávání, což vede k snadnějšímu nalezení problémů[1].

## 2.2 Statistické NER

Tento přístup je zpravidla složitější, neboť je založen na strojovém učení. Díky tomu je možné rozpoznávat entity i podle kontextu nebo složitějších příznaků. Nevýhodou však je, že je potřeba mít velké množství označených dat (viz. 2.2.1), která často nejsou k dispozici.

### 2.2.1 Strojové učení

Strojové učení je odvětvím umělé inteligence, které se zabývá studiem učení systému podle dat. Podle předem daných dat se program naučí rozpoznávat data nová. V případě NER se podle označených entit v textu naučí rozpoznávat nové entity. Základem těchto algoritmů je *klasifikátor*, který je nutné natrénovat. Tyto klasifikátory se dělí podle způsobu trénování a typu trénovacích dat:

- Učení s učitelem - Současná nejpoužívanější technika[1]. Podle označených trénovacích dat klasifikátor určuje výstup. V těchto trénovacích

[1]Tato práce se zabývá pouze tímto typem klasifikátorů.

datech musí být označeny všechny entity, které mají být rozpoznány. V tomto případě se učí klasifikátor pouze z trénovacích dat a dílčí rozpoznávání jeho znalosti nerozšiřuje. Do této třídy spadají techniky jako Maximum Entropy[3] (viz. 2.2.2), Conditional Random Fields[10] (viz. 2.2.3) nebo Decision Trees[15] (viz. 2.2.4).

- Částečné učení s učitelem - V tomto případě algoritmus nepotřebuje tolik rozsáhlá trénovací data. Jako základ se použije malý počet příkladů, takzv. *semínka*. Systém tato slova hledá ve větách a zároveň hledá slova podobná podle kontextu. Tato nalezená slova přidá mezi semínka a hledání opakuje.

- Učení bez učitele - Typický přístup k učení bez učitele je seskupování. Například seskupování entit do skupin podle kontextu. Existují i jiné metody učení bez učitele. U tohoto způsobu nemusí být trénovací data žádná, avšak vnitřní struktura klasifikátoru po natrénování bývá často pro člověka nesrozumitelná.

Klasifikátor potřebuje mít k natrénování definované *příznaky*, podle kterých je možné NE rozpoznat. Příznak je matematické vyjádření určitého jevu, který pomáhá rozpoznat danou entitu. Klasifikátor těmto jednotlivým příznakům přiřadí různé váhy podle toho, jak jsou užitečné. Při rozpoznávání je schopen o určitém slovu říci, o jakou entitu se s největší pravděpodobností jedná.

Například kdybychom chtěli rozpoznávat názvy firem, jako příznaky bychom mohli určit:

- Velká písmena na začátku slov - Firmy začínají velkým písmenem a často obsahují jména vlastníků nebo název města, ve kterém sídlí.

- Předcházející slovo - Před názvem firmy se často objevují slova jako firma, společnost atd.

- Následující slovo - Po názvu firmy se můžou opakovat slova jako kupuje, vlastní, atd.

- Zkratky - Firmy často obsahují v názvu zkratky jako s.r.o., spol., atd.

### 2.2.2 Maximum Entropy

Maximum Entropy (ME) sestavuje model, který vyhoví všem určeným omezením a nezakládá se na žádných jiných předpokladech. Pro určení omezení definujeme příznaky. Obvykle se používají binární příznaky, ale obecně je možná libovolná nezáporná funkce. Máme-li příznakovou funkci pro entitu typu MÍSTO, $f(x, y)$, která pro $x$ začínající velkým písmenem a $y$ rovno MÍSTO vrací hodnotu 1 a jinak 0. Omezení bude definováno jako rovnost středních hodnot pro daný příznak

$$E_p(f_i(x, y)) = E_{\tilde{p}}(f_i(x, y)),$$

kde $E_{\tilde{p}}(f_i(x, y))$ je střední hodnota příznaku vypočítaná z trénovacích dat a $E_p(f_i(x, y))$ je střední hodnota modelu. Je zaručené, že takový model bude existovat a navíc bude jedinečný. Řídí se pravděpodobností rozdělení maximální věrohodnosti a je v následujícím tvaru[13].

$$p(y|x) = \frac{1}{Z(x)} exp \sum_i \lambda_i f_i(x, y)$$

$$Z(x) = exp \sum_i \lambda_i f_i(x, y)$$

$Z(x)$ je normalizační faktor a zaručuje, že $p(y|x)$ je pravděpodobnostní rozdělení. $\lambda_1 ... \lambda_n$ jsou parametry modelu, které určují váhu daného příznaku. Tyto parametry se liší pro různé trénovací algoritmy, mezi které patří například BFGS (L-BFGS)[12].

### 2.2.3 Conditional Random Fields

Conditional Random Fields (CRF) je neorientovaný grafický model používaný k počítání podmíněných pravděpodobností výstupních uzlů podle vstupních uzlů. Ve speciálním případě, kde výstupní uzly grafického modelu jsou spojeny hranami do lineárního řetězce, CRF tvoří Markovův řetězec, a tudíž může být chápán jako podmíněně-trénovatelný konečný stavový automat (KSA). Podstata tohoto algoritmu je silně založena na ME. ME rozpoznává jednu entitu po druhé na rozdíl od CRF, kde jsou klasifikovány všechny entity najednou.

Nechť $o = \langle o_1, o_2, ... o_T \rangle$ jsou pozorovaná vstupní data, jako je sekvence slov v textu nějakého dokumentu, (hodnoty na $n$ vstupních uzlech grafického modelu). Nechť $s$ je množina stavů KSA a každý je spojen s jedním typem

entity, $l \in L$, (např. firma). Nechť $s = \langle s_1, s_2, ...s_T \rangle$ je sekvence stavů, (hodnoty na $T$ výstupních uzlech). CRF definuje pravděpodobnosti stavů podle vstupů jako

$$P_\Lambda(s|o) = \frac{1}{Z_o} exp(\sum_{t=1}^{T} \sum_{k} \lambda_k f_k(s_{t-1}, s_t, o, t)),$$

kde $Z_o$ je normalizační faktor přes všechny stavy, $f_k(s_{t-1}, s_t, o, t)$ je příznaková funkce s argumenty a $\lambda_k$ je naučená váha pro jednotlivé funkce.

Příznaková funkce může být definovaná například pro většinu možností jako hodnota 0 a hodnota 1 právě když $s_{t-1}$ je stav #1 (což může být OSTATNÍ), $s_t$ je stav #2 (což může být MÍSTO), vstup $o$ na aktuální pozici $t$ je obsažen v seznamu názvů míst. Vyšší hodnoty $\lambda$ zvyšují možnost přechodu KSA. Obecně se příznaková funkce ptá na důležité otázky ohledně vstupní sekvence slov, včetně otázek na předchozí slova, následující slova, souvislosti mezi všemi atd.

### 2.2.4 Decision Trees

Decision Trees (DT) je stromově založený klasifikátor. Listy reprezentují jednotlivé entity a větve reprezentují budoucí rozdělení, které vede k jednotlivým entitám. Během průchodu grafu se mohou cesty rozdělovat, ale i spojovat. Sekvencí otázek se dostáváme do listu, který nám určí, o jakou entitu se jedná. Nalezení nejkratšího DT je těžký optimalizační problém[7].

## 2.3 Předzpracování

Předzpracování je příprava dat pro jejich snadnější zpracování. Data jsou obvykle v počítači neznámém formátu a je nutné je pro něj připravit. Mezi procesy předzpracování patří např. tokenizace, dělení vět, stemming, lemnatizace atd.

### 2.3.1 Tokenizace

Rozpoznávaná data jsou obvykle v podobě dlouhých textů, které se programu jeví jako jeden dlouhý řetězec. Proto je většinou nutné tento řetězec rozdělit na tzv. tokeny. Tyto tokeny jsou nejmenší možný samostatný celek, který

dává smysl. Jako token je možné brát jednotlivá slova, zkratky, čísla atd.

Obvykle se k určení hranice tokenu používají bílé znaky a interpunkce. Problémy nastávají v případě, že se jedná například o zkratky. Zkratky často obsahují několik teček, kvůli kterým je zkratka rozdělena na několik tokenů, které však samostatně nedávají smysl. V těchto případech se používají sofistikovanější algoritmy, které těmto problémům zamezují.

S tokenizací úzce souvisí proces dělení vět (sentence segmentation), který se zabývá rozdělením textu do jednotlivých vět[5].

### 2.3.2 Stemming

Stemming je proces používající se ke zredukování počtu skloňovaných nebo jinak upravených slov. Během tohoto procesu se nechávají ze slova pouze jejich kořeny. Například máme-li slova *policie*, *policejní*, *policista* nebo *policisty*, nesou tato slova stejnou informaci a je možné je brát jako slova stejná. V případě, že bychom vyhledávali policii, chceme, aby se nám zobrazily všechny její výskyty, i když jsou jinak modifikovány.

Problémy nastávají v případě, kdy dvě významově odlišná slova mají stejný kořen. Po stemmingu se tato slova budou považovat za slova stejná. Například slovo *leden* a *led* jsou slova významově naprosto odlišná, ale mají stejný kořen, tudíž se po stemmingu budou považovat za slova totožná.

Existuje několik algoritmů pro stemming, které se liší účinností ořezávání. Mezi nejčastěji používané algoritmy patří ořezávání přípon. Tento algoritmus je založen na základě nadefinovaných pravidel, která od slov ořezávají přípony. Některé příklady pravidel:

- jestliže slovo končí -atech, odstraň -atech
- jestliže slovo končí -ětem, odstraň -ětem
- jestliže slovo končí -ou, odstraň -ou

Tato pravidla se mohou aplikovat na jedno slovo několikrát. Podobně lze uzpůsobit pravidla i na předpony. Tento algoritmus je velice jednoduchý a dosahuje dobrých výsledků. Nevýhodou však je závislost na jazyce. Mezi další algoritmy patří například n-gramová analýza.

