# Specifikace softwarového projektu Textan

*Automatický analyzátor textů*

**Vedoucí projektu:** RNDr. Ondřej Bojar, Ph.D. (bojar@ufal.mff.cuni.cz)
**Řešitelé:** Petr Fanta, Duc Tam Hoang, Adam Huječek, Václav Pernička, Jakub Vlček
**Platforma:** primárně Windows (ale počítá se s platformovou nezávislostí)
**Očekávané obhájení:** září 2014

## Motivace

V dnešním světě existuje velké množství nestrukturovaných písemných zdrojů, které obsahují strukturované informace. Jako příklad můžeme uvést policejní zprávy. Ty obsahují informace, entity (jména, adresy, telefonní čísla, zbraně, drogy atd.) a vztahy mezi nimi, jež je třeba převést do podoby lépe zpracovatelné pomocí počítačů, například je zanést do databáze a přiřadit je k již existujícím objektům, pokud mají stejný význam. V současnosti probíhá extrahování těchto informací z policejních zpráv ručně, což je velmi zdlouhavé. Podobně časově náročné zpracování dokumentů lze jistě najít i v jiných institucích či společnostech.

## Cíl projektu

Cílem softwarového projektu Textan je vytvořit nástroj, který co nejvíce zefektivní dolování strukturovaných informací z nestrukturovaných dokumentů a umožní získané informace vhodným způsobem uložit a procházet. V jednotlivých dokumentech nástroj rozpozná entity a případně i vztahy mezi nimi. Jelikož entity označují nějaké objekty a tyto objekty se mohou vyskytovat ve více dokumentech, bude systém propojovat informace z různých dokumentů vztahující se k danému objektu. Nástroj má být co nejvíce univerzální, proto uživatel bude moci určovat druh rozpoznaných entit (respektive objektů) a vztahů a též, který úsek textu je reprezentuje. K získávání informací bude využívat automatické zpracování textu a strojové učení, avšak těžištěm projektu není dosažení co nejvyšší úspěšnosti.
Protože rozpoznávání nemůže být dokonale přesné, například kvůli nejednoznačnostem, bude aplikace umožňovat uživatelům opravovat výsledky automatického zpracování. Ručně potvrzené záznamy lze poté využít při automatickém zpracování dalších dokumentů a vylepšovat tak postupně výsledky automatického rozpoznávání, i kdyby počáteční přesnost nebyla dostatečná.
Data získaná z dokumentů budou moci uživatelé pomocí aplikace rovněž procházet, například formou grafu, jehož uzly budou tvořit objekty a hrany vztahy mezi nimi (viz dále).
Projekt Textan počítá s napojením na policejní systém zpracování zpráv, ale měl by být natolik obecný a samostatný, aby ho bylo možno použít i v jiných odvětvích.

## Charakteristika projektu

Hlavní náplň práce spočívá v návrhu a implementaci analyzátoru dokumentů, databáze informací získaných z dokumentů a aplikace umožňující pohodlnou kontrolu automatického zpracování a procházení databáze získaných informací.

Protože se předpokládá, že se systémem bude pracovat více uživatelů současně, bude použita architektura klient-server, aby bylo možné sdílet zpracované dokumenty a informace z nich získané. Server bude zprostředkovávat data uložená v databázi a také na něm bude umístěn analyzátor, aby se mohl průběžně vylepšovat podle textů přidávaných všemi uživateli a toto vylepšení se co nejrychleji projevilo u všech uživatelů. Komunikace klientské aplikace se serverem bude probíhat pomocí Webových služeb[1] (protokol SOAP). Vyhovíme tak požadavkům ze strany policie jakožto budoucího potenciálního uživatele systému, jelikož plánují využít pouze serverovou část projektu bez tlustého klienta, který bude součástí projektu.
Analyzátor je komponenta, která bude řešit jednak úlohy pro zpracování přirozeného jazyka, jednak porovnávání rozpoznaných entit s objekty v databázi. Počítá se s využitím již existujících nástrojů pro zpracování přirozeného jazyka, případně s tvorbou vlastního nástroje, pokud se ukáže, že vhodné nástroje neexistují, nebo je není možné použít.
Klientská aplikace běžící na lokálním zařízení umožní uživatelům snadno zpracovávat dokumenty a nabídne přívětivé prostředí pro kontrolu a opravu automatického rozpoznávání. Dále bude přehledně zobrazovat objekty a jejich vztahy uložené v databázi ve formě grafů a seznamů.

