



TEXTAN

Bc. Petr Fanta

Duc Tam Hoang, B.Sc.

Bc. Adam Huječek

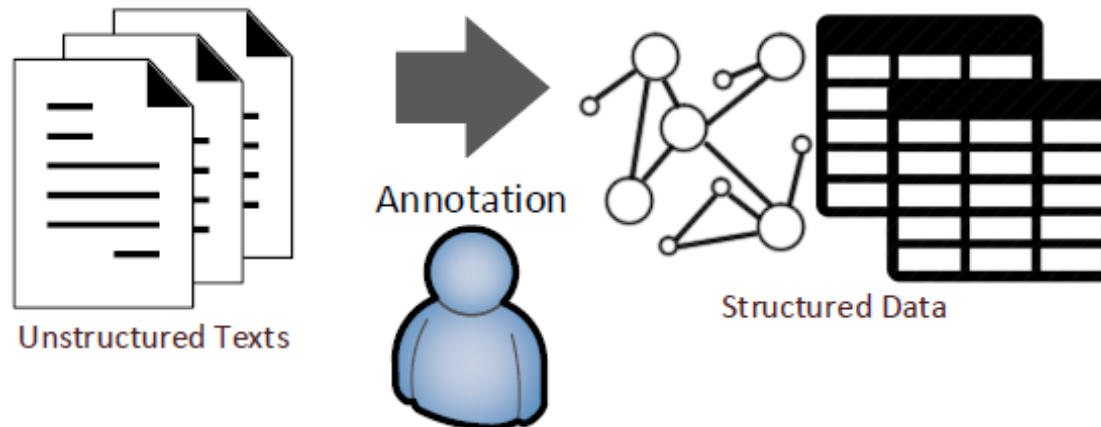
Bc. Václav Perníčka

Bc. Jakub Vlček

Vedoucí projektu: RNDr. Ondřej Bojar, Ph.D.

Motivace

- Velké množství textů v různých oblastech
- Nestrukturované dokumenty X strukturované informace
- Objekty a vztahy zachycené v dokumentech
- Příklad
 - Policejní zprávy
 - Proces zpracování zpráv v Policii ČR



Vzorová zpráva

Č.j.: NPC-707-37/2005

14.09.2005

Ú ř e d n í z á z n a m

Dnešního dne byl ve věznici PRAHA 4 PANKRÁČ vytěžen obv. Kuka David, nar. 11.6.1976. Předmětem vytěžení bylo získání Kuky ke spolupráci a získání informací k mezinárodnímu obchodu s OPL. Kuka se asi po hodinovém rozhovoru rozhodl spolupracovat s NPC jak na objasnění v jeho trestní věci, tak i na odhalení dalších skupin a osob, které dováží do Afriky kokain, hašiš, marihuanu a vyrábí a vyváží drogu extázi. Jako důvod spolupráce uvedl snahu o snížení trestu, který určitě dostane, a také proto, že mu ve vězeňské nemocnici zjistili, že pravděpodobně trpí diabetem. Schůzka byla omezena časově, z toho důvodu není vytěžení úplné, avšak na dalších návštěvách sdělí další informace.

Vzorová zpráva - informace

Č.j.: NPC-707-37/2005

14.09.2005

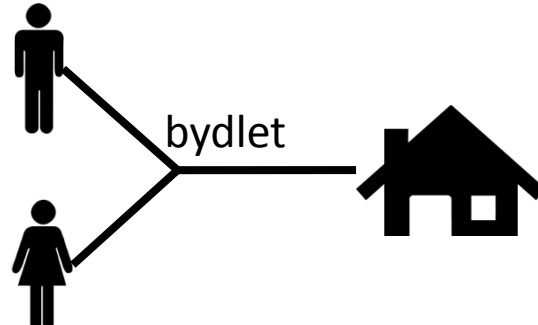
Ú ř e d n í z á z n a m

Dnešního dne byl ve věznici **PRAHA 4 PANKRÁC** vytěžen obv. **Kuka David**, nar. **11.6.1976**. Předmětem vytěžení bylo získání **Kuky** ke spolupráci a získání informací k mezinárodnímu obchodu s OPL. **Kuka** se asi po hodinovém rozhovoru rozhodl spolupracovat s NPC jak na objasnění v jeho trestní věci, tak i na odhalení dalších skupin a osob, které dováží do **Afriky kokain**, **hašiš**, **marihuanu** a vyrábí a vyváží drogu **extázi**. Jako důvod spolupráce uvedl snahu o snížení trestu, který určitě dostane, a také proto, že mu ve vězeňské nemocnici zjistili, že pravděpodobně trpí diabetem. Schůzka byla omezena časově, z toho důvodu není vytěžení úplné, avšak na dalších návštěvách sdělí další informace.

Terminologie TextAnu

- „Adam a Eva žili v Ráji.“
- Entita
 - Adam a Eva žili v Ráji.