Se stemmingem je úzce spojen proces lemnatizace. Lemnatizace na rozdíl od stemmingu pracuje s kontextem a nehledá pouze kořeny slov, ale jejich základní tvary. Například slovo *horší* má základní tvar *špatný*, což stemmer nemůže bez kontextu poznat[4].

## 2.4 Evaluace

Pro NER systémy je evaluace základem jejich vývoje. Systémy jsou obvykle hodnoceny podle toho, jaký je jejich výstup ve srovnání se správným výstupem. Vezmeme v úvahu následující správně označený text[2]:

```
Narozdíl od <ENTITA TYP="OSOBA">Roberta</ENTITA>, <ENTITA
TYP="OSOBA">John Briggs Jr</ENTITA> kontaktoval <ENTITA
TYP="SPOLEČNOST">Wonderfull Stockbrockers Inc</ENTITA> v
<ENTITA TYP="MÍSTO">New Yorku</ENTITA> a nařídil jim prodat
všechen jeho podíl v <ENTITA TYP="SPOLEČNOST">Acme</ENTITA>.
```

Nyní uvažujme, že systém by text označil následovně

```
<ENTITA TYP="MÍSTO">Narozdíl</ENTITA> od Roberta, <ENTITA
TYP="SPOLEČNOST">John Briggs Jr</ENTITA> kontaktoval
Wonderfull <ENTITA TYP="SPOLEČNOST">Stockbrockers</ENTITA>
Inc <ENTITA TYP="OSOBA">v New Yorku</ENTITA> a nařídil jim
prodat všechen jeho podíl v <ENTITA TYP="SPOLEČNOST">
Acme</ENTITA>.
```

Systém provedl pět různých chyb[3](vysvětleny v Tabulka č. 2.1). V tomto případě systém správně rozpoznal jednu entitu: (SPOLEČNOST Acme). Otázkou zůstává, jaké skóre systému přiřadit. V následujících sekcích je ukázáno, jak výsledky řeší jednotlivé typy evaluací.

### 2.4.1 MUC evaluace

Na MUC událostech, je systém hodnocen na dvou osách: na ose, která hodnotí schopnost najít správný typ (TYP) entity a na ose, která hodnotí schopnost najít přesný obsah (TEXT) entity. Správný TYP je započten, když je přiřazen správný typ entity a přitom nezáleží na rozmezí textu entity. Správný

[2]Přeložená ukázka z MUC.

[3]Neformální publikace od Christopher Manning: http://nlpers.blogspot.com/2006/08/doing-named-entity-recognition-dont.html.

| **Správné řešení** | **Výstup systému** | **Chyba** |
|---|---|---|
| Narozdíl | <ENTITA TYP="MÍSTO"> Narozdíl</ENTITA> | Systém rozpoznal entitu, kde žádná není. |
| <ENTITA TYP="OSOBA"> Roberta</ENTITA> | Robert | Entita byla zcela vynechána. |
| <ENTITA TYP="OSOBA"> John Briggs Jr </ENTITA> | <ENTITA TYP="SPOLEČNOST"> John Briggs Jr </ENTITA> | Systém správně rozpoznal rozsah entity, ale uvedl špatný typ. |
| <ENTITA TYP="SPOLEČNOST"> Wonderfull Stockbrockers Inc </ENTITA> | <ENTITA TYP="SPOLEČNOST"> Stockbrockers </ENTITA> | Systém správně rozpoznal typ entity , ale uvedl špatný rozsah. |
| <ENTITA TYP="MÍSTO"> New Yorku</ENTITA> | <ENTITA TYP="OSOBA"> v New Yorku</ENTITA> | Systém špatně označil rozsah i typ entity. |

Tabulka 2.1: Chyby provedené při rozpoznávání

TEXT je započten, jestliže jsou dobře označeny hranice entity a nezáleží na typu. Pro obě hodnoty jsou uchovávány tři hodnoty: počet správných odpovědí (COR), aktuální počet odhadů systému (ACT) a počet možných entit v textu (POS).

Finální MUC skóre je průměrem mikro f-míry[4] (MAF), které je harmonickým průměrem přesnosti a pokrytí spočtené ze všech entit a z obou os. Měření je prováděno na všech typech entit bez rozdílu (chyby i úspěchy se počítají dohromady).

Přesnost se počítá jako COR / ACT a pokrytí jako COR / POS. Pro předchozí příklad je tedy COR = 4, ACT = 10 a POS = 10. Z toho vyplívá, že přesnost je 40%, pokrytí je 40% a MAF je také 40%.

Toto měření má výhodu, že bere v úvahu všechny možné typy chyb z Tabulky 1. Dává částečné body za chyby pouze na jedné ose. Díky tomu, že jsou dvě osy, každý celkový úspěch dává dva body a za největší chyby ztrácí dva body.

### 2.4.2 Přesná evaluace

Systémy jsou porovnávány na základě MAF s přesností, která je procento správně nalezených NE a pokrytí, které je procento všech entit v dokumentu. NE je správně rozpoznaná, právě když je správně typ i rozsah entity.

V minulém příkladu bylo správně 5 entit. Systém jich rozpoznal 5, ale pouze jedna byla správně. Proto přesnost je 20%, pokrytí je 20% a MAF je také 20%.

### 2.4.3 ACE evaluace

ACE používá složitější evaluační metodu, která zahrnuje mechanizmy pro hodnocení různých evaluačních problémů (částečná shoda, špatný typ atd.). Každý typ entity má svoji váhu a přispívá do celkové maximální hodnoty (MAXVAL) finálního skóre (např. jestliže osoba má váhu 1 bod a organizace 0.5 bodu, pak ve finálním skóre je potřeba dvě organizace k vyrovnání jedné osoby).

Finální skóre, nazývané Detekce Entit a Rozpoznávání Hodnot (EDR[5]), je 100% mínus hodnota chyb. Například pro předchozí příklad je EDR skóre

[4] Z anglického micro-averaged f-measure.

[5] Z anglického Entity Detection and Recognition Value.

31.3%. Vypočítáno následovně pomocí ACE parametrů z roku 2004[6]. Každá z pěti entit přispívá do finálního skóre. Pomocí standardních ACE parametrů je MAXVAL pro entity typu osoba 61.54%, organizace mají MAXVAL 30.77% a místa 7.69%. Tyto hodnoty dají dohromady 100%. Pomocí parametrů se pro každý typ entity vypočítá hodnota chyb, která se odečte od finálních 100% a dostaneme finální EDR 31.3%.

ACE evaluace může být nejúčinější evaluační metoda díky modifikovatelné váze parametrů. Avšak díky tomu je také problematická, protože finální skóre je porovnatelné, právě když se použijí stejné váhy. Navíc složité metody nejsou intuitivní a díky nim se stávají analýzy chyb složité.[11]

## 2.5 Cross validace

Cross validace je způsob testování používaný k zjištění úspěšnosti statistických metod. Klasifikátory je nutné testovat na neznámých datech (tzn. dokumentech, které nejsou součástí trénovacích dat). Cross validace se používá v případech, kdy není dostatečně velký počet označkovaných dat pro rozdělení na trénovací a testovací skupiny. Obecně se používá $n$-fold cross validace, při které se $n$-tice dokumentů vynechá z trénovacích dat a použijí se k testování. Tento proces se opakuje pro všechna možná rozdělení dat na trénovací a testovací skupiny. Velikost $n$ se volí podle počtu trénovacích dat. Ve speciálním případě, kdy $n$ je rovno jedné, se mluví o tzv. leave-one-out cross validaci.

Podle hodnoty $n$ se snižuje doba potřebná k testování. Při leave-one-out validaci je nutné pro každý testovaný dokument klasifikátor znovu natrénovat. V případě 10-fold validaci je čas potřebný na testování přibližně desetkrát kratší.

Díky této metodě jsou všechny dokumenty použity zároveň na testování i trénování.[14]

[6] http://www.nist.gov/speech/tests/ace/ace04/index.htm.

# 3 Analýza

V první části kapitoly je popsána struktura dokumentů a poskytnutá trénovací data. V další části jsou popsány entity, které se budou rozpoznávat, jejich ukázky a odůvodnění způsobu jejich klasifikace. Konec kapitoly se zabývá normalizací entit a textu pro jednotlivé metody. Na závěr jsou uvedeny důvody výběru programovacího jazyka.

## 3.1 Struktura dokumentu

Na začátku každého dokumentu se nachází základní metadata o tomto dokumentu:

> Právní věta: Spojit věci ke společnému řízení lze i tehdy, není-li k jejich projednání a rozhodnutí podle rozvrhu práce příslušný tentýž soudce (senát).
> Soud: Krajský soud v Brně
> Datum rozhodnutí: 04/21/2011
> Spisová značka: 18 Co 297/2010
> Typ rozhodnutí: ROZSUDEK
> Heslo: Spojení věcí ke společnému řízení
> Dotčené předpisy: § 112 odst. 1 o. s. ř.
> Kategorie rozhodnutí: A
> Publikováno ve sbírce pod číslem: 118 / 2011

Poté následuje rekapitulace z předešlých rozhodnutí od nižších soudů. V následující ukázce je vidět, že se jedná o rozsudek Okresního soudu v Brně a z metadat víme, že se jedná o rozhodnutí Krajského soudu v Brně, který je samozřejmě nad okresním soudem.

