Objavovanie vzťahov medzi kľúčovými slovami

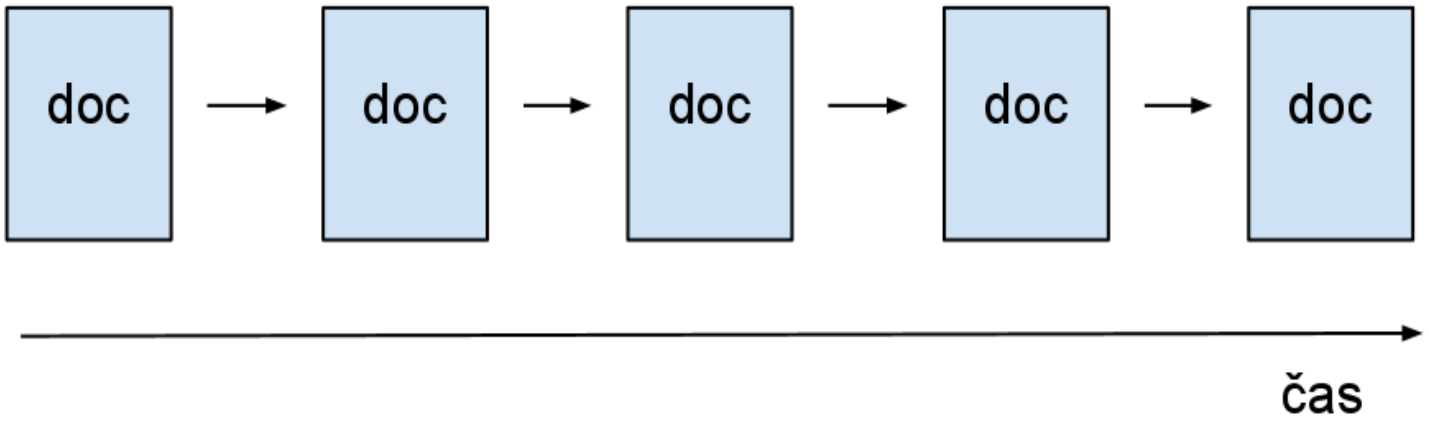
Peter Kajan
Vedúci: Michal Barla

Pripomenutie

- Hľadanie vzťahov
- z analýzy logov prezerania dokumentov
- a z nich automaticky získaných kl. slov

