Zápisnica zo stretnutia č. 7

Zapísal: Bronislava Pečíková

November 6, 2017

Predmet: Tímový projekt
Názov projektu: Deep search
Názov tímu: Neverest
Číslo tímu: 26

Čas: 11:00 - 14:00
Miestnosť: FIIT STU 3.28

Prítomní:
Peter Berta, Erik Jankovič, Matej Adamov, Michal Krempaský, Oliver Macko, Bronislava Pečíková
Neprítomní:

Obsah stretnutia:

- Michal nám odprezentoval naštudované možnosti pre lematizáciu textu v českom jazyku
  - ElasticSearch plugin: obsahuje tokenizáciu, lematicáciu odstránenie stop slov, priraďovanie morfologických značiek
  - dokumentácia k analýze dostupných technológii: `https://docs.google.com/document/d/1o7CnIyK6U4Ph1l7eq90xXWSmJAEFYvmYEuwETgGkC-A/edit`
- Erik nás upozornil, že sa mu nepodarilo lematizovať text v ElasticSearch plugine - LemmaGen a bude potrebné overiť, či to bude možné
- Michal doplní naštudované poznatky do dokumentácie
- Erik nám odprezentoval naimplementovanú lematizáciu:
  - Treba dorobiť: tokenizácia dátumov, tokenizácia viet, odstránenie nechcených znakov
- Matej nám odprezentoval simuláciu testovania identifikovaných kvalifikátorov

Závery vyplývajúce zo stretnutia:

- Neštruktúrované texty budeme ukladať v ElasticSearch

- Tokenizovať, lematizavať, odstráňovať stop slová a priraďovať morfologické značky budeme pomocou príslušných pluginov k ElasticSearch

Úlohy vyplývajúce zo stretnutia:

- Doplniť dokumentáciu inžinierského diela o poznatky na ktoré prišiel počas štúdia dostupných technológii na spracovanie českého jazyka - Michal
- Doplniť dokumentáciu zo stretnutí tímu o závery a úlohy - Broňa
- Vytvorenie šablón pre dokumentáciu v Latexe - Peter
- Predspracovanie dokumentov: tokenizácia dátumov, tokenizácia viet, odstránenie nechcených znakov
- Otestovať ElasticSearch plugin - LemmaGen pre český jazyk
- Otestovat ElasticSearch plugin Morphodita - Michal