



MVCRX037TIV1
prvotní identifikátor

Č E S K Á R E P U B L I K A
M I N I S T E R S T V O V N I T R A

Praha 7, Nad Štolou 936/3, IČ: 00007064

zastoupená

JUDr. Petrem Novákem, Ph.D.

ředitelem odboru bezpečnostního výzkumu a vzdělávání

Kontaktní adresa: Nad Štolou 3, 170 34 Praha 7

Č. j. MV-152905-8/OBVV-2015

Praha 9. listopadu 2016

**Veřejná zakázka – Systém pro analýzu textových dat pro potřeby Policie ČR –
dodatečná informace č. 1**

K UVEŘEJNĚNÍ NA PROFIL ZADAVATELE

Zadavatel obdržel od účastníka řízení následující dotazy k výše uvedené veřejné zakázce.

Dotaz č. 1:

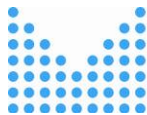
Dotaz zní, zda a jakým způsobem budou k dispozici data, která bude systém zpracovávat? Budou tato data již předem anotována, tj. označena jednotkami, které se budou hledat? Předpokládáme, že tato data budou velmi důvěrného charakteru, nicméně jejich poskytnutí je klíčové pro úspěšné vyřešení úlohy.

Odpověď na dotaz č. 1:

Po zhotoviteli se požaduje vyvinout systém, který bude pracovat s daty z obecně známých, česky otagovaných korpusů. Následně bude třeba systém přizpůsobit (natrénovat) na korpus policejních textů. Tento si musí zhotovitel vytvořit sám na základě anonymizovaných vzorových příkladů, které mu budou objednatelem poskytnuty. Výsledná evaluace řešení a korigování výstupů (ladění) bude probíhat pouze v prostředí objednavatele v oddělené síti na HW objednavatele. Tento režim je nutný z hlediska ochrany citlivých dat.

Dotaz č. 2:

Jakým způsobem se bude prokazovat, že systém splňuje požadovanou funkcionalitu (tj. hodnocení přesnosti identifikace pojmenovaných entit, přesnost hledání



“podobných” případů se stejnou skutkovou podstatou, vyhodnocení výsledků sumarizace, apod.)?

Odpověď na dotaz č. 2:

Předpokládá se vytvoření sumarizačního reportu jaké entity a vazby byly v textu nalezeny a možnost jejich porovnání s očekávaným výstupem (vstupní kontrolní report). Základem je spárování nalezených a kontrolních entit a vazeb, určení poměru false pozitive nalezených entit a vazeb k očekávanému, dále pak komplexnost nalezených dat (např. provázání jména, příjmení a data narození u osob nebo ulice, číslo orientační/popisné, obec u adres) a přesnost (např. druh vazby).

Řešiteli budou dodány vzorové dokumenty i s očekávaným výstupním reportem, použitými číselníky k vývoji. Ověření realizace pak proběhne nad reálnými texty, které budou již ručně zpracovány analytikem a dojde k posouzení efektivnosti řešení.

Objektivní zadání je, aby:

- systém zrychlil manuální činnost analytika textů při extrakci entit a jejich vazeb. Pokud by manuální zpracování textů bylo rychlejší než automatické s následnými manuálními opravami analytikem, systém zadání nesplnil,
- při hledání podobných případů bylo předloženo co nejvíce relevantních výsledků a co nejméně falešných (F skóre),
- výsledkem sumarizace bylo co nejvýstižnější shrnutí dokumentu.

Dotaz č. 3:

Má být podporována vizualizace nalezených výsledků fulltextového vyhledávání formou orientovaného interaktivního grafu pojmenovaných entit. Co má tento graf v nalezených výsledcích zobrazovat a zda uvažujete i k-partitní graf (různé typy uzlů)?

Odpověď na dotaz č. 3:

Výstupní vizualizace ve formě grafu musí vycházet ze standardního strukturovaného formátu dat. Lze využít veřejně známy formáty XML, CSV, GEXF, GDF, GML, Pajek NET nebo komerční formát ANB (IBM i2 Analyst Notebook). Následná vizualizace tohoto formátu je na volbě zhotovitele. Každý typ entity vyžaduje svůj vlastní typ uzlu. Každý typ vazby mezi entitami vyžaduje vlastní typ hrany.



Dotaz č. 4:

Je požadováno vytváření souhrnu nejdůležitějších faktů. Jak má takový souhrn vypadat? Jaká fakta jsou zadáním myšlena?

Odpověď na dotaz č. 4:

Sumarizace vychází z definice z oboru Natural Language Processing, např. https://en.wikipedia.org/wiki/Automatic_summarization.

Dotaz č. 5:

Jako přínos projektu uvádíte usnadnění vyhledávání za pomoci fulltextového nástroje, kde je dále upřesněno, že bude použit "inteligentní dotazovací jazyk s logickými operátory a uvážením ohebnosti českých slovních tvarů i nalezených entit". Můžete prosím toto blíže specifikovat?

Odpověď na dotaz č. 5:

Zadavatel vyžaduje systém, jehož součástí bude i fulltextové jádro jako implementace volně dostupných fulltext engines, např. Lucene či SOLR, jejichž součástí je vlastní dotazovací jazyk, který je třeba doplnit o parametry hledání jednotlivých typů entit. Fulltext musí rovněž nabízet funkce stemmingu/lemmatizace na český jazyk.

Dotaz č. 6:

Ve specifikaci výsledků je uvedeno, že vstup bude podporovat "libovolné" elektronické dokumenty. Prosím specifikujte formát těchto požadovaných dokumentů.

Odpověď na dotaz č. 6:

Libovolnými elektronickými dokumenty se myslí dokumenty, které obsahují textovou informaci, tudíž hlavní obsah dokumentu a dále připojená metadata. Systém bude interně pracovat s formátem čistého textu (plain text v režimu kódování UTF-8), do něhož je třeba vstupní formáty dokumentů převést (např. volně dostupný nástroj Apache Tika či Google Tesseract).

za zadavatele
JUDr. Petr Novák, Ph.D.
ředitel
v z. PaedDr. Jan Vykoukal