> Rozsudkem ze dne 23. 9. 2010, který byl napaden odvoláním, O k r e s n í s o u d Brno - venkov určil, že žalobkyně J. H. a Ing. M. N. (dříve H.), jsou dědičkami ze zákona po svém otci zůstaviteli Ing. J. J., zemřelém 13.10.2007 (výrok I.). Výrokem II. uložil žalované povinnost zaplatit žalobkyním na náhradě nákladů řízení částku ve výši 39 840 Kč do tří dnů od právní moci tohoto rozsudku k rukám právního zástupce žalobkyň...

Nyní přichází důležitá část dokumentu, což je rozsudek soudu (doposud soudu nejvyššího), od kterého rozhodnutí pochází. V případě, že by se tento případ

dostal k vyššímu soudu, bude shrnutí této části obsaženo v rozhodnutí vyššího soudu. Proto výše zmíněná rekapitulace není příliš významná.

> Krajský soud v Brně, jako soud odvolací (§ 10 odst. 1 zákona č. 99/1963 Sb., občanského soudního řádu, ve znění pozdějších předpisů), po zjištění, že odvolání bylo podáno osobou k tomu oprávněnou – účastníkem řízení (§ 201 o. s. ř.) v zákonem stanovené lhůtě (§ 204 odst. 1 o. s. ř.) a směřuje proti rozsudku, proti němuž je odvolání přípustné (§ 202 a contrario), přezkoumal napadený rozsudek, jakož i řízení, které jeho vydání předcházelo, a dospěl k závěru, že odvolání není důvodné...

## 3.2 Trénovací data

Firma e-Lectum nám dodala trénovací data v podobě označených textů ve formátu *.doc*. Jednotlivé entity jsou označeny různými barvami. Tyto dokumenty bude nutné upravit, aby je bylo možno zpracovat. Nejvhodnější by bylo přeformátovat je do formátu *XML* nebo obyčejného textu a barvy nahradit značkami.

V tabulce č. 3.1 jsou uvedeny počty trénovacích dat. Počet označených dokumentů s entitami typu zákonná norma a jiné rozhodnutí je podstatně menší než u entit typu předěl, jelikož se jich v jednom dokumentu nachází příliš mnoho a je velice časově náročné je označit. Jestliže se entity budou rozpoznávat pomocí pravidlových metod, není nutné mít těchto dat tolik, jako pro metody statistické. V tomto případě se použijí pouze pro kontrolu, na rozdíl od statistických metod, které je potřebují k natrénování klasifikátoru.

> ...Krajský soud v Brně, jako soud odvolací (§ 10 odst. 1 zákona č. 99/1963 Sb., občanského soudního řádu, ve znění pozdějších předpisů), po zjištění, že odvolání bylo podáno osobou k tomu oprávněnou – účastníkem řízení (§ 201 o. s. ř.) v zákonem stanovené lhůtě (§ 204 odst. 1 o. s. ř.) a směřuje proti rozsudku, proti němuž je odvolání přípustné (§ 202 a contrario), přezkoumal napadený rozsudek, jakož i řízení, které jeho vydání předcházelo, a dospěl k závěru, že odvolání není důvodné. Žalobkyně J. H. a Ing. M. N. podaly na základě usnesení Okresního soudu Brno-venkov ze dne 28. 5. 2008, č. j. 21 D 948/2007-100, které bylo potvrzeno usnesením Krajského soudu v Brně ze dne 13. 10. 2008, č. j. 18 Co 326/2008-109, každá samostatně žalobu 10. 11. 2008, přičemž ...

| **Typ entity** | **Počet dokumentů** | **Počet entit** |
|---|---|---|
| Zákonné normy | 13 | 501 |
| Jiné rozhodnutí | 13 | 169 |
| Předěl | 125 | 161 |

Tabulka 3.1: Počty trénovacích dat

## 3.3 Typy entit

V dokumentech se budou rozpoznávat 3 typy entit.

### 3.3.1 Zákonné normy

Zákonná norma je jakýkoliv zmíněný zákon (v trénovacích datech označeny žlutou barvou). Zákony se skládají podle stupně zanoření ze sbírky, paragrafu, odstavce a písmena. Aby zákon měl smysl, je nutné, aby vždy obsahoval hierarchicky vyšší část. Jestliže budeme vědět, že se jedná o paragraf 2, ale nevíme, z jaké sbírky, je tato informace zcela bezcenná. Příklad zákona:

§ 204 odst. 1 písm. c) 99/1963 Sb.

Z ukázky je vidět, že sbírka je vždy úplně vpravo, paragraf vždy nalevo a další zanoření jde od paragrafu směrem doprava. V některých případech dochází ke sdružení zákonů, kdy se zmíní pro jednu sbírku více podúrovní. Tyto zákony je nutné rozdělit na samostatné entity.

Sdružená entita: § 2 odst. 1, odst. 2 písm. c) a d), § 3 513/1991 Sb.
Samostatné entity: § 2 odst. 1 513/1991 Sb.
§ 2 odst. 2 písm. c) 513/1991 Sb.
§ 2 odst. 2 písm. d) 513/1991 Sb.
§ 3 513/1991 Sb.

### 3.3.2 Jiné rozhodnutí

Rozhodnutí mají své specifické a jedinečné označení (v trénovacích datech označeny zelenou barvou). Vždy začínají číslem a spisovou značkou, které určují, o čem dané rozhodnutí pojednává a o jakou právní oblast se jedná (civilní, trestní, správní). Speciálním případem jsou rozhodnutí ústavního soudu, které obsahují místo čísla římskou číslici, označující ze kterého senátu rozhodnutí pochází. Jejich spisová značka je vždy ÚS. Všechna rozhodnutí dále obsahují dvě čísla oddělená lomítkem. Druhé číslo označuje rok, ze kterého rozhodnutí pochází a první číslo je pořadové číslo rozhodnutí v daném roce. V případech, kdy je navíc uvedeno za těmito čísly pomlčkou oddělené další číslo, odkazuje toto číslo přesně na část rozhodnutí, o kterém se mluví. Příklad rozhodnutí:

| |
|---|
| Obecné rozhodnutí: 18 Co 326/2008-109<br>Rozhodnutí ústavního soudu: III.ÚS 1528/11 |

### 3.3.3 Vyrozumění soudu

Vyrozumění soudu je oblast textu na konci dokumentu, kde se nachází vyrozumění daného soudu (viz sekce 3.1). Tato část rozpoznávání je na pomezí mezi NER a dělení textu (text segmentation)[2]. Avšak tato dvě odvětví zpracování textu spolu úzce souvisí a používají podobné metody.

Vyrozumění je odlišeno pomyslným kontextovým předělem, který se musí najít (v trénovacích datech azurová barva). Bohužel, tento předěl, je pro každý typ soudu jiný. Je možné ho definovat slovními spojeními, která se v něm nachází. Nejvíce používané obraty nám firma e-Lectum sepsala (obr. 3.1).

Jako entita je brána věta, která se nachází v předělu. V některých dokumentech je jako předěl označeno více vět za sebou, proto se v tabulce 3.1 vyskytuje více entit než dokumentů. To však neznamená, že by jeden dokument mohl mít více předělů. Ten je vždy pouze jeden. Příklad předělu:

| |
|---|
| Krajský soud se v souladu se zásadou procesní ekonomie nejprve zabýval odvolací námitkou nesprávného posouzení důvodnosti žalovaným vznesené námitky promlčení uplatněného nároku, přičemž tuto námitku shledal důvodnou, a to... |

| Typická spojení |
|---|
| § 237 o.s.ř.; dovolání je / není přípustné; bylo / nebylo shledáno přípustným |
| § 239 o.s.ř.; dovolání je / není přípustné; bylo / nebylo shledáno přípustným |
| § 240 o.s.ř.; dovolání je podáno včas; dovolání je podáno opožděně |
| § 267 tr.ř.; nejvyšší soud přezkoumal podle; zákon porušen byl / nebyl; porušení zákona shledal / neshledal |
| nejvyšší soud přezkoumal zákonnost a odůvodněnost; zákon porušen byl / nebyl; porušení zákona shledal / neshledal |
| § 265a t.ř.; splněny podmínky přípustnosti dovolání; dovolání je / není přípustné |
| § 265c t.ř.; Nejvyšší soud jako soud dovolací zkoumal |
| § 265d t.ř.; podané osobami oprávněnými k podání dovolání |
| § 265e t.ř.; podané ve lhůtě |
| § 265f t.ř.; splňuje náležitosti obsahu dovolání |
| dovolání podala včas oprávněná osoba |
| dovolání je zjevně neopodstatněné |
| trestní kolegium Nejvyššího soudu zaujalo shora uvedené stanovisko |
| při jednání trestního kolegia Nejvyššího soudu převážil názor |
| Nejvyšší soud České republiky jako soud nejblíže společně nadřízený příslušnému soudu |
| § 11 o.s.ř.; Nejvyšší soud České republiky, jemuž byla věc v předložena k rozhodnutí o určení místní příslušnosti |
| § 11 o.s.ř.; Nejvyšší soud určí, který soud věc projedná a rozhodne |
| jsou / nejsou splněny zákonné podmínky pro to, aby věc byla přikázána jinému než příslušnému soudu |
| občanskoprávní a obchodní kolegium Nejvyššího soudu ČR zaujalo následující stanovisko |

Obrázek 3.1: Ukázka typických spojení objevujících se v předělu.

