Radionica metoda i alata udaljenog čitanja u Galveju 2018

Mihailo Škorić

Seminar društva za jezičke resurse i tehnologije 20.12.2018.

Galway Training School 2018

https://www.distant-reading.net/events/galway-2018

- **COST Action**
  Distant Reading for European Literary History
- Od 5. do 7. decembra 2018. godine
- Nacionalni Univerzitet Irske u Galveju (NUIG)

- Dve paralelne radionice
  1. **Metodi i tehnike udaljenog čitanja prilagođeni na više evropskih jezika**
  2. Teoretski koncepti i njihova konfrontacija sa Kompjutacionim metodom

Metodi i alati udaljenog čitanja

- Textometrie

  http://textometrie.ens-lyon.fr/?lang=en
- DARIAH-DE Topics explorer

  https://github.com/DARIAH-DE/TopicsExplorer
- Palladio

  http://hdlab.stanford.edu/palladio/
- Gephi

  https://gephi.org/
- Stylo R

  https://sites.google.com/site/computationalstylistics

Postavka?

- Instalirati sve softvere na laptop
- Doneti laptop u Galvej
- Pratiti upustva predavača
- Isprobati softver
- Smisliti praktičnu primenu

Prvi dan

Wednesday, 5 December 2018: "Corpus management and analysis with TXM", taught by Serge Heiden (École normale supérieure de Lyon, France)

08:30-9:00 Installation fix session (for software needed at either of 3 workshop days)

09:00-09:30 Welcome (for both Training School workshops)

09:30-10:30 Corpus management and analysis with TXM 1

10:30-11:00 Coffee break

11:00-12:30 Corpus management and analysis with TXM 2

12:30-13:30 Lunch break

13:30-15:45 Corpus management and analysis with TXM 3

15:45-16:15 Coffee break

16:15-17:30 Lecture for participants of both Training School workshops by Professor Christof Schöch (Trier University, Germany)

Textometrie (TXM) Serge Heiden

- Softver za kreiranje i obradu korpusa
- Podržava TEI, pa samim tim i XML i TXT format

- Instalacija?

Šta TXM radi?

- Kreiranje novog korpusa na osnovu TXT, XML ili TEI datotoeka
- Uobličenje teskta korpusa u stranice knjige koje je prelistavaju
- Tokenizacija i obeležavanje tokena korišćenjem TreeTagger-a
- Kreiranje leksikona sa konkordancama za korpus, pri čemu se linije mogu grupisati prema atributima (npr: leksikon imenica, prideva)
- Eksportovanje leksikona u CSV formatu ili kopiranje tabela direktno u Excel ili drugi program za upravljanje tabelama
- Upoređivanje leksikona dva korpusa

Šta TXM radi?

- Kreiranje upita nad korpusom korišćenjem CQL-a
  - Primer 1:> korpus svih tokena čije leme sadrže love [enlemma=".*love.*"]
  - Primer 2: korpus svih imenica čije leme sadrže love [enpos="NN" & enlemma ="love"]
  - Primer 3: "government"%c – za ignorisanje kapitalizacije
  - Primer 4: "State"%d – ignorisanje dijakritika
  - Primer 5: "franc.*"%cd – kombinacija prethodna 2 primera
- Kreiranje podkorpusa na osnovu CQL upita
- Kreiranje nasumičnih particija nad nekim korpusom

Šta TXM radi?

- Pronalaženje specifičnosti korišćenjem ugrađenih R alata
- Na osnovu secifičnosti korpusa kreiranje tabele verovatnoća da se neki token nađe u nekom korpusu
- Proveravanja zajedničkog pojavljivanja, gde jedinice mogu biti reči ali i CQL upiti
- Grafički prikaz ovih izlaza i njihov eksport u SVG

Zaključak prvog dana

- TXM radi puno stvari, a preleteli smo ih za jedan dan
- Ako vam je potreban neki alat za obradu korpusa, moguće je da se nalazi u okviru TXM-a
- Moguće je da će TXM biti od koristi za ELTEC projekat