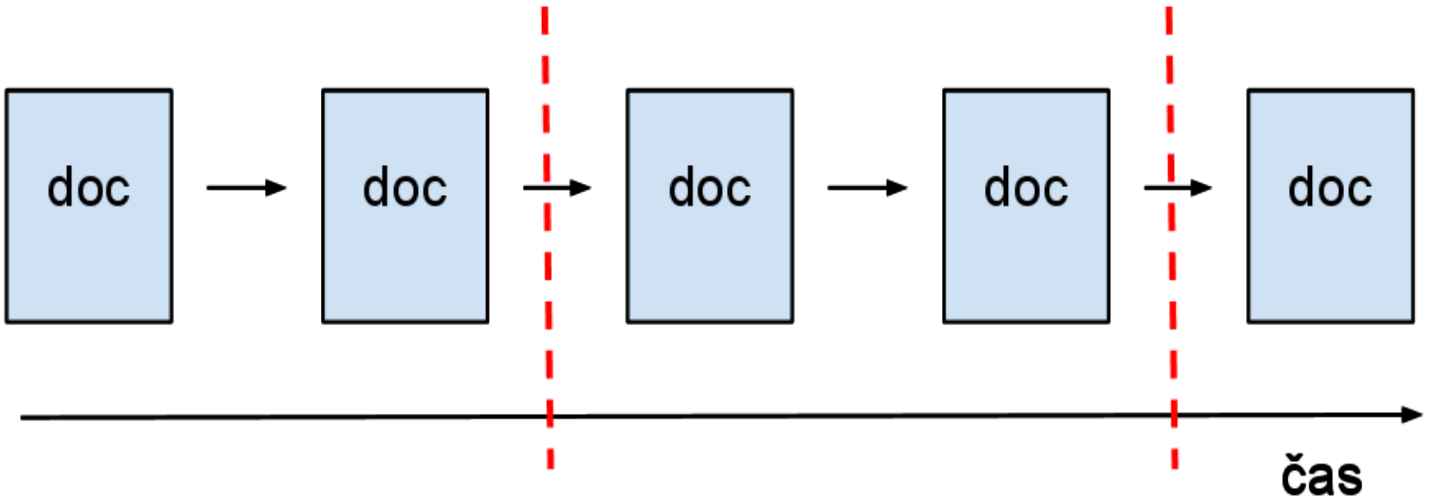
- Cieľ: zistenie vplyvu časovej zložky

Logy



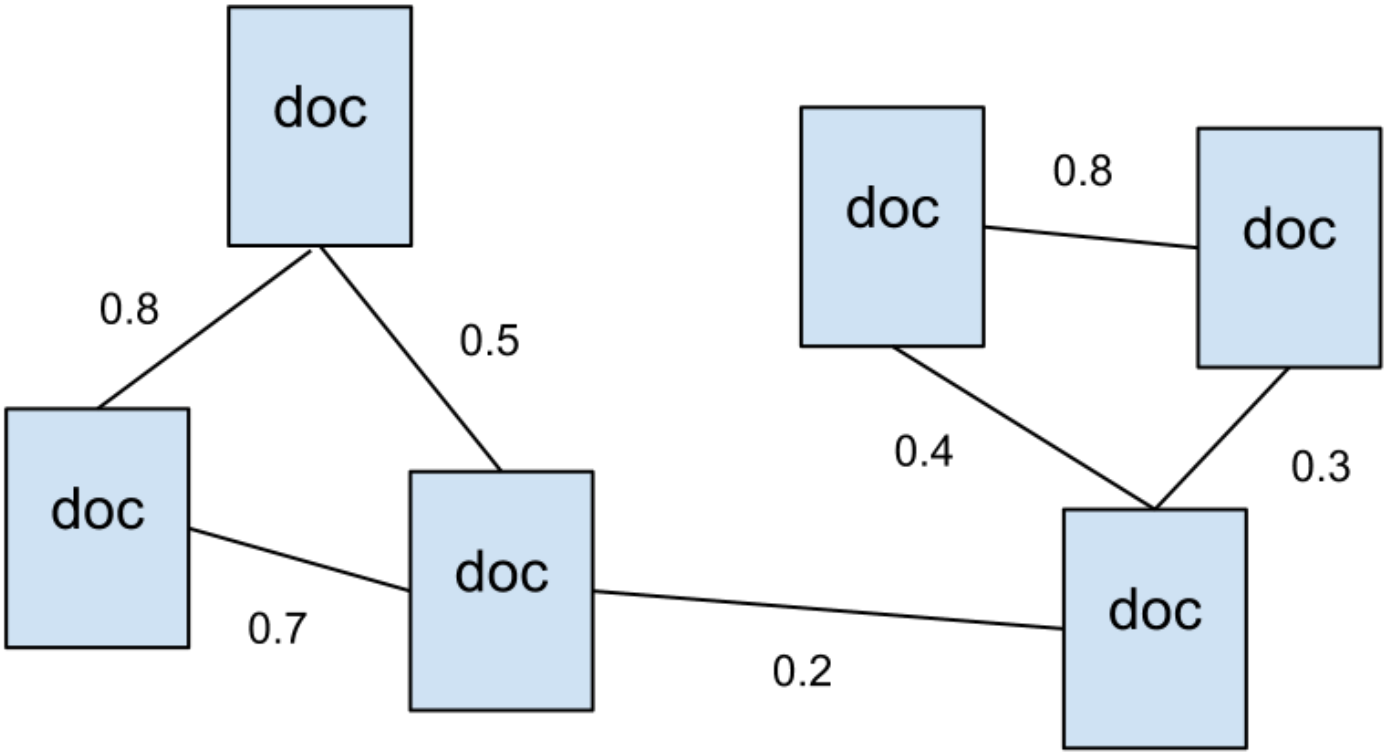
- Počet záznamov: 560 000
- Preprocesing: 13 100
- Odhodenie mail, facebook, youtube, rovnaké posebe idúce stránky: cca. 5000