## 3.4 Metody klasifikace

Pro rozpoznávání zákonných norem a jiných rozsudků se pokusíme použít pravidlové metody. Jejich tvary mají velice specifický tvar, který je možné popsat regulárními výrazy. U zákonných norem bude nejspíše problém se sdružováním zákonů u jedné sbírky, ale složitější regulární výraz by na to měl stačit. U jiných rozsudků je to jednodušší, k žádnému shlukování nedochází a jsou vždy ve stejném tvaru. Jestliže by pravidlové metody nestačily, museli bychom přistoupit k statistickým metodám.

V případě vyrozumění soudu už si bez statistických metod nevystačíme, nebot' rozdělení vychází pouze z kontextu a bez rozsáhlé databáze výrazů by pravidlové metody neuspěly. Předěl se pokusíme najít především pomocí zmiňovaných slov uvnitř samotného předělu.

Na základě konzultace s vědeckým pracovníkem Ing. M. Konkolem, zabývajícím se využitím strojového učení v NER, bylo stanoveno, že nejlepších výsledků v oblasti NER dosahuje klasifikátor CRF. Bohužel, tento klasifikátor zpracovává celou sekvenci dat najednou a není možné zjistit přesné pravděpodobnosti jednotlivých prvků sekvence (viz. 2.2.3). V našem případě potřebujeme najít větu, u které je největší pravděpodobnost, že je předě-

lem. V případě použití CRF bychom dostali více kandidátů, avšak nepoznali bychom, který je z nich ten nejlepší. Proto jsme se rozhodli použít klasifikátor ME, který nám toto umožní.

## 3.5 Normalizace

### 3.5.1 Pravidlový NER

V právních textech se u často používaných zákonů zaměňují číselná označení sbírek za jejich slovní označení, a proto je nutné udělat mapování slovních označení na označení číselná. Bohužel se tato slovní označení používají v libovolných zkratkách. Například zákon č. 99/1963 Sb. má slovní označení občanský soudní řád a zkratku o.s.ř. Bez mapování bychom po rozpoznání entit měli tři různé zákony namísto jednoho.

Dále je nutné normalizovat velikosti písmen a počty mezer. V zákonech často dochází ke zdvojení mezer nebo naopak k vynechání některé z nich. K největším rozporům dochází u zkratky Sb., kde občas chybí závěrečná tečka nebo dochází k právě zmiňované záměně velikostí písmen. Proto se při rozpoznávání ignoruje počet mezer a velikost písmen. Až po rozpoznání se tyto entity normují.

### 3.5.2 Statistický NER

Předěl v dokumentech budeme rozpoznávat podle slov, a tak je dobré, abychom se zbavili bezvýznamových slov. Mezi tato slova patří spojky, předložky, podmiňovací způsob slovesa být (by, bychom) atd. Obecně jsou tyto slova nazývána jako stop slova (stopwords).

V případě slov se stejným významem je dobré nahradit je jedním zastupujícím slovem pro všechny. V případě synonym bychom potřebovali rozsáhlý slovník, který však nemáme k dispozici. Ale je možné nahradit čísla a nalezené entity stejnou značkou (např. #NUMBER, #ZAKON, #JINEROZHODNUTI) . Není důležité, jaké číslo se v předělu nachází, ale že se tam nachází. To samé platí v případě zákonů nebo jiných rozhodnutí.

K ještě většímu omezení rozmanitosti termínů se používá tzv. lowercasing, kdy se všechna velká písmena vymění za malá. Tento způsob může být problémem v případě, kdy hledáme vlastní názvy (pro tyto případy se pou-

žívá truecasing[8]). V našem případě nehrají velká písmena žádnou roli, takže je můžeme nahradit.

## 3.6 Programovací jazyk

Pro implementaci projektu byl zvolen programovací jazyk Java. Hlavním důvodem je budoucí uplatnění NER ve vyhledávacím serveru Apache Solr[1], který je také napsán v Javě. Dalším důvodem je dodaná knihovna s prostředky strojového učení od Ing. M. Konkola, která je také napsána v Javě.

[1] http://lucene.apache.org/solr/

# 4 Popis

## 4.1 Pravidlové NER

Tato část práce je implementována pouze pomocí regulárních výrazů, které lze bez zásahu do kódu upravit pomocí konfiguračního souboru.

### 4.1.1 Zákonné normy

Na obrázku č. 4.1 je znázorněn regulární výraz pro nalezení zákonných norem. Největším problémem byly zanořené paragrafy, odstavce a písmena. Pro obsáhnutí největšího množství zákonů je možný mezi jednotlivými slovy libovolný počet bílých znaků, které se pak jednoduše nahradí za jeden.

Červenou barvou je označena část, která rozpoznává jednotlivé paragrafy. Vždy musí obsahovat znak paragrafu a číslo. Písmeno za číslem paragrafu už je pouze volitelné. Při zmiňování několika paragrafů v rámci jedné sbírky jsou tyto paragrafy odděleny čárkou nebo písmenem *a*. Těchto dalších paragrafů může být libovolný počet. Regulární výraz je uzpůsoben i pro označení rozmezí spojkou *až*.

§1 § 1, 2a a 3D § 1 až 3

Zelenou barvou je označena část, která označuje odstavce. Tato část je téměř stejná jako pro paragrafy. V případě písmen, která jsou modrou barvou, je rozdíl pouze ve výběru mezi znakem *)* a */*, které se nachází za písmenem.

Barva fialová zajišťuje případ, kdy je zmíněno pro jednu sbírku více zanoření.

§ 1 odst. 2 písm. a) a § 2, § 3 odst. 1 písm c/

Pro případ, že před číslem nebo slovním označením sbírky je zmíněno, že se jedná o zákon nebo předpis, je černá část regulárního výrazu.

zákona č. 82/1988 Sb. § 32 odst. 1 předpisu č. 120/2001Sb.

```
((§\s*(\d+[a-zA-Z]?\s*([a,]\s*|až\s*
§?\s*\d+[a-zA-Z]?\s*)?)*)?
(odst\.\s*(\d+\s*([a,]\s*|až\s*\d+\s*)?)*)?
(písm\.\s*([a-zA-Z](\)|/)\s*
([a,]\s*|až\s*[a-zA-Z](\)|/)\s*)?)*)?
([a,]\s*)?)*
((zákona|zák\.)\s*č\.\s*|předpisu č\.\s*)?
(\d+/\d+\s*(Sb|sb|SB)\.?
|o\.\s*s\.\s*ř\.?|občanského\s*soudního\s*řádu
|(obč|obc|občanského)\.\s*(zákona|zák\.?)
|\s+OZ|(obch\.?|obchodního|obchod\.)
\s*(Z|zákona|zák\.?)|e\.ř\.|exekučního
řádu|OBZ|soudního\s*řádu\s*správního|správního
\s*řádu\s*soudního|s\.\s*ř\.\s*s\.?|správního
\s*řádu|SŘ|(ins\.|insolvenčního)\s*
(zák\.?|zákona)|(tr\.|trestního)\s*(ř\.?|řádu)
|(tr\.|trestního)\s*(zák\.?|zákona))
```

Obrázek 4.1: Regulární výraz pro zákony

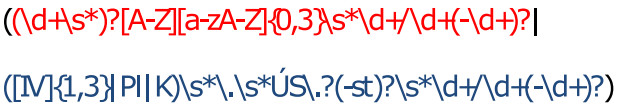
Zbytek regulárního výrazu označuje sbírku. Vybírá mezi azurovou a oranžovou částí. Azurová část reprezentuje číselné označení sbírky a oranžová část slovní označení sbírky. V této oranžové části musí být zmíněny všechny slovní názvy z mapování zákonů, jinak se k mapování nedostanou a budou opomenuty.

Mapování zákonů je také možné modifikovat a přidávat bez zásahu do kódu pomocí konfiguračního souboru. Avšak jak bylo řečeno, je nutné po přidání do mapování přidat dané slovní spojení i do konfiguračního souboru pro rozpoznávání. Nastává zde omezení, kdy je nutné mít seznam zákonů. Ale i kdyby se nám povedlo vymyslet algoritmus, který označí slovně reprezentované sbírky bez seznamu, budou tyto zákony bez mapování pro právníky stejně bezcenné. Není tedy nutné vymýšlet žádný sofistikovanější algoritmus. Příklad mapování:

| |
|---|
| Občanský soudní řád;99/1963 Sb.;<br>(o\.\s*s\.\s*ř\.?\|občanského\s*soudního\s*řádu) |

### 4.1.2 Jiná rozhodnutí

Regulární výraz pro jiná rozhodnutí, jak můžeme vidět na obrázku č. 4.2, je mnohem jednodušší než v případě zákonů. Tyto odkazy mají vždy stejný




Obrázek 4.2: Regulární výraz pro jiná rozhodnutí

tvar.

V případě obecných rozhodnutí (červená barva) může být na začátku číslo, poté je vždy sekvence jednoho až čtyř písmen, kdy první je vždy velké. Na konci jsou dvě čísla oddělená lomítkem a může je následovat ještě jedno číslo oddělené pomlčkou.

V případě rozhodnutí ústavního soudu (modrá barva) je na začátku vždy římská číslice nebo písmena *Pl* nebo *K*. Následují další dvě písmena a tečka *ÚS.*, kde se občas objeví navíc *-st*. Nakonec jsou opět čísla jako u obecných rozhodnutí.

## 4.2 Statistické NER

Tato část práce je založena na strojovém učení a na knihovně od Ing. M. Konkola[1]. Jako klasifikátor je použito ME a rozpoznávaná data jsou v podobě jednotlivých vět. Věta obsahuje informace o své pozici, délce, obsahuje pole slov, která se v ní nacházejí a jednoduché informace o dokumentu (celková délka, typ soudu). Na základě těchto atributů klasifikátor hodnotí, zda se jedná o větu předělovou nebo ne. V následujícím textu jsou popsány jednotlivé typy featur.