Drugi dan

Thursday, 6 December 2018: "Distributional Semantics + Network Analysis", taught by Steffen Pielström (Würzburg University, Germany) and Meliha Handzic (International Burch University, Bosnia and Herzegovina)

09:00-10:30 Topic Modeling 1 (Steffen Pielström)

10:30-11:00 Coffee break

11:00-12:30 Topic Modeling 2 (Steffen Pielström)

12:30-13:30 Lunch break

13:30-15:00 Using tools (Palladio) for network analysis (Meliha Handzic)

15:00-15:30 Coffee break

15:30-17:00 Using tools (Gephi) for network analysis (Meliha Handzic)

DARIAH-DE Topic Modeling Steffen Pielström

- Topic Modeling je odgovor na problem tematike i tematske klasifikacije teksta
- Ideja je da se kompjutacionim metodom naprave tj razdvoje ponavljajuće teme, u koje se tekstovi zatim razvrstavaju
- Fokusira se na više reči kako bi se izbegao problem „višetematskih“ dokumenata
- Zasnovan je na pretpostavci da se reči iz iste temečešće javljaju zajedno ili na maloj udaljenosti

Šta je Topics explorer i šta on radi?

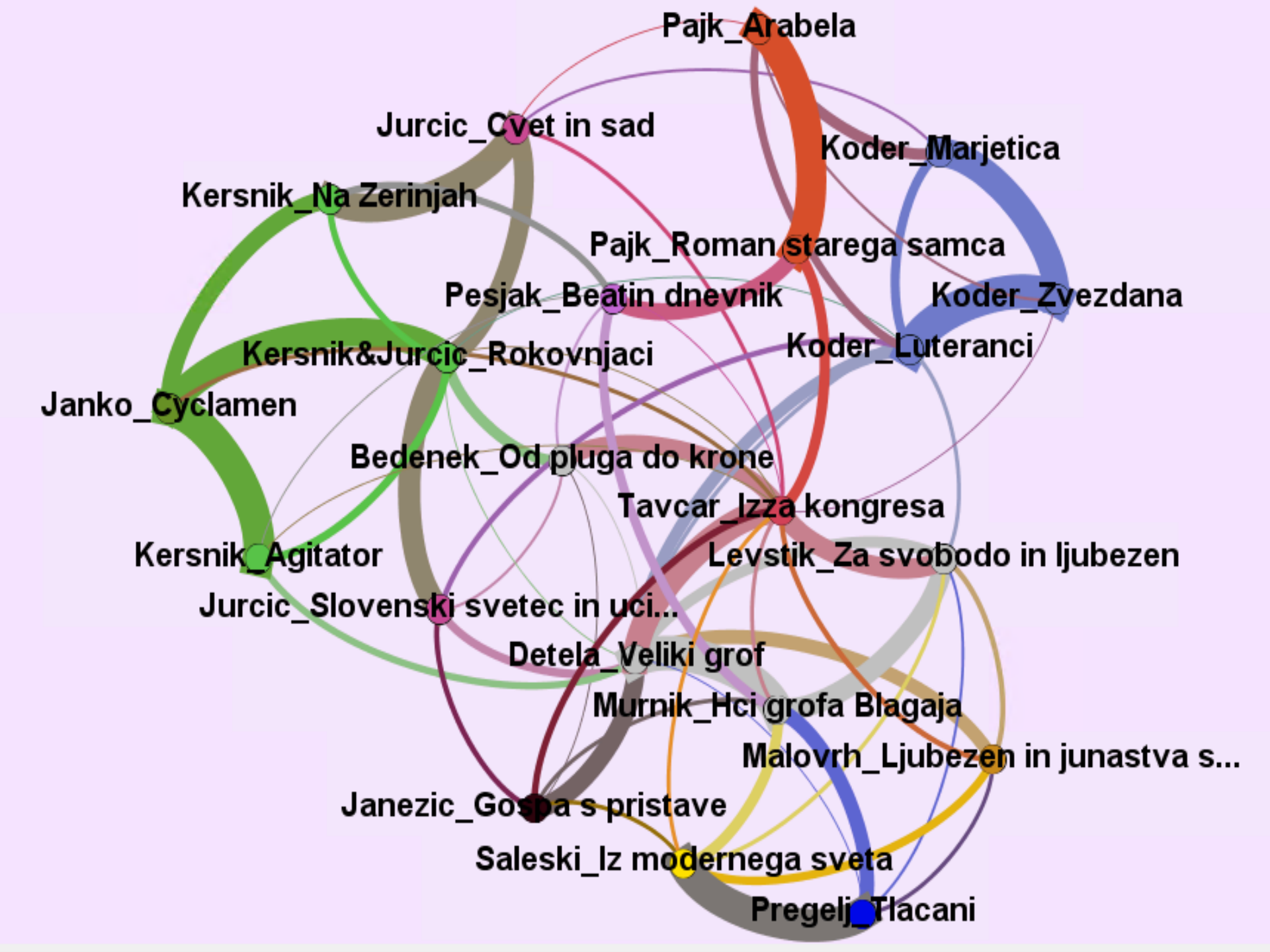
- Veoma jednostavan alat koji olakšava korisnicima pristup topic modeling-u
- Potrebno je samo:
  1. Odabrati tekstove za analizu
  2. Ukoliniti funkcionalne reči (2 načina)
  3. Odrediti broj tema i iteracija (jedino problematično)
  4. Proučiti rezultate i po potrebi izmeniti parametre
- Dobijaju se vrlo zanimljivi rezultati, poput toga koja dela se bave sličnim temama ili koji pisci uopšte