- Objekt
 - Osoby: Adam  a Eva 
 - Adresa: Ráj 

- Relace

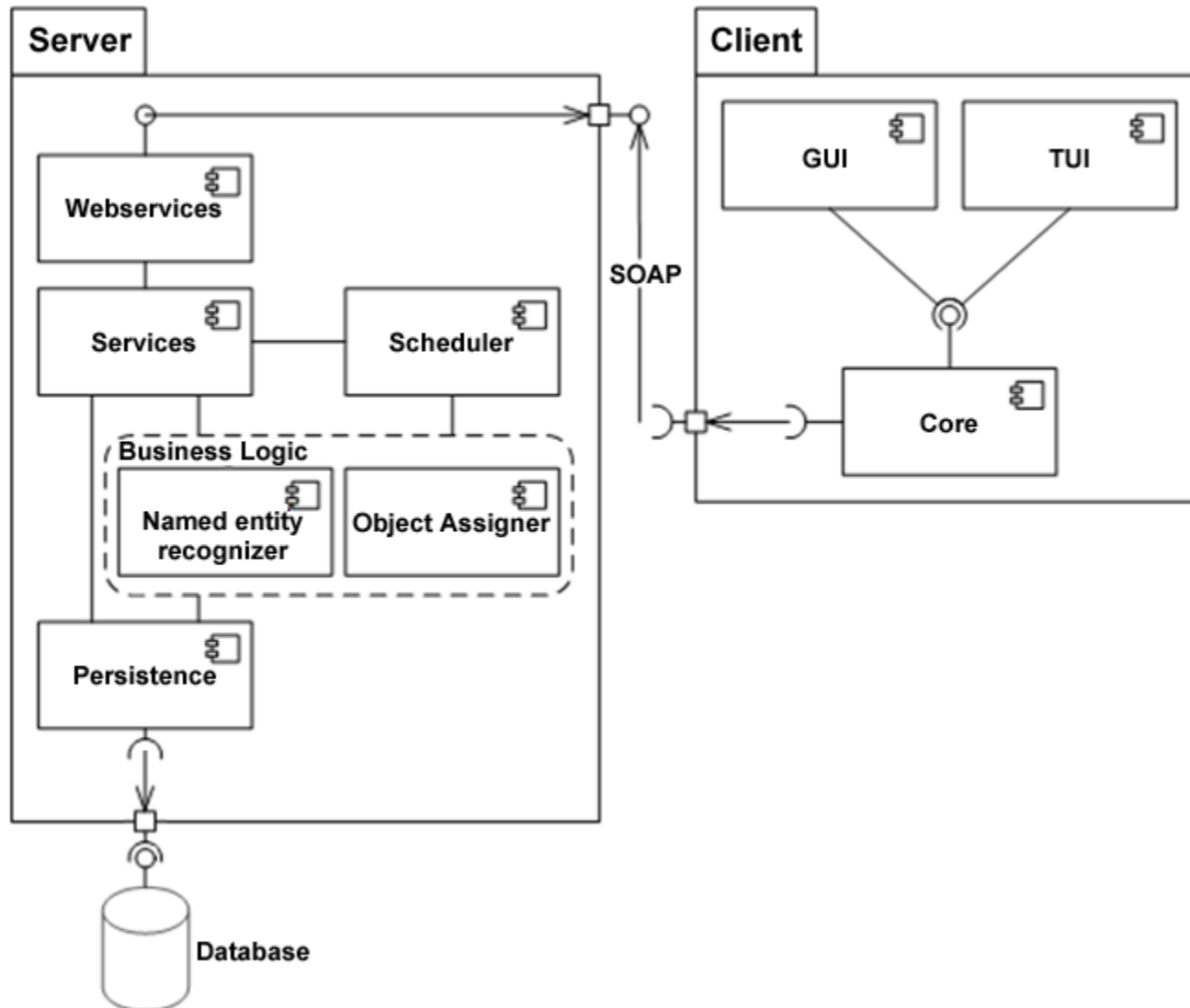
The diagram illustrates the relationship between the entities and the action. On the left, there are two person icons: a male figure (Adam) and a female figure (Eva). On the right, there is a house icon. Two lines originate from the person icons and converge at a point, from which a single line extends to the house icon. The word 'bydlet' (live) is written above the convergence point, indicating the relationship between the subjects and the location.

Co je to TextAn?

- (Polo)automatický analyzátor textů a databáze textů
- Jazyk – čeština (další jazyky konfigurovatelné)
- Cílová doména
- Zpracování dokumentu
 - Entity (automatické rozpoznání)
 - Objekty (automatické přiřazení)
 - Vztahy (manuální vyznačení)
- Procházení dat
 - Sekvenční
 - Grafy



Architektura



Technologie

- Server
 - NameTag & MorphoDiTa
 - WEKA
 - Spring Framework
 - Apache CXF
 - Jetty
 - Hibernate ORM
 - Hibernate Search
 - SLF4J & Logback
- Klient
 - ControlsFX
 - JFXtras
 - JUNG & PretopoLib
 - JCommander

Požadavky na spuštění

- Server
 - Java 8 & C++11
 - Podporované platformy
 - Windows (32 bit / 64 bit)
 - Linux (32 bit / 64 bit)
- Klient
 - Platformově nezávislý
 - Java 8

Zpracování dokumentu - Pipeline

- Výběr zdroje zprávy
- Editace textu
- Editace entit
- Editace objektů
- Editace vztahů

Ukázka



Rozpoznávání pojmenovaných entit

- NameTag (a MorphoDiTa)
 - Strojové učení s učitelem (MEMM a Viterbi decoder)
- Přetrénování modelu rozpoznávače
 - Generování trénovacích dat
 - Asynchronní zpracování
 - Konfigurovatelnost

Přiřazení entit k objektům

- Problémy
 - Jednoznačnost
 - Neuplnost dat
- Klasifikace (kNN) i ranking
- Features
- Trénování

Budoucí vývoj

- Automatické rozpoznávání relací
- Vyhledávání v grafech
- Podpora více jazyků najednou
- Omezení pro objekty v relacích
- A mnoho dalších vylepšení...

Testování a praktické využití

- Testování na skutečných policejních zprávách
- 28000+ zpráv
- Napojení na existující policejní systémy přes webové služby
- Příprava dat pro jiné projekty

Děkujeme za pozornost

Více na [http:// github.com/PreXident/TextAn](http://github.com/PreXident/TextAn)