Logy

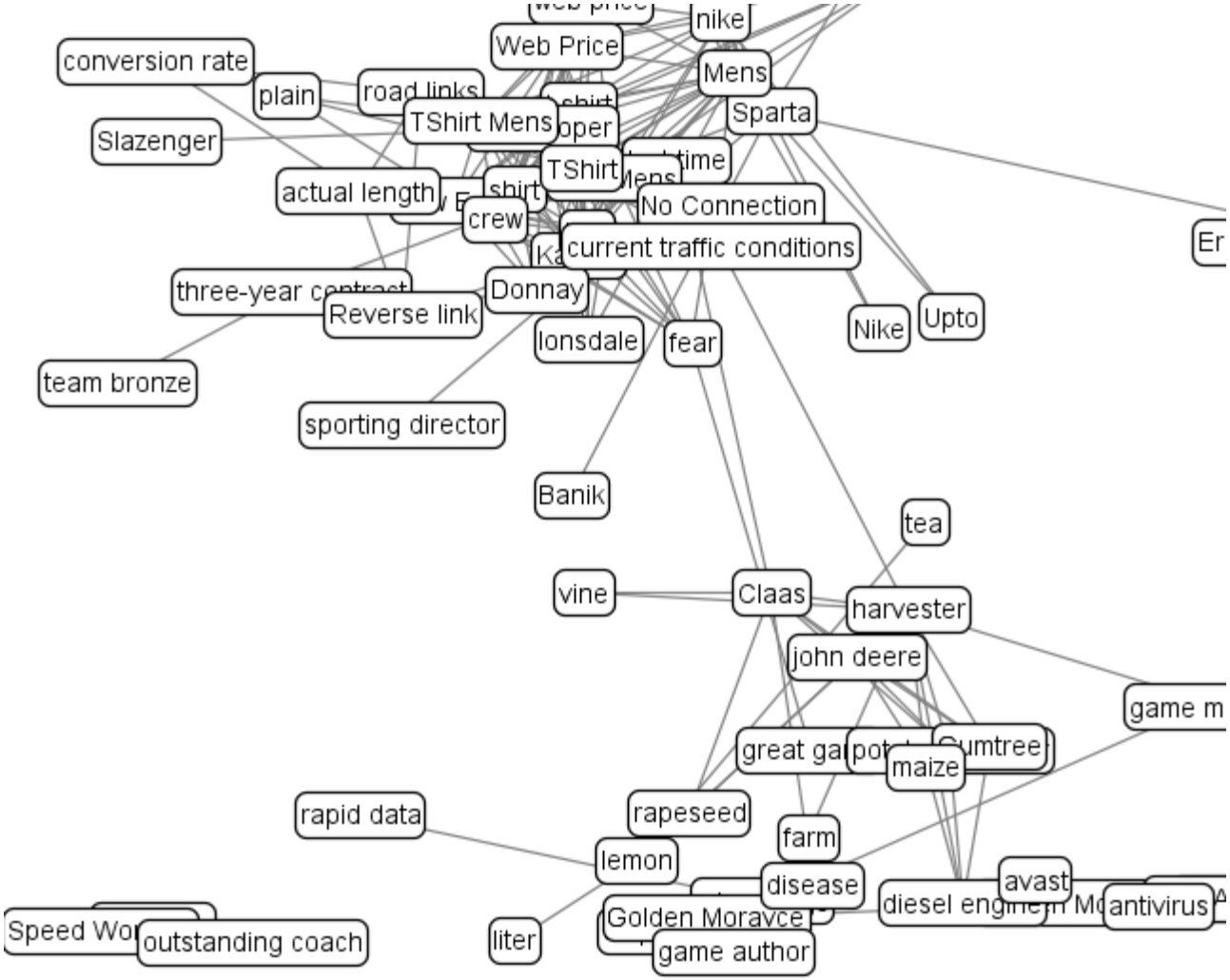


- Počet záznamov: 560 000
- Preprocesing: 13 100
- Odhodenie mail, facebook, youtube, rovnaké posebe idúce stránky: cca. 5000