Palladio
Meliha Handžić

- Softver za grafičko predstavljanje podataka na mapi
- Moguće je prikazati
  1. Mesto
  2. Vreme
  3. Intezitet
  4. Veze između entiteta

Gephi
Meliha Handžić

- Softver za vizuelizaciju podataka korišćenjem grafova kompleksnih mreža
- Najveći je posao pripremiti Tabelu sa podacima
- U tabeli je potreno prikazati intezitet veze između svaka dva objekta, kako bi Gephi napravio graf
- Sve ostalo se svodi na podešavanja izgleda grafa


Pajk_Arabela
Jurcic_Cvet in sad
Koder_Marjetica
Kersnik_Na Zerinjah
Pajk_Roman starega samca
Pesjak_Beatin dnevnik
Koder_Zvezdana
Kersnik&Jurcic_Rokovnjaci
Koder_Luteranci
Janko_Cyclamen
Bedenek_Od pluga do krone
Tavcar_Izza kongresa
Kersnik_Agitator
Levstik_Za svobodo in ljubezen
Jurcic_Slovenski svetec in uci...
Detela_Veliki grof
Murnik_Hci grofa Blagaja
Malovrh_Ljubezen in junastva s...
Janezic_Gospa s pristave
Saleski_Iz modernega sveta
Pregelj_Tlacani

Zaključak drugog dana

- Topic modeling je zanimljiva ali nedorađena metoda
- Gephi je verovatno najbolji besplatni softver za kreiranje vizuelizacija relacionih podataka

Treći dan

Friday, 7 December 2018: "Stylometry with R", taught by Joanna Byszuk (Polish Academy of Sciences, Poland)