### 4.2.1 Obsahové featury

Tyto featury jsou základem celého klasifikátoru. Porovnávají věty na základě obsažených slov. Při trénování se spočtou výskyty jednotlivých slovních n-gramů[2] v předělech. Z těchto spojení se vyřadí ta, která se nacházejí v méně než jedné desetině předělů. Při rozpoznávání se pak určují předěly podle obsažených klíčových slov.

V předělu je v prvních slovech často zmíněn soud, od kterého rozhodnutí

[1] http://liks.fav.zcu.cz/ml/

[2] n-gram je uspořádaná $n$-tice po sobě jdoucích slov.

pochází. Tento soud je zjistitelný z metadat na začátku dokumentu. Jedna z důležitých featur porovnává první slova věty s tímto soudem. Další pak porovnává, jestli je ve větě tento soud obsažen kdekoliv.

Posledním typem featur jsou ty, které zjišťují, jestli věta obsahuje typická spojení poskytnutá zadavatelem. Tato spojení se načtou z konfiguračního souboru. Mohou obsahovat jedno až tři slova. Featury porovnávají, zda věta obsahuje unigramy, bigramy a trigramy těchto spojení.

### 4.2.2 Strukturální featury

Tyto featury porovnávají věty na základě polohy a délky. Porovnávají, v jaké poměrné části se tato věta v dokumentu nachází, např. jestli daná věta leží za půlkou dokumentu. Důležitá je také délka věty. Předělové věty většinou obsahují více slov.

## 4.3 Analyzační modul

Jelikož jsou právnické texty velice specifické a je potřeba je připravit pro klasifikátor, bylo nutné naimplementovat analyzer. Tento analyzer se používá při trénování klasifikátoru, ale také při rozpoznávání. Skládá se z několika komponent: preprocesing, dělení vět, tokenizer, stemmer a filtry.

Nejdříve se pomocí regulárního výrazu najde první výskyt libovolného soudu, tento výskyt je v dokumentu obsažen v metadatech, kde určuje, od kterého soudu toto rozhodnutí pochází. Tento soud je pak přiřazen do atributů každé věty tohoto dokumentu. Poté se pomocí pravidlového NER nahradí slova entitami. Je použit stejný NER jako v 4.1, pouze byly přidány navíc entity pro osoby, data[3] a čísla. V rozhodnutích se jako zvýraznění často používá slovo, kde jsou písmena oddělena mezerami (např. O D Ů V O D N Ě N Í). Z těchto slov se odeberou mezery.

Dělení vět je realizováno pomocí standartní třídy Javy *BreakIterator*. Funkcionalita této třídy není dostačující a bylo nutné rozdělené věty ještě poupravit. Největším problémem byly zkratky, protože tečka ve zkratce způsobí rozdělení na dvě věty. V každé větě se zkontroluje, jestli není na konci zkratka pomocí vytvořeného seznamu zkratek[4]. V případě, že se tam nachází, tak se spojí s následující větou.

[3]Myšleno jako kalendářní údaj.

[4]Tento seznam je pomocí konfiguračního souboru modifikovatelný.

Díky budoucímu uplatnění v Solr se jako tokenizer použil *StandartTokenizer* z Lucene[5] (součást Solr). Tento tokenizer pracuje podle následujících pravidel:

- Rozdělí slova podle znamínek a mezer, které odstraní. Avšak čárka, kterou nenásleduje mezera je považována za součást tokenu.
- Rozdělí slova podle pomlček
- Rozpozná e-mailové adresy a internetové domény

Tyto tokeny se vyfiltrují pomocí následujících filtrů:

- Lowercase filter - Upraví všechna písmena v tokenech na malá.
- Stopwords filter - Odstraní všechny tokeny, které jsou obsaženy v seznamu stop slov.

Na zbylé tokeny se použije stemmer, který nechá ze slov pouze kořeny. Jako stemmer je použit český stemmer z University of Neuchâtel[6].

[5]http://lucene.apache.org/
[6]http://members.unine.ch/jacques.savoy/clef/index.html.

# 5 Testování

## 5.1 Výsledky Pravidlového NER

Pro hodnocení výsledků pravidlového NER byla zvolena přesná evaluace, kde musí být označen přesně jak rozsah tak i pokrytí (viz. 2.4.2). V případě špatně označené meze nebo prohození typu entit, jsou tyto entity velice matoucí a nepoužitelné.

### 5.1.1 Zákonné normy

Výsledky je možné vidět na obrázku č. 5.1. Program dosahuje vysoké přesnosti, ale doplácí na malé pokrytí. To je způsobeno výskytem zákonů, ke kterým nemáme poskytnuté mapování. Tyto zákony jsou v dokumentech označeny, ale pro našeho zadavatele nejsou důležité, proto nám k nim mapování nevypracoval.

Další problém nastává v případě výskytu paragrafů, odstavců a písmen bez označení sbírky, ze které pocházejí. Tyto entity je možné označit, ale není možné k nim jednoznačně přiřadit jejich původ a jejich informace je zcela bezcenná (viz. 3.3.1). Tyto entity jsou však v dokumentech označeny a snižují hodnotu pokrytí.

### 5.1.2 Jiná Rozhodnutí

Výsledky je možné vidět na obrázku č. 5.2. Problémy nastaly v případě, kdy se v dokumentu vyskytovaly u rozhodnutí také zmíněny jejich publikace ve sbírkách rozhodnutí. Tyto sbírky nejsou důležité, protože rozhodnutí je snáze dohledatelné samostatně než v této sbírce. Bohužel tyto sbírky mají velice podobnou strukturu názvu a jsou rozpoznávány jako samostatná rozhodnutí. Tyto sbírky se v datech často neobjevují, proto jsou entity rozpoznávány téměř se 100% úspěšností.

| Soubor | POS | COR | ACT | Přesnost | Pokrytí | MAF |
|---|---|---|---|---|---|---|
| NSS-9 Ans 14-2012 | 52 | 28 | 28 | 100,00% | 53,85% | 70,00% |
| KS-18 Co 297_2010 | 10 | 9 | 9 | 100,00% | 90,00% | 94,74% |
| KS-75 Co 19_2011 | 46 | 39 | 39 | 100,00% | 84,78% | 91,76% |
| NS-4 Nd 50_2008 | 11 | 10 | 10 | 100,00% | 90,91% | 95,24% |
| NS-8 Tdo 562_2011 | 68 | 66 | 66 | 100,00% | 97,06% | 98,51% |
| NS-25 Cdo 4053_2008 | 33 | 28 | 28 | 100,00% | 84,85% | 91,80% |
| NS-29 NSCR 5_2010 | 21 | 13 | 13 | 100,00% | 61,90% | 76,47% |
| NSS-Komp 3-2011 | 72 | 29 | 29 | 100,00% | 40,28% | 57,43% |
| NSS-Pst 32-2011 | 24 | 10 | 10 | 100,00% | 41,67% | 58,82% |
| ÚS-1-3016-11 | 32 | 9 | 9 | 100,00% | 28,13% | 43,90% |
| ÚS4-403-11 | 7 | 6 | 6 | 100,00% | 85,71% | 92,31% |
| ÚS4-2005-09 | 75 | 19 | 19 | 100,00% | 25,33% | 40,43% |
| VS-Ncp 2831_2009 | 50 | 40 | 40 | 100,00% | 80,00% | 88,89% |
| **Celkově:** | 501 | 306 | 306 | 100,00% | 66,50% | 76,95% |

Obrázek 5.1: Výsledky testů pro zákonné normy (ZN).

| Soubor | POS | COR | ACT | Přesnost | Pokrytí | MAF |
|---|---|---|---|---|---|---|
| NSS-9 Ans 14-2012 | 15 | 14 | 14 | 100,00% | 93,33% | 96,55% |
| KS-18 Co 297_2010 | 11 | 11 | 11 | 100,00% | 100,00% | 100,00% |
| KS-75 Co 19_2011 | 23 | 23 | 23 | 100,00% | 100,00% | 100,00% |
| NS-4 Nd 50_2008 | 5 | 5 | 5 | 100,00% | 100,00% | 100,00% |
| NS-8 Tdo 562_2011 | 17 | 17 | 17 | 100,00% | 100,00% | 100,00% |
| NS-25 Cdo 4053_2008 | 7 | 7 | 7 | 100,00% | 100,00% | 100,00% |
| NS-29 NSCR 5_2010 | 11 | 10 | 10 | 100,00% | 90,91% | 95,24% |
| NSS-Komp 3-2011 | 9 | 9 | 10 | 90,00% | 100,00% | 94,74% |
| NSS-Pst 32-2011 | 7 | 6 | 6 | 100,00% | 85,71% | 92,31% |
| ÚS-1-3016-11 | 19 | 19 | 23 | 82,61% | 100,00% | 90,48% |
| ÚS4-403-11 | 6 | 6 | 6 | 100,00% | 100,00% | 100,00% |
| ÚS4-2005-09 | 30 | 30 | 30 | 100,00% | 100,00% | 100,00% |
| VS-Ncp 2831_2009 | 9 | 9 | 9 | 100,00% | 100,00% | 100,00% |
| **Celkově:** | 169 | 166 | 171 | 97,89% | 97,69% | 97,64% |

Obrázek 5.2: Výsledky testů pro jiné rozhodnutí (JR).

## 5.2 Statistické NER

### 5.2.1 Způsob testování

Jelikož trénovacích dat není takové množství, aby se daly rozdělit na trénovací a testovací data, byla použita metoda testování Leave-One-Out cross validace, viz 2.5.