Súvislosť dokumentov

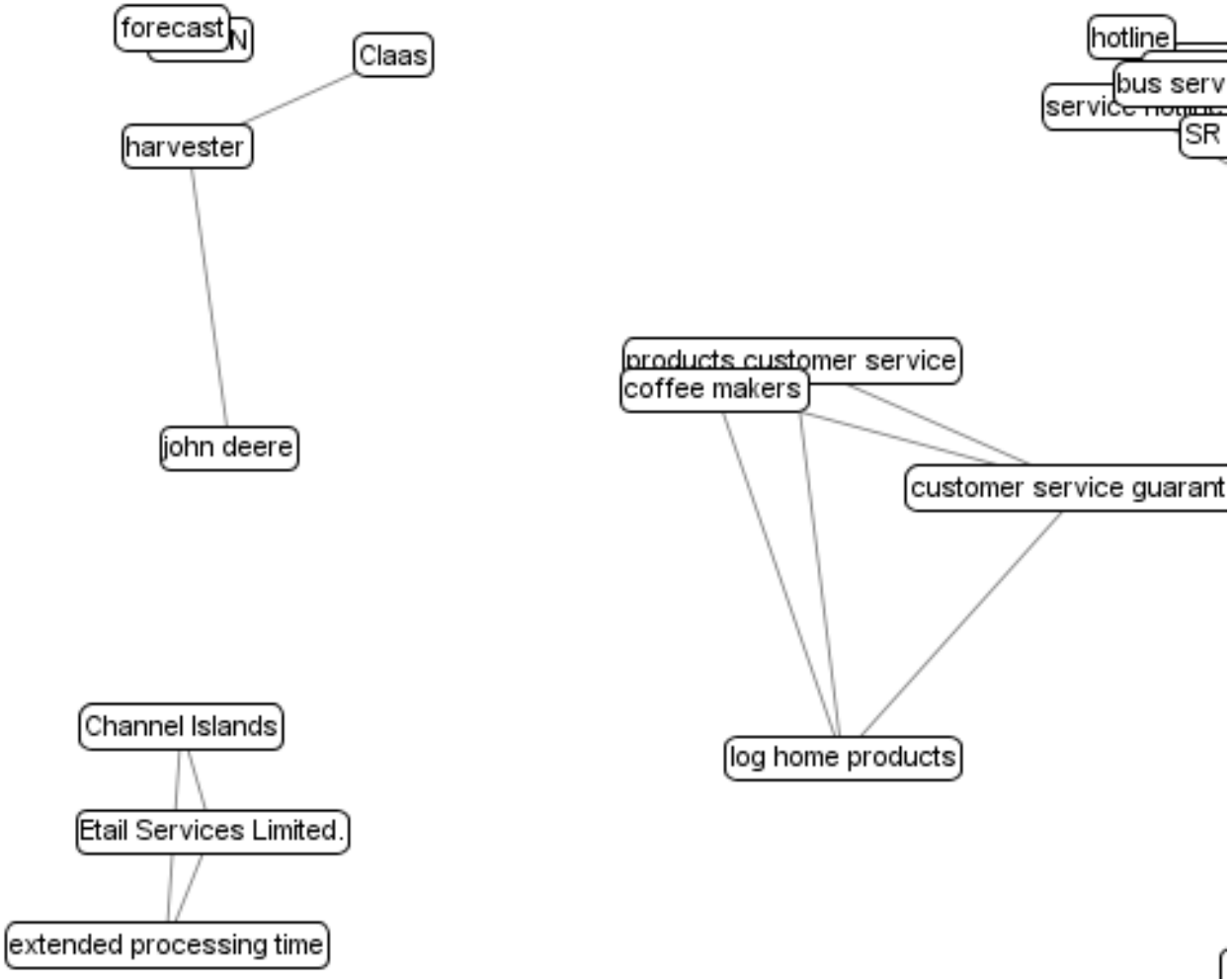


Súvislosť kľúčových slov



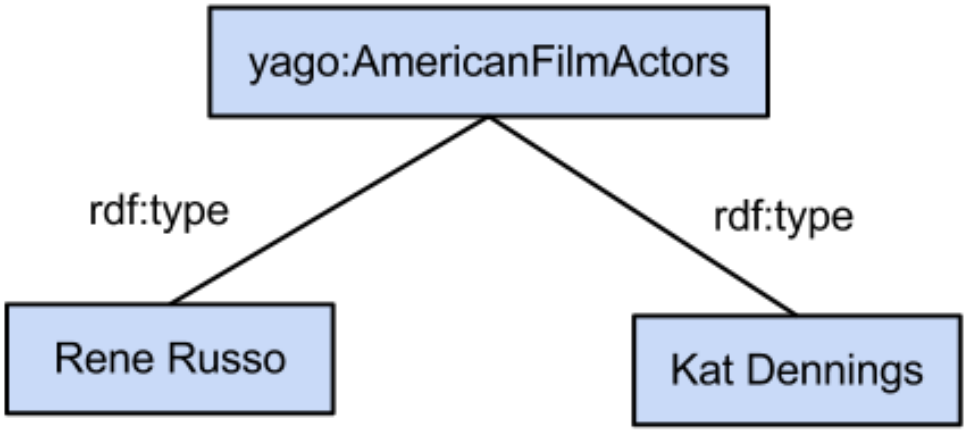
Obmedzenia

- Musí sa nachádzať aspoň 5 krát
- Hodnota súvislosti viac ako stanovený prah
- z 2 000 000 na 2 000


forecast
Claas
harvester
john deere
hotline
bus serv
SR
products customer service
coffee makers
customer service guarant
log home products
Channel Islands
Etail Services Limited.
extended processing time

