Zápisnica 5

Dátum	17.10.2018
Miesto	FIIT STU, 4.26
Zapisovateľ	Dávid Csomor

Zhodnotenie 1. šprintu + poznámky vedúceho

- Import je vyriešený
- Anotácia je vybavená (ale je problém s obmedzeným počtom znakov zo strany API vedúceho)

Organizácia + návrhy na zlepšenie:

- Bude sa treba vážne venovať validácií
- My ako tím sme mali zistiť koľko story pointov vieme dodať za týždeň
- Môžeme to robiť benevolentnejšie v zmysle, že ak bude "TO" hlavné vybavené, tak môžeme robiť aj veci z/do budúcnosti (nejaké veci , ktorým sa budeme venovať v budúcnosti)
- Mali by sme robiť 9h/ týždeň na člena
- V rámci toho tímáku sa robí viacero parserov (anotácia, invertované indexy)
    - vieme si to rozdeliť a robiť ako keby dopredu
- Bude sa treba vážne venovať validácií
- Zváženie zapojenia sa do "blbec dňa":
    - zadávateľ sa stretne a pokúsi sa vybaviť napr. nacrawlované dáta
- Streda 31.10 nebude stretnutie → rozdelíme šprinty na 1T a 2T

Diskusia k taskom:

- Velká relačná tabuľka, kde bude:
    - word
    - lemma
    - POS
    - NER (pole)
    - count
    - tfidf-document (frekvencia výskytu slova z hľadiska dokumentu)
    - tfid-corpus
    - Ngrams (pole ngramov)
- Pre každý spracovaný článok si vytvoriť dátum
- Zatiaľ netreba riešiť tfidf
- Drilldown je slovo ←→ článok tak, že budeme môcť postupne konkretizovať filter a zmenšovať zobrazené výsledky
- Pri 2 článkoch rozdelíme obrazovku na 2 časti a štatistiky budú pod tým
- Najjednoduchšie na parsovanie sú webnoviny (neexistuje rozumná alternatíva)

- Reportovať zlé správanie poskytovateľa API cez slack, ak by dobre nefungovalo
- Vedúci má slovník, ktorý nám vie poslúžiť na hľadanie synoným + slovník azet
- Slovník sa bude budovať iba na kolekciu top 1000 POTOM, čospravíme invertovaný index
-

TODO:

- Pridať do datasetu mestá, vrchy, osobnosti (osobnosti z wiki)
    - tie osobnosti bude asi treba manuálne naklikať
- Treba začať robiť výstupy zo šprintov ala „dokumentácia riadenia“
- Začať robiť invertovaný index
- **Dokumenty:**
    - dokumentácia k riadeniu (-) - (vždy musí byť po šprinte)
    - dokumentácia k inžinierskemu dielu (architektúra a pod.) - (na konci semestra)
- Od 2. šprintu treba lepšie písať user stories
    - pridávať aj akceptačné kritériá
    - pridávať opis
    - user stories robíme my ako tím/členovia
    - vedúci sa stará a FEATURES nie U.S.
- Pridať dátum „kedy sme článok pridali“, „kedy sme ho oanotovali“, „kedy sme vytvorili invertovaný index“
- Články by sa mali dať reanotovať
- Byť schopný pridať dátum a anotáciu automaticky napr. články v korpuse Slovensko sa pridá dátum a tag Slovensko pre všetko v corpuse Slovensko automaticky
- Webcrawler pre Webnoviny – šport
    - treba aj dokument k tomu crawleru
    - Treba, aby bola zachovaná hierarchia
    - dátum, že kedy sa to stiahlo
- Vyriešiť taxonómiu (ako budovať slovník)
    - Ísť na stránku s cudzím jazykom a odtiaľ si stiahnuť/ inšpirovať sa „slová + vzťah“
- Pozrieť fast-text word2vec
- Spraviť kolekciu pre vety