09:00-10:30 Getting started with stylometry

10:30-11:00 Coffee break

11:00-12:30 Stylometry for literary explorations

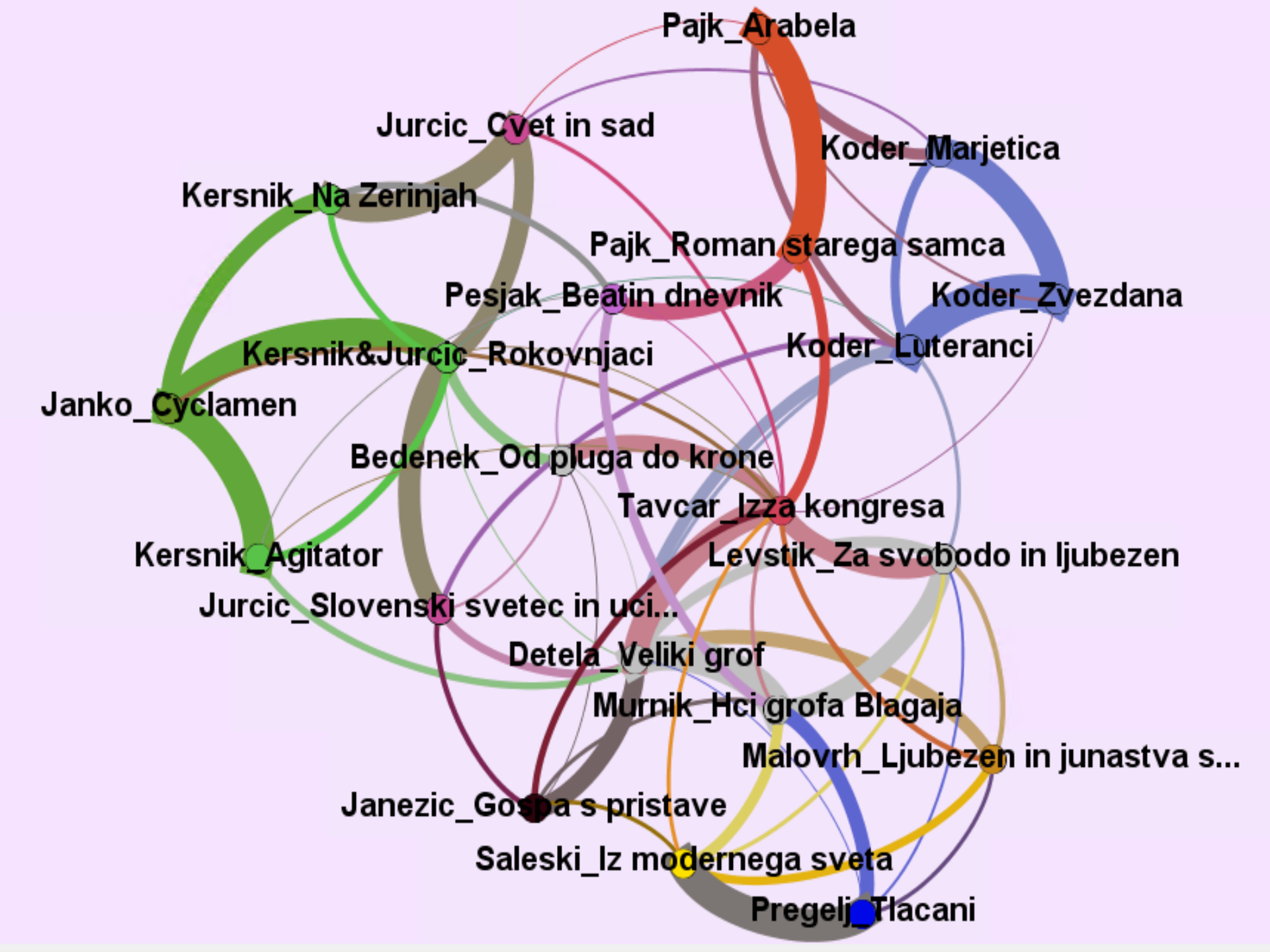
12:30-13:30 Lunch break

13:30-15:00 Authorship attribution

15:00 Closing (for both Training School workshops)

