Zápisnica 4

Dátum	09.10.2018
Miesto	FIIT STU, 4.26
Zapisovateľ	Dávid Csomor

Krátke zhrnutie

Alan s Danielom spravili API pre pridávanie a updatovanie korpusov a článkov a naštudovali MongoDB. Peťo spravil manuál pre TFS a upravil celkové rozloženie práce v TFS. V TFS sa vďaka tomu už dá pekne orientovať. Bol vyriešený tiež problém s GITom. Júlia pokračovala v práci s TP-CUPom. Paťo zdokonaľoval frontend, Krištof riešil úpravu skriptov na ťahanie dát. Adam si spravil prehľad o projekte.

Úlohy vyplývajúce zo stretnutia

- vytvoriť kolekciu synoným (synonymické množiny/ sety)
- z textu potrebujeme vyťahovať features, na základe ktorých vieme text klasifikovať
- zamyslieť sa aké features by sme používali:
- distribúcia podstatných mien,
- subj. obj. slov
- percentuálny podiel interpunkcií
- medián (pre vylúčenie outlierov)
- priemerná dĺžka vety
- podiel slov, ktoré sú lokácia
- podiel slov, ktoré nevieme identifikovať
- uchovávanie pomocou invertovaného indexu a pre naše účely aj naopak (treba mať „inverzný“ skript ku inverznej indexácií)
- prípadne spracovať inú sekciu Wiki (zoznam miest, zoznam vrchov nad 400m+-, rieky nám vie poskytnúť vedúci)
- synonymá hľadať po lematizácií
- existuje slovník „wordnet na JULS“, reprezentuje synonymické sety pomocou ID
- STOP slová by sme si mali vybudovať samy (z invertovaného indexu to vieme zistiť)