Hľadanie ďalších vzťahov

- Mapovanie na Linked data
- Hierarchia
- Ďalšie pomocou Linked data

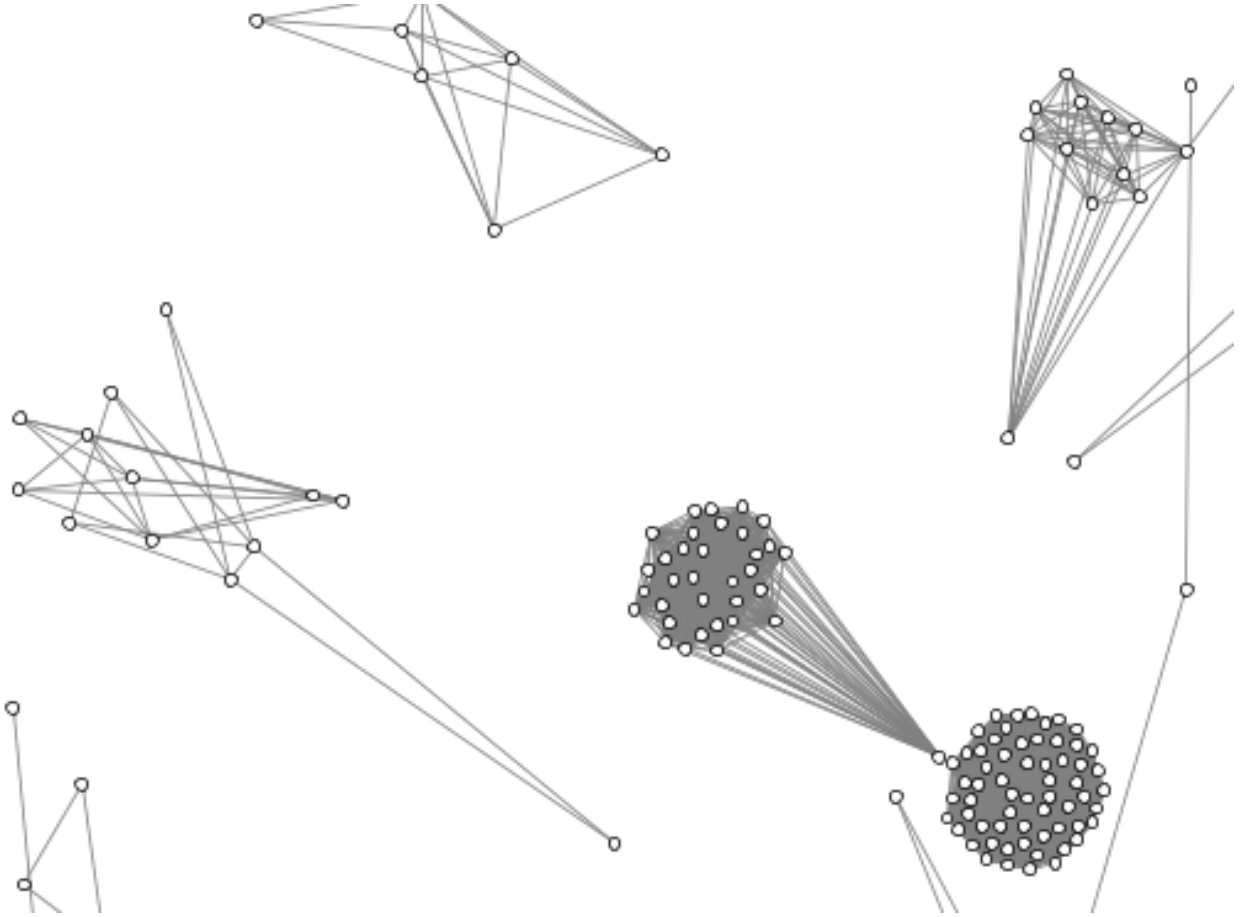


Problémy

- Z veľkých dát málo párov, málo pohľadov na jedno kľúčové slovo
- Vzťahy sú kontextovo závislé
- Vzťah väčšinou nie je očividný, záleží od kontextu (ostatných blízkych slov)