### 5.2.2 Evaluace

Při testování bylo použito několik způsobů evaluace. Jako základ byla použita přesná evaluace (viz. 2.4.2), kde musí věta být určena přesně. Avšak tento způsob je velice striktní, proto byly zavedeny další způsoby evaluace s tolerancí. V tomto případě, vzhledem k rozsahu dokumentů a významu této entity, je velkým úspěchem určení této entity i s odchylkou (vzdáleností od pravého předělu) dvou vět. Význam předělené oblasti zůstane téměř nepoškozený nebo přibude nevýznamná část, která tam nepatří. Za částečný úspěch lze počítat i entity s odchylkou do pěti vět, kdy předělená část nese stále velkou část informace a v opačném případě není přebytečná část stále ještě tak velká.

### 5.2.3 Postprocessing

K dosažení lepších výsledků je po klasifikaci vět proveden postprocesing. Během rozpoznávání byla přiřazována velká pravděpodobnost krátkým větám a větám nacházejících se v metadatech. Tyto věty obsahují velké množství klíčových slov, v některých případech jsou složeny pouze z klíčových slov. Proto jsou tyto věty odfiltrovány. Ukázka jedné z těchto vět nacházející se v metadatech:

Soud: Nejvyšší správní soud.

| **Přesná evaluace** | **Tolerance 2 věty** | **Tolerance 5 vět** | **Průměrná odchylka** | **Průměrná poměrná odchylka** |
|---|---|---|---|---|
| 45% | 58% | 66% | 10 vět | 8.32% |

Tabulka 5.1: Výsledky statistického NER

### 5.2.4 Výsledky

V tabulce č. 5.1 je možné vidět výsledky rozpoznávání předělové věty. Bylo dosaženo výsledků od 45% do 66% podle přesnosti určení věty. Průměrná odchylka nepřesahuje 10 vět. Poměrná odchylka je počítána jako podíl odchylky ku celkovému počtu vět. I hůře rozpoznané věty mohou být přínosem. Stále je velká pravděpodobnost, že údaje, které právník hledá, mohou být obsaženy v odděleném rozhodnutí a nemusí procházet celý dlouhý dokument.

# 6 Závěr

Program rozpoznává entity s průměrnou úspěšností 87%, přičemž zbývající procenta jsou většinou zákony, které nejsou pro zadavatele důležité. Oddělení vyrozumění daného soudu se pohybuje mezi 45% až 66% podle přesnosti určení oblasti. Avšak odchylka průměrně není větší než jedna dvanáctina celého dokumentu.

Budoucí rozšíření by spočívalo ve zlepšení přesnosti rozpoznávání předělu. Lepších výsledků by šlo dosáhnout rozdělením klasifikátorů podle jednotlivých soudů. Klasifikátor by se natrénoval pouze na rozhodnutích, která bude rozpoznávat. Avšak tento přístup by potřeboval větší množství dat. Další možné rozšíření by spočívalo v doplnění mapování slovních označení sbírek, aby byla možná použitelnost mezi všemi právnickými oblastmi.

Práce splňuje všechny body zadání. Po konzultaci se zadavatelem JUDr. Jakubem Svobodou, Ph.D. jsme dospěli k závěru, že práce je pro právnickou oblast velkým přínosem. První část, kde se využívá pravidlových metod, se již používá ve vyhledávacím serveru a má mezi právníky velký ohlas.

# Seznam zkratek

CRF Conditional Random Fields

DT Decision Trees

EDR Detekce entit a rozpoznávání hodnot (Entity Detection and Recognition Value)

KSA Konečný stavový automat

MAF průměrná mikro f-míra (micro-averaged f-measure)

ME Maximum Entropy

MUC Konference Message Understanding (Message Understanding Conference)

NE Pojmenovaná entita (Named Entity)

NER Rozpoznávání pojmenovaných entit (Named Entity Recognition)

# Literatura


[1] Laura Chiticariu, Rajasekar Krishnamurthy, Yunyao Li, Frederick Reiss, and Shivakumar Vaithyanathan. Domain adaptation of rule-based annotators for named-entity recognition tasks. In *In EMNLP (To appear*, 2010.

[2] Freddy Y. Y. Choi. Advances in domain independent linear text segmentation. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, NAACL 2000, pages 26–33, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics.

[3] James R. Curran and Stephen Clark. Language independent ner using a maximum entropy tagger. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 164–167, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[4] Ljiljana Dolamic and Jacques Savoy. Advances in multilingual and multimodal information retrieval. chapter Stemming Approaches for East European Languages, pages 37–44. Springer-Verlag, Berlin, Heidelberg, 2008.

[5] Uwe D.Reichel and Hartmut R. Pfitzinger. Text preprocessing for speech synthesis, 2006.

[6] R. Grishman and B. Sundheim. Message Understanding Conference-6: A Brief History. In *Proceedings of the 16th International Conference on Computational Linguistics*, Copenhagen, June 1996.

[7] Laurent Hyafil and Ronald L. Rivest. Constructing optimal binary decision trees is np-complete. *Inf. Process. Lett.*, 5(1):15–17, 1976.

[8] Lucian Vlad Lita, Abe Ittycheriah, Salim Roukos, and Nanda Kambhatla. truecasing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 152–159, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[9] Diana Maynard, Valentin Tablan, Cristian Ursu, Hamish Cunningham, and Yorick Wilks. Named Entity Recognition from Diverse Text Types. Submitted to Recent Advances in Natural Language Processing 2001 Conference, Tzigov Chark, Bulgaria., 2001.

[10] Andrew McCallum and Wei Li. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 188–191, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[11] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1):3–26, January 2007. Publisher: John Benjamins Publishing Company.

[12] J. Nocedal. Updating quasi-Newton matrices with limited storage. *Mathematics of Computation*, 35:773–782, 1980.

[13] Stephen Della Pietra, Vincent J. Della Pietra, and John D. Lafferty. Inducing features of random fields. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(4):380–393, 1997.

[14] Massimiliano Pontil. Leave-one-out error and stability of learning algorithms with applications. *International Journal of Systems Science*, 2002.

[15] J. R. Quinlan. Induction of decision trees. *Mach. Learn.*, 1(1):81–106, March 1986.

# A Využití

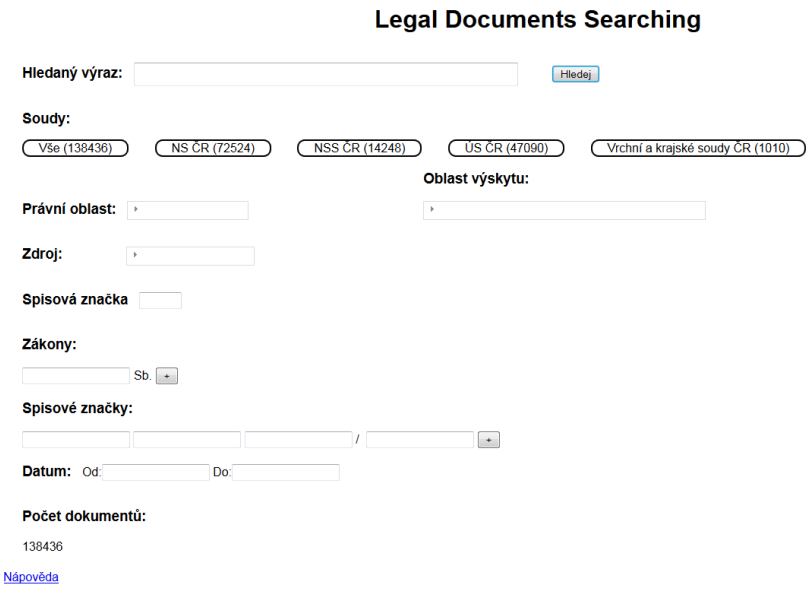
Integrace NER do vyhledávacího serveru Apache Solr a front-end aplikace pro komunikaci s tímto serverem byla součástí smluvního výzkumu, avšak není součástí této bakalářské práce. První část práce, kde se využívá pravidlových metod, je již součástí serveru a je plně funkční.

Entity typu Zákonná norma jsou využity ve vyhledávacím formuláři, kde je možné omezit výsledky vyhledávání pouze na dokumenty obsahující v textu tyto zákony (obrázek č. A.1[1]). Dále jejich využití spočívá ve facetové navigaci (obrázek č. A.2), kde jsou zákony reprezentovány stromovou strukturou a je možné hierarchicky zobrazovat další části sbírky. Po kliknutí na jakoukoliv část se vyhledávání opět omezí na výsledky s touto zákonnou normou.

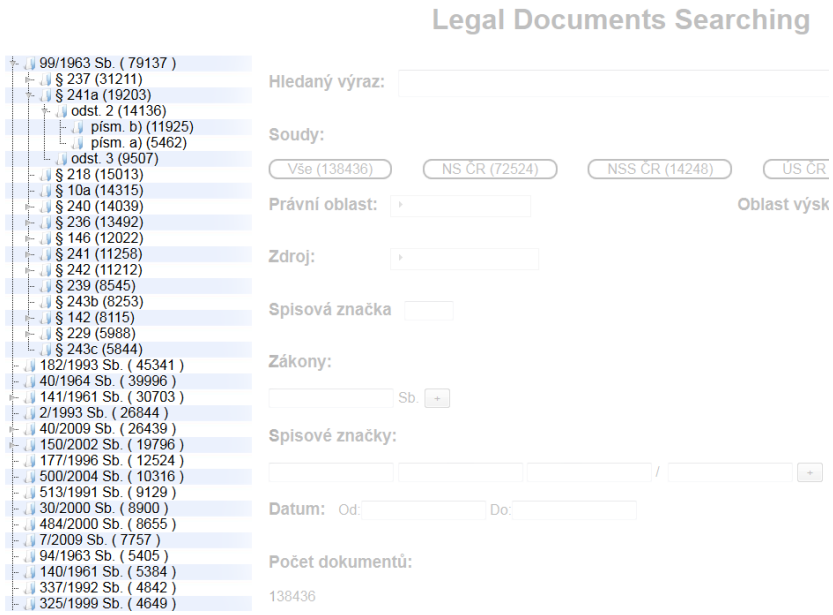
Entity typu Jiné rozhodnutí mají stejné využití jako zákonné normy, ale jsou jen součástí vyhledávacího formuláře.