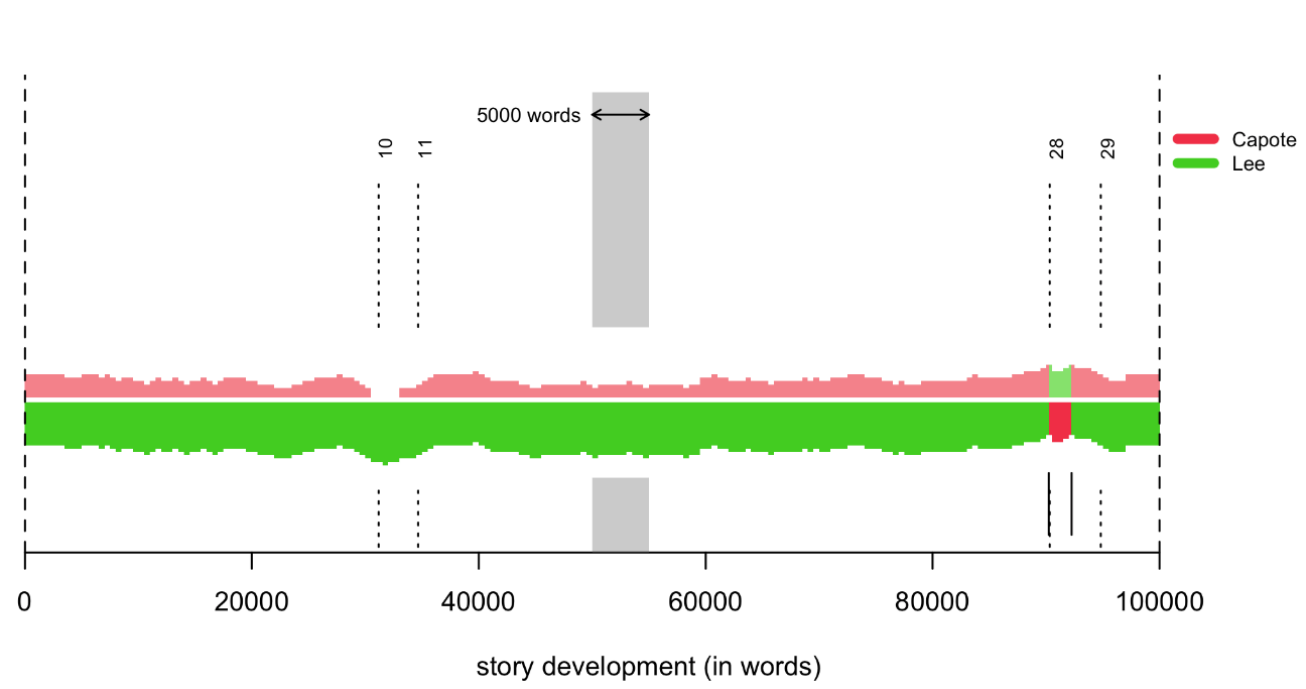
Stylometry paket za R

- Šta je ideja stilometrije?
- Najpre, određivanje autorstva
- Upoređivanje stila pisaca na osnovu kvantitativnih metoda
  1. Koje reči ili n-grami se upotrebljavaju
  2. Na kojoj razdaljini
- Procesiranje se vrši kroz R putem komandi iz biblioteke Stylo dok se vizuelizacija dobijenih veza vrši kroz Gephi


Pajk_Arabela
Jurcic_Cvet in sad
Koder_Marjetica
Kersnik_Na Zerinjah
Pajk_Roman starega samca
Pesjak_Beatin dnevnik
Koder_Zvezdana
Kersnik&Jurcic_Rokovnjaci
Koder_Luteranci
Janko_Cyclamen
Bedenek_Od pluga do krone
Tavcar_Izza kongresa
Kersnik_Agitator
Levstik_Za svobodo in ljubezen
Jurcic_Slovenski svetec in uci...
Detela_Veliki grof
Murnik_Hci grofa Blagaja
Malovrh_Ljubezen in junastva s...
Janezic_Gospa s pristave
Saleski_Iz modernega sveta
Pregelj_Tlacani

Stylometry paket za R

- Stil je takođe moguće trenirati i potom testirati
- Poznati primer: Go set a Watchman
  Da li je knjigu napisao Capote ili Harper Lee?



Stylometry paket za R

- Osim stilova kojim se služe autori, može se proučavati i sintaktički stil, na primer u različitim jezicima

Zaključak trećeg dana

- Stylo paket za R i stilometrija uopšte imaju različite mogućnosti i moguće je da je potrebna samo dobra ideja kako bi se dobili zanimljivi rezultati
- Stylo je veoma koristan za analizu u vizuelizaciju celokupne književosti, a može biti koristan i u neke druge svrhe