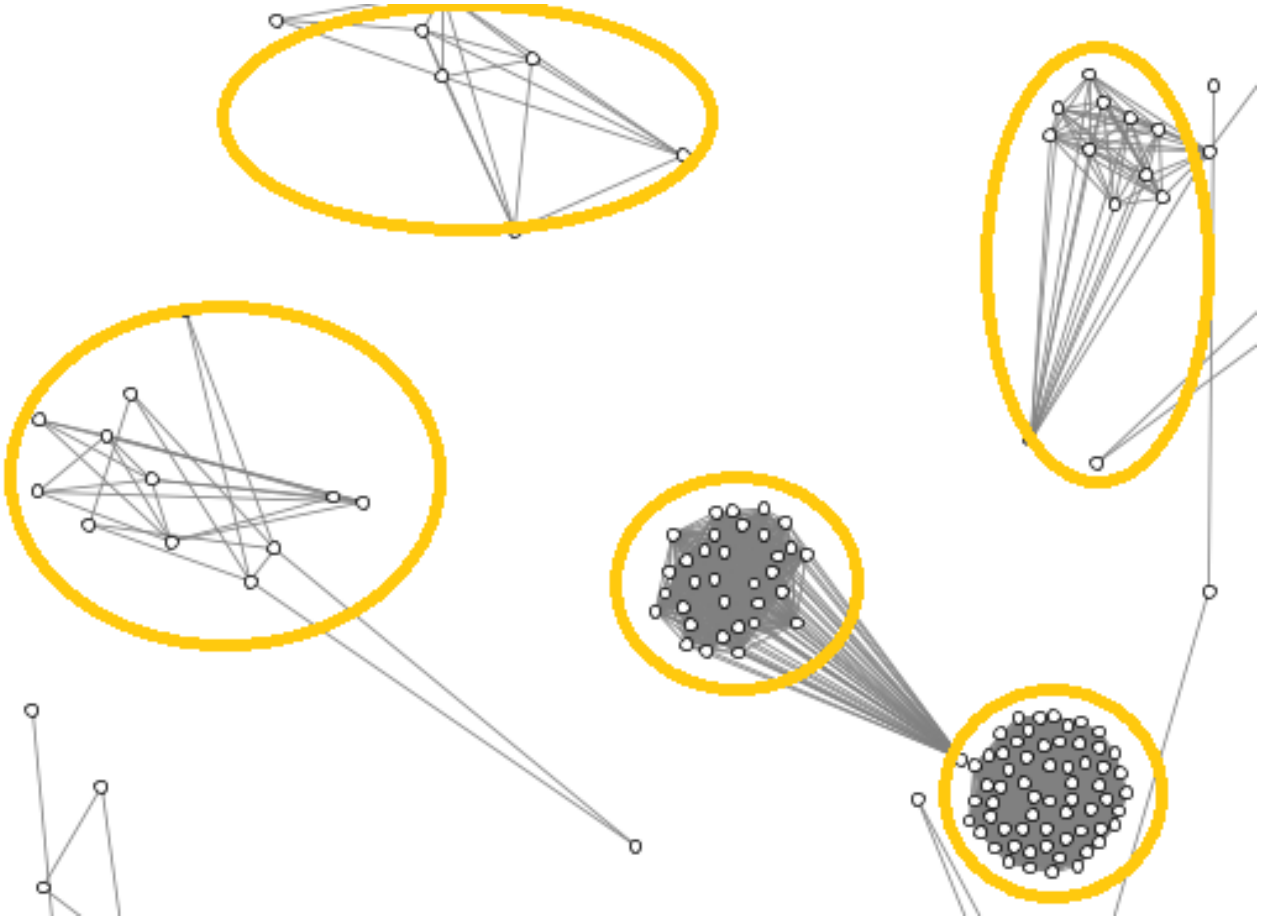
Návrh experimentu

- Treba pridať kontext – identifikácia klastrov



Návrh experimentu

- Treba pridať kontext – identifikácia klastrov



Návrh experimentu

- Tribune
- Copa América winger
- UEFA
- Marek Hamsik
- Lyon
- Saritamu

Návrh experimentu

- Tribune
- Copa América winger
- UEFA
- Marek Hamsik
- Lyon
- Saritamu
- Residence
- USB

Návrh experimentu

- Saritamu
- USB
- Copa América winger
- Tribune
- Lyon
- UEFA
- Residence
- Marek Hamsik

Návrh experimentu 2

- Saritamu
- USB
- Copa América winger
- Tribune
- Lyon
- UEFA
- Residence
- Marek Hamsik

- Vyhodenie 0 až size / 2 výrazov (granularita súvisl.)

Presnosť a pokrytie

- False positive – ak sa škrtne slovo z klastra
- False negative – ak sa neškrtne náhodné slovo

Kvalitatívny experiment - prototyp

- 4 ľudia
- 10 klastrov na základe postupností
- 10 klastrov na základe spoločného výskytu v dok.

- Feedback:
  - Bolo potrebné uviezť do problematiky
  - Veľa googlenia
  - Dohadovanie sa o význame slova
  - Vyznačenie istého slova
  - Vyhodenie celého klastra
  - Potreba motivácie

Väčší experiment

- Hra, bude sa bodovať
- Kto sa najviac priblíži k algoritmu vyhrá ☺
- Možnosť googlenia slova, viacerých slov
- Odkaz na wikipédiu, ak sa podarí

Návrh 2. experimentu

- Overenie vzťahov – presnejšie vzťahy
- Dvojica kľúčových slov
- Výber 0 až všetky z ponúknutých vzťahov
- Je možné určiť ďalšie vzťahy?
- Vymenovanie vzťahov, zadanie počtu