Zápisnica 2

Dátum	26.09.2018
Miesto	FIIT STU, 4.26
Zapisovateľ	Dávid Csomor

Analýza trackovacích nástrojov:

- logovanie hodín: toggl alebo clock
- percentuálne hodnotenie zo 100% + vyjadriť kto koľko spravil
- používať scrum poker cards (odhaľovať naraz)
- TP CUP (hlasovanie: 2 áno, 6 neutrálnych)

MongoDB overview:

- alternatívy: mongoVUE, RockMongo (na 2. treba PHP)
- treba vytvoriť štruktúru mongo databázy
- collection
- corpus: zhluk/kolekcia článkov (mal by byť ako cudzí kľúč, mongo štandardne nemá join)
- články z jedného zdroja (články zo stránok wiki a tie vyfiltrovať)
- pridávať tagy (budeme ich tam písať manuálne ako zoznam slov oddelených čiarkou)
    - dkoument
    - bude jedna inštancia (článok) už spracovaný text (rozdelený na slová, lemy)
    - SLOVO (uchovávanie duplicít je v poriadku (aj keď majú iné kľúče)
    - kategória - POS (part of speech) – gramatické kategórie/ slovné druhy
    - slovo → NER (named entity recognition) – názvoslovná entita

Zdroje:

- wiki, žurnály (žurnály neskôr), SITA/TASR (zdroj správ zo zahraničia, možno majú nejaké RSS, ktoré by sme mohli parsovať)

Spracovanie:

- 1. možnosť: wikipédia umožňuje vrátiť stránku v JSONE ak pridáme nejaký špeciálny tag k URL
- 2. možnosť: vlastný parser a parsovať rovno z URL
- na úvod napr. zobrať štáty z wiki a spracovať tie
- Spraviť rozhranie na vizualizáciu dát z databázy
- bude obsahovať:
    - zoznam článkov
    - priemerná dĺžka článkov
    - počet znakov, slov, plnovýznamových slov, slovies, ...
    - robiť štatistiky nad článkami s vybraným tagom
    - histogram (slov, lem, slovných druhov, názvoslovných entít, tagov, N-gramy (možno v budúcnosti))

    - KWIC (keyword in content) – N-gramy
    - tag-cloud
    - frekvencia slov
    - TFIDF – metrika, ktorá počíta s frekvenciou slova ?? (používa sa napr., aby sa nezvýhodňovali dlhšie články)
    - prideľovanie váhy slovám podľa toho ako často sa používajú
- treba vedieť odpovedať na dotazy: histogram mužských rodov, ...
- treba vedieť ku každému článku pridať tag, aby sa potom dali hľadať podobné

Rozhranie pre import:

- môže byť CLI (script bez GUI)
- musí obsahovať:
    - import: názov článku, text
    - tokenizácia (rozsekať a označiť)
    - export

Do budúceho stretnutia:

- štruktúra aplikácie
- štruktúra DB
- plagát
- low fidelity návrh pre UI
- ako stiahnuť články z wiki
- nájsť/pozrieť nástroje na spracovanie textu a spraviť report (čo tam je, či to funguje a či to vieme použiť):
    - text.fiit.stuba.sk
    - nametag, morphodita – Masarykova univerzita (vedúci pošle odkaz)

MISC:

- TFIDF odfiltruje často používané slová a ostanú nám slová špecifické pre danú tématiku
- "word2cec" – prevádza slová do vektoru - algoritmus na predpovedanie postupnosti slov (vieme pomocou neho hľadať synonymá)

Linky:

- SAV morfológia - https://korpus.sk/morpho.html
- SAV slovník - http://slovniky.juls.savba.sk/

Zoznam high-level úloh (p-backlog)

- vytvorenie webu tímu (+ administrácia, Paťo) 5-8
- vytvoriť plagát !!! (Júlia) 2-3
- navrhnutie štruktúry mongoDB !!! (danko, alan, kiko) 1-5
- zistiť možné klientske apps pre mongoDB 1-3
- zireframe-y pre web (Paťo) 5

- zistiť možnosti importu článkov z wiki (Peťo, Dávid) 5-8
- nájsť iniciálny insert pre vzorové články (return json) (danko, alan, peto) 2-11
- nájsť nástroje na spracovanie textu (text.fiit.stuba.sk -> report / name tag + morphodita -> masarykova) (Adam, David + ?) 5
- navrhni architektúru systému (Dankova skica môže byt helpful) !!! 3
- rozhranie pre webovu appku (zobrazenie dát z Db do user-friendly formy)
- vytvoriť crawler na získavanie žurnálových článkov (sme, aktuality, sita, tasr)

mongoDB

- vytvorenie štruktúry
- collection: články (názov, text (origin text), tagy -> tagy o článku (šport, etc.), tokeny -> pole slov, ktoré uchováva pole informácií)
- korpus (zhluk článkov) – napr. všetky články z wiki
- document: článok
- uchovávanie spracovaného textu (rozdelený na slova, pre každé slovo uchovať jeho lemu (peknej -> pekný))
- uchovať gramatické kategórie (POS -> part of speech (rod, číslo, pád, vzor))
- uchovať rozpoznanie entít (NER -> name entity recognition (podstatne etc.))
- kategorie NER (osoba, (person from locality -> bratislavčan (variable)), lokalita, organizácia, other, dátum, čas, mena (currency))
- môžeme uchovať duplicity (ak to ma význam -> selecty etc.)
- compass, mongovue, mongo rock (alebo iný client)

Rozhranie pre vizualizaciu

- zoznam korpusov
  - zoznam článkov
  - informácie o dĺžke článku / počet slov/ počet plnovýznamových slov / počet slovies/ etc
  - umožniť pridať k článkom tagy
  - zobrazenie N-gramov (toto je future work) , najčastejšie vyskytujúce sa N-tice slov
  - WORD 2 VEC (pozri co to je, môže byt užitočné)
  - Funkcie and korpusom:
    - histogram (podľa ľubovoľnej kategórie) + v budúcnosti histogram N-gramov
    - KWIC – keyword in content
    - tagcloud
    - frekvencia slov (Therm frequency) – vzorec na to, ako často sa používa slovo a aká je jeho priorita
    - TF-IDF

Rozhranie pre import

- Môže byť cmd script
- Iniciálny upload do DB (názov článku + text)
- Tokenizácia (otagovať článok, rozdeliť na slova etc.)

- Export (do rozumného formát)

Zdrojové dáta

- Zdroje: Wikipedia , žurnálové články (na wiki robiť export článkov)
- Pre wiki spraviť research ako chceme získavať údaje (ci export, pomocou URL získať JSON (asi wiki api -> treba nájsť), crawler etc.)

Linky

- https://korpus.sk/morpho.html → SK language
- https://dumps.wikimedia.org/skwiki/latest/ → wiki db dump SK
- http://ufal.mff.cuni.cz/morphodita --> morphoDiTa