## Složení týmu

- 2 lidé na vývoj a implementaci serverové části, zvláště pak automatického zpracovávání dokumentů (analyzátor)
- 2 lidé na implementaci klientské aplikace
- 1-2 lidé na databázi a síťové služby zpřístupňující výkonné části (nejde o webové uživatelské rozhraní)

## Funkční požadavky (povinná část zadání)

- systém rozpozná v textu entity a určí jejich druh
- systém přiřadí rozpoznané entity v textu k objektům z databáze
- uživatel bude moci upravit rozpoznané entity (přidat, smazat, změnit typ, či rozsah entity)
- uživatel bude moci upravit přiřazení rozpoznaných entit k objektům
- uživatel bude moci mezi entitami vyznačit vztahy
- uživatel bude moci přidávat nové typy entit (respektive objektů) a vztahů
- systém se bude průběžně učit rozpoznávat entity a přiřazovat objekty podle zpracovaných dokumentů potvrzených uživatelem
- uživatel bude moci ztotožnit objekty uložené v databázi
- uživatel bude moci procházet objekty uložené v databázi včetně vztahů mezi nimi, jednak jako seznamy, jednak v podobě grafu
- uživatel bude moci k procházení databáze využít tyto druhy vyhledávání:
    - fulltextové vyhledávání v dokumentech
    - vyhledávání objektů, či vztahů podle typu
    - vyhledávání objektů dle jejich pojmenování
    - kombinace těchto metod
- uživatel může kdykoliv přerušit práci na aktuálním dokumentu

[1] http://en.wikipedia.org/wiki/Webservices

- průběžné změny při zpracování dokumentu se budou uchovávat lokálně, do databáze se budou ukládat hromadně po dokončení práce na dokumentu
- databáze musí logovat (žurnálovat) všechny provedené změny

## Rozšiřující funkční požadavky (nepovinná část)

- ztotožnění objektů bude moci uživatel zrušit s případným následným poloautomatickým zjednoznačněním výskytů rozpoznaných v době, kdy byly objekty ztotožněny
- systém určí vztahy mezi rozpoznanými entitami v dokumentu podle vztahů, které dříve v jiných dokumentech uživatel vyznačil
- uživatel bude moci upravit rozpoznané vztahy mezi rozpoznanými entitami v textu
- uživatel bude moci k procházení databáze využít i složitější vyhledávání:
    - vyhledávání spojení mezi objekty
    - vyhledávání vztahů podle slov navázaných ke vztahu (kotvy) v textu

## Mimofunkční požadavky (povinná část)

- systém musí podporovat paralelní připojení více klientů
- systém se musí umět zotavit z pádu serveru, či klienta
- systém musí být propojitelný s jinými systémy
- systém musí rozumně řešit nedostupnost serveru
- aktualizace automatického rozpoznávání (učení z nových příkladů) může trvat delší dobu, během učení se bude používat předchozí model
- systém musí být použitelný bez připojení k Internetu (například v uzavřeném intranetu)

## Pevně daná omezení

- entity v textu tvoří vždy souvislý úsek a nemohou se překrývat
- uživatel nebude moci editovat údaje uložené v databázi (jedinou formou aktualizace záznamů je vložení nového dokumentu a jeho anotace)
- souběžné anotování jedné zprávy více uživateli najednou nebude podporováno
- systém nebude řešit autentizaci a autorizaci uživatelů
- systém nebude řešit zálohování

## Předpokládané milníky

1. Seznámení s technologiemi (1. měsíc)
2. Úprava architektury vzhledem k použitým technologiím (2. měsíc)
3. Prototyp rozhraní (2. měsíc)
4. Prototyp analyzátoru a klientské aplikace (3. měsíc)
5. Propojení serverových komponent (3. měsíc)
6. Příprava dat pro analyzátor (4. měsíc)
7. Implementace rozhraní (4. měsíc)
8. Implementace analyzátoru a klientské aplikace (7. měsíc)
9. Testování (8. měsíc)
10. Dokumentace (8. měsíc)