Všechny nalezené entity je možné vidět při zobrazení vybraného dokumentu (obrázkek č. A.3 a obrázek č A.4).

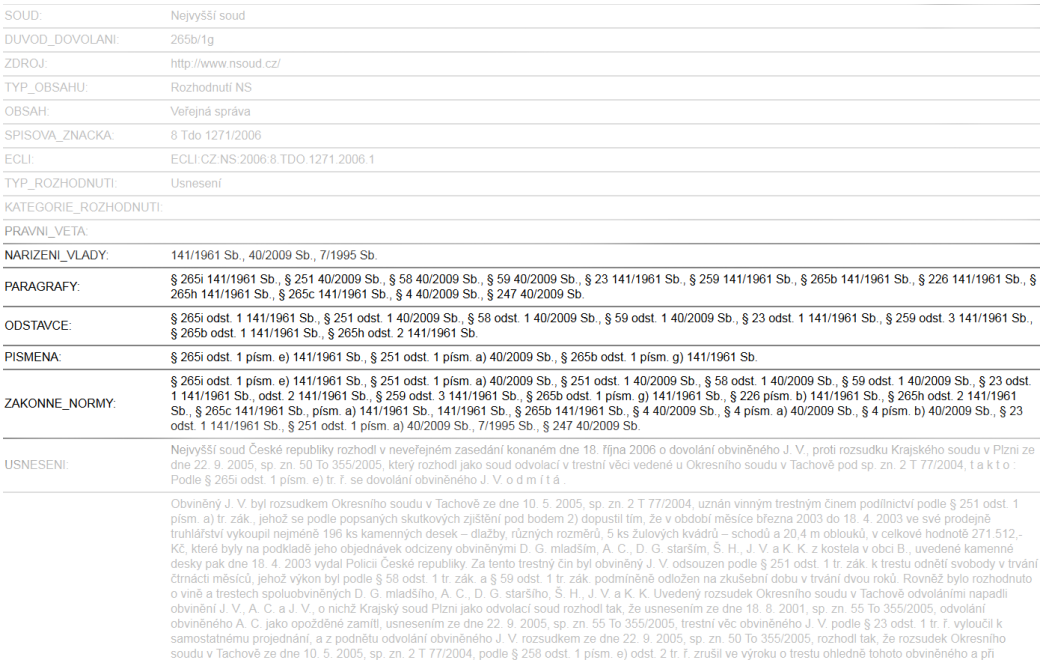
[1] Aplikace není zatím graficky stylizována.




Obrázek A.1: Screenshot z aplikace s ukázkou formuláře.




Obrázek A.2: Screenshot z aplikace s ukázkou facetové navigace.



| | |
|---|---|
| SOUD: | Nejvyšší soud |
| DUVOD_DOVOLANI: | 265b/1g |
| ZDROJ: | http://www.nsoud.cz/ |
| TYP_OBSAHU: | Rozhodnutí NS |
| OBSAH: | Veřejná správa |
| SPISOVA_ZNACKA: | 8 Tdo 1271/2006 |
| ECLI: | ECLI:CZ:NS:2006:8.TDO.1271.2006.1 |
| TYP_ROZHODNUTI: | Usnesení |
| KATEGORIE_ROZHODNUTI: | |
| PRAVNI_VETA: | |
| NARIZENI_VLADY: | 141/1961 Sb., 40/2009 Sb., 7/1995 Sb. |
| PARAGRAFY: | § 265i 141/1961 Sb., § 251 40/2009 Sb., § 58 40/2009 Sb., § 59 40/2009 Sb., § 23 141/1961 Sb., § 259 141/1961 Sb., § 265b 141/1961 Sb., § 226 141/1961 Sb., § 265h 141/1961 Sb., § 265c 141/1961 Sb., § 4 40/2009 Sb., § 247 40/2009 Sb. |
| ODSTAVCE: | § 265i odst. 1 141/1961 Sb., § 251 odst. 1 40/2009 Sb., § 58 odst. 1 40/2009 Sb., § 59 odst. 1 40/2009 Sb., § 23 odst. 1 141/1961 Sb., § 259 odst. 3 141/1961 Sb., § 265b odst. 1 141/1961 Sb., § 265h odst. 2 141/1961 Sb. |
| PISMENA: | § 265i odst. 1 písm. e) 141/1961 Sb., § 251 odst. 1 písm. a) 40/2009 Sb., § 265b odst. 1 písm. g) 141/1961 Sb. |
| ZAKONNE_NORMY: | § 265i odst. 1 písm. e) 141/1961 Sb., § 251 odst. 1 písm. a) 40/2009 Sb., § 251 odst. 1 40/2009 Sb., § 58 odst. 1 40/2009 Sb., § 59 odst. 1 40/2009 Sb., § 23 odst. 1 141/1961 Sb., odst. 2 141/1961 Sb., § 259 odst. 3 141/1961 Sb., § 265b odst. 1 písm. g) 141/1961 Sb., § 226 písm. b) 141/1961 Sb., § 265h odst. 2 141/1961 Sb., § 265c 141/1961 Sb., písm. a) 141/1961 Sb., 141/1961 Sb., § 265b 141/1961 Sb., § 4 40/2009 Sb., § 4 písm. a) 40/2009 Sb., § 4 písm. b) 40/2009 Sb., § 23 odst. 1 141/1961 Sb., § 251 odst. 1 písm. a) 40/2009 Sb., 7/1995 Sb., § 247 40/2009 Sb. |
| USNESENI: | Nejvyšší soud České republiky rozhodl v neveřejném zasedání konaném dne 18. října 2006 o dovolání obviněného J. V., proti rozsudku Krajského soudu v Plzni ze dne 22. 9. 2005, sp. zn. 50 To 355/2005, který rozhodl jako soud odvolací v trestní věci vedené u Okresního soudu v Tachově pod sp. zn. 2 T 77/2004, t a k t o : Podle § 265i odst. 1 písm. e) tr. ř. se dovolání obviněného J. V. o d m í t á . |
| | Obviněný J. V. byl rozsudkem Okresního soudu v Tachově ze dne 10. 5. 2005, sp. zn. 2 T 77/2004, uznán vinným trestným činem podílnictví podle § 251 odst. 1 písm. a) tr. zák., jehož se podle popsaných skutkových zjištění pod bodem 2) dopustil tím, že v období měsíce března 2003 do 18. 4. 2003 ve své prodejně truhlářství vykoupil nejméně 196 ks kamenných desek – dlažby, různých rozměrů, 5 ks žulových kvádrů – schodů a 20,4 m obloukú, v celkové hodnotě 271.512,- Kč, které byly na podkladě jeho objednávek odcizeny obviněnými D. G. mladším, A. C., D. G. starším, Š. H., J. V. a K. K. z kostela v obci B., uvedené kamenné desky pak dne 18. 4. 2003 vydal Policii České republiky. Za tento trestný čin byl obviněný J. V. odsouzen podle § 251 odst. 1 tr. zák. k trestu odnětí svobody v trvání čtrnácti měsíců, jehož výkon byl podle § 58 odst. 1 tr. zák. a § 59 odst. 1 tr. zák. podmíněně odložen na zkušební dobu v trvání dvou roků. Rovněž bylo rozhodnuto o vině a trestech spoluobviněných D. G. mladšího, A. C., D. G. staršího, Š. H., J. V. a K. K. Uvedený rozsudek Okresního soudu v Tachově odvoláními napadli obvinění J. V., A. C. a J. V., o nichž Krajský soud Plzni jako odvolací soud rozhodl tak, že usnesením ze dne 18. 8. 2001, sp. zn. 55 To 355/2005, odvolání obviněného A. C. jako opožděné zamítl, usnesením ze dne 22. 9. 2005, sp. zn. 55 To 355/2005, trestní věc obviněného J. V. podle § 23 odst. 1 tr. ř. vyloučil k samostatnému projednání, a z podnětu odvolání obviněného J. V. rozsudkem ze dne 22. 9. 2005, sp. zn. 50 To 355/2005, rozhodl tak, že rozsudek Okresního soudu v Tachově ze dne 10. 5. 2005, sp. zn. 2 T 77/2004, podle § 258 odst. 1 písm. e) odst. 2 tr. ř. zrušil ve výroku o trestu ohledně tohoto obviněného a při |

Obrázek A.3: Screenshot z aplikace se zobrazením entit typu Zákonná norma.

