Zápisnica 18

Dátum	12.03.2019 – 08:00
Miesto	FIIT STU, 4.26
Zapisovateľ	Dávid Csomor

Úvod – pripomienky

- Niektoré slová sú zle otokenované a označené
- Odtlačok článku:
    - Plnovýznamové slová
    - Neplnovýznamové slová
    - STOP slova – budú také, ktoré sa budú nachádzať vo veľa článkov – z tfidf
    - TOP.global 1000 – zatiaľ nie
    - TOP.$corpus 1000 – zatiaľ nie
    - Pozitívne, negatívne, neutrálne – zatiaľ nie

Ďalší šprint

- **Odtlačok/TFIDF korpusu vs článku**
- Každý článok by mal mať odtlačok
- Každý korpus by mal mať odtlačok
- Porovnanie odtlačku korpusu vs článku
- Vyberieme nejaký token článku a porovnáme pravdepodobnosť výskytu tokenu v iných článkoch (prípadne korpusov)
- "Článok X, sa podobá na 20% článku Y"
- „Článok X má takéto slová a na základe toho bol priradený do korpusu Z"
- Pri porovnaniach, porovnávať články s článkami a korpusy s korpusmi
- Vybrať 150 slov z každého článku a počítať
- 1 tabuľka pre článok a 7 tabuliek pre korpus

- **Miniaplikácia**
- Pouvažovať nad interaktívnou miniaplikáciu za účelom zbierania dát od používateľov