**8 Tdo 1271/2006**

| | |
|---|---|
| DREREFERENCE: | http://www.nsoud.cz/Judikatura/judikatura_ns.nsf/WebSearch/4DB771BC0B81D49BC1257A4E0068FCC1?openDocument |
| JINE_ROZSUDKY: | 8 Tdo 1271/2006, 50 To 355/2005, 2 T 77/2004, 55 To 355/2005, 50 T 355/2005 |
| DRETITLE: | 8 Tdo 1271/2006 |
| DREDATE: | 2006-10-18T00:00:00Z |
| DATUM_ROZHODNUTI: | 2006-10-18T00:00:00Z |
| DATUM_DEN: | 18 |
| DATUM_TYDEN: | 42 |
| DATUM_MESIC: | 10 |
| DATUM_ROK: | 2006 |
| OBLAST: | Trestní |
| SPZN: | Tdo |
| SPZN_POPIS: | dovolání proti pravomocným rozhodnutím soudů druhého stupně ve věcech trestních |
| SOUD: | Nejvyšší soud |
| DUVOD_DOVOLANI: | 265b/1g |
| ZDROJ: | http://www.nsoud.cz/ |
| TYP_OBSAHU: | Rozhodnutí NS |
| OBSAH: | Veřejná správa |
| SPISOVA_ZNACKA: | 8 Tdo 1271/2006 |

Obrázek A.4: Screenshot z aplikace se zobrazením entit typu Jiné rozhodnutí.

# B Ukázka celého dokumentu

Spisová značka: 4 Cmo 504/1998
Právní věta: Bylo-li řízení zastaveno podle ust. § 43 odst. 2 o.s.ř. a žalobce se podaným odvoláním domáhá, aby usnesení bylo zrušeno a řízení zastaveno z důvodu nezaplacení soudního poplatku, pak se jedná o odvolání směřující do důvodů napadeného usnesení, které je třeba odmítnout jako nepřípustné.
Soud: Vrchní soud v Olomouci
Datum rozhodnutí: 08.06.2000
Forma rozhodnutí: Usnesení
Heslo: Zastavení řízení
Kategorie rozhodnutí: C
Shora označeným usnesením soud prvého stupně zastavil řízení a rozhodl, že žádný z účastníků nemá právo na náhradu nákladů řízení. Výrok o zastavení řízení je odůvodněn ust. § 43 odst. 2 o.s.ř. Žalobce byl vyzván, aby opravil - doplnil neúplnou žalobu, byl vyzván, aby doplnil a uvedl správné a úplné obchodní jméno žalobce a správné a úplné sídlo žalovaného. Jelikož žalobce na výzvu soudu nereagoval, řízení bylo zastaveno.

Proti tomuto usnesení si podal žalobce včas odvolání a namítl, že řízení mělo být zastaveno podle zákona o soudních poplatcích, konkrétně podle ust. § 9 odst. 2 zákona č. 549/1991 Sb. a nikoliv podle ust. § 43 odst. 2 o.s.ř. Žalobce uvedl, že současně s výzvou k opravě žaloby, mu byla doručena i výzva k zaplacení soudního poplatku ve výši 500,- Kč. Jelikož soudní poplatek ve stanovené lhůtě nebyl zaplacen, a to ani dodatečně, měl soud, který doposud nezačal jednat ve věci samé, řízení zastavit, podle zákona o soudních poplatcích. Žalobce proto navrhl, aby odvolací soud napadené usnesení zrušil a rozhodl o zastavení řízení z důvodu nezaplacení soudního poplatku.
Odvolací soud se nejdříve zabýval otázkou, zda odvolání žalobce je přípustné.

Podle ust. § 201 o.s.ř. účastník může napadnout rozhodnutí soudu prvního stupně odvoláním, pokud to zákon nevylučuje.

Podle ust. § 202 odst. 3 o.s.ř. odvolání jen proti důvodům rozhodnutí není přípustné.

Z obsahu předloženého spisu odvolací soud zjistil, že žalobou podanou u Krajského obchodního soudu v Brně 2. 7. 1997 se žalobce po žalovaném domáhá zaplacení částky ve výši 5 442,- Kč. Usnesením z 25. 11. 1997 soud vyzval žalobce, aby opravil svou žalobu tak, že uvede správné a úplné obchodní jméno žalobce ke dni zahájení soudního řízení a uvede správné a úplné sídlo žalovaného ke dni zahájení řízení v souladu se zápisem v obchodním rejstříku. S touto výzvou byla žalobci současně doručena i výzva k zaplacení soudního

poplatku z žaloby ve výši 500,- Kč. Žalobce na žádnou z výzev nereagoval, a proto soud řízení zastavil s odůvodněním, jak výše uvedeno.

Po zhodnocení zjištěných skutečností odvolací soud dospěl k závěru, že odvolání žalobce je třeba odmítnout, neboť odvolání směřuje pouze proti důvodům napadeného rozhodnutí. Soud prvého stupně zastavil řízení podle ust. § 43 odst. 2 o.s.ř. pro neodstranění vad žaloby. Žalobce v odvolání namítá, že dle jeho názoru měl soud zastavit řízení dle ust. § 9 odst. 2 zák.č. 549/1991 Sb. o soudních poplatcích a nikoliv dle ust. § 43 odst. 2 o.s.ř., neboť současně s výzvou k opravě žaloby byla doručena i výzva k zaplacení soudního poplatku, který žalobce ve lhůtě stanovené soudem nezaplatil, a to ani dodatečně. Žalobce se proto podaným odvoláním domáhá, aby odvolací soud napadené usnesení zrušil a zastavil řízení z důvodu nezaplacení soudního poplatku podle zákona o soudních poplatcích. Je nepochybné, že odvolání žalobce směřuje toliko do důvodů napadeného usnesení, když žalobce nenapadá výrok o zastavení řízení, nesouhlasí pouze s důvody, pro které soud prvního stupně řízení zastavil a domáhá se toho, aby řízení bylo zastaveno z jiného důvodu, než který uvedl soud prvého stupně, konkrétně se domáhá toho, aby řízení bylo zastaveno pro nezaplacení soudního poplatku. Odvolání, kterým není napadán výrok rozhodnutí, ale pouze důvody, které vedly soud k jeho vydání, není přípustné.

S ohledem na výše uvedené proto odvolacímu soudu nezbylo, než odvolání žalobce podle ust. § 218 odst. 1 písm. c) o.s.ř. jako odvolání nepřípustné odmítnout.

Výrok o nákladech odvolacího řízení je odůvodněn ust. § 142 odst. 1 o.s.ř. za použití ust. § 224 odst. 1 o.s.ř. Žalobce nebyl se svým odvoláním procesně úspěšný a žalovanému v rámci odvolacího řízení žádné náklady řízení nevznikly, a proto o nákladech odvolacího řízení bylo rozhodnuto tak, jak uvedeno ve výroku tohoto rozhodnutí.

# C Návod k programu

## C.1 Požadavky

Pro správné přeložení a spuštění vyžaduje aplikace následující požadavky:

- Java[1] $\geq$ 1.6.0
- Ant[2] $\geq$ 1.9.0

Knihovny třetích stran jsou přiloženy v adresáři *lib* (viz. kapitola D). Není nutné s nimi manipulovat, při sestavování se k nim cesta nastaví automaticky.

## C.2 Sestavení

Sestavení je možné provézt pomocí nástroje Ant. Pro vytvoření spustitelné aplikace stačí otevřít adresář build se skriptem *build.xml* a zadat do příkazové řádky:

```
ant
```

V adresáři bin se vytvoří právě přeložená aplikace *NER.jar*.

## C.3 Spuštění

Po sestavení je aplikace je přeložena do spustitelného JAR archivu, který se nachází v adresáři *bin*. Aplikaci lze spustit následujícím příkazem:

*java -jar NER.jar arg1 arg2*,

[1] http://www.oracle.com/technetwork/java/javase/downloads/index.html.
[2] http://ant.apache.org.

kde *arg1* je první argument programu udávající akci, která má být provedena. Seznam akcí:

- 1 - Pravidlový NER. Musí následovat jako další argument cesta k vstupnímu souboru (*arg2*).
- 2 - Statistický NER. Musí následovat jako další argument cesta k vstupnímu souboru (*arg2*).
- 3 - Spuštění testu klasifikátoru (neobsahuje žádný další argument!).

V případě špatného spuštění je vypsán dialog s nápovědou pro správné spuštění.

## C.4 Výstup

Výstup programu je v kódování *utf-8*. V případě spuštění s argumentem *1* je výstupem programu výčet všech nalezených entit pomocí pravidlových metod. S argumentem *2* se zobrazí informace o tom, s jakou pravděpodobností se jedná o větu předělovou a věta samotná.

Výstupem programu v případě spuštění s argumentem *3* je vždy předělová věta s dodatečnými informacemi oddělenými dvojtečkou v následujícím formátu:

```
XML\1 Aps 6_2012.xml:true:0:0.0:0.9998839235880919:Posouzení
věci Nejvyšším správním soudem [ #cislo ]  Nejvyšší správní
soud při posuzování kasační stížnosti hodnotil, zda jsou
splněny podmínky řízení, přičemž dospěl k závěru, že kasační
stížnost má požadované náležitosti, byla podána včas a osobou
oprávněnou, a není důvodné kasační stížnost odmítnout pro
nepřípustnost.
```

- 1. - cesta k souboru
- 2. - true/false, věta je správně/špatně nalezená
- 3. - vzdálenost od pravé věty

- 4. - poměrná vzdálenost od pravé věty
- 5. - pravděpodobnost s jakou se jedná o předělovou větu
- 6. - rozpoznaná předělová věta

# D Obsah přiloženého DVD

K této práci je přiložen DVD disk, na kterém jsou uloženy zdrojové kódy, knihovny a nástroje pro přeložení aplikace. Jeho struktura je následující:

- bachelor-thesis - Elektronická verze (PDF) této práce
- bin - Adresář se spustitelnou verzí (generuje Ant).
- build - Adresář se skriptem *build.xml* pro Ant.
- config - Adresář s konfiguračními soubory
- javadoc - JavaDoc programátorská dokumentace
- lib - Externí knihovny
- src - Zdrojové kódy
- TXT - Adresář se vstupnímy soubory
- XML - Adresář pro trénovací XML soubory