Prepoznavanje autora knjige upotrebom perzistentne homologije

Aleksandra Ružić

Petar Zečević

Sadržaj

Uvod

- Analiza teksta - obično pomoću "bag of words" modela
- Mana - ne uzima u obzir redosled reči i stil pisanja
- TDA (Topological Data Analysis)
  - Za visokodimenzionalne podatke sa dosta šumova
  - Hvata oblike u podacima
  - Uloge: klasterovanje, smanjenje dimenzionalnosti
- Bavimo se upotrebom TDA u prepoznavanju autora

Algoritam

- Gledaju se odnosi između likova
- Faze:

1. Izdvajanje pozicija pojavljivanja likova
2. Merenje rastojanja između likova u romanu
3. Pravljenje dijagrama i distanci između romana
4. Predikcija autora

Izdvajanje pozicija pojavljivanja likova

- StanfordCore NLP - tokenizacija
- Pamti se ime lika i redni broj tokena (pozicija lika)
- Uzima se 10 najčešćih likova
- Ne uzimaju se u obzir koreference

Merenje rastojanja između likova u romanu

1. Prave se vektori pozicija istih dužina

$$I = (i_1, i_2, ..., i_n)$$

$$J = (j_1, j_2, ..., j_m), m > n$$

$$J^* = (j_1^*, j_2^*, ..., j_n^*)$$

$$j_x^* = \min_{j_y} |j_y - i_x|, \forall x \in \{1, ..., n\}, y \in \{1, ..., m\}$$

$$j_x^* \neq j_t^*, \forall t \in \{1, ..., x-1\}$$

2. Vaserštajnova distanca između vektora

Vaserštajnova distanca

- Earth mover's distance
- Cena potrebna da se jedna raspodela verovatnoće pretvori u drugu
- Koristi se za traženje slike po sadržaju (računa se rastojanje između histograma boja dve slike)

Merenje rastojanja između likova u romanu

- Distanca između dva lika:

$$Distance_t(A, B) = WD_{0.5}(I^{(1+t)}, J^{(1+t)})$$

  - I i J normalizovani vektori pozicija
  - Prave se 3 matrice distanci (za $t \in \{0, 0.1, -0.1\}$)
  - 0.5 je red Vaserštajnove distance

Pravljenje dijagrama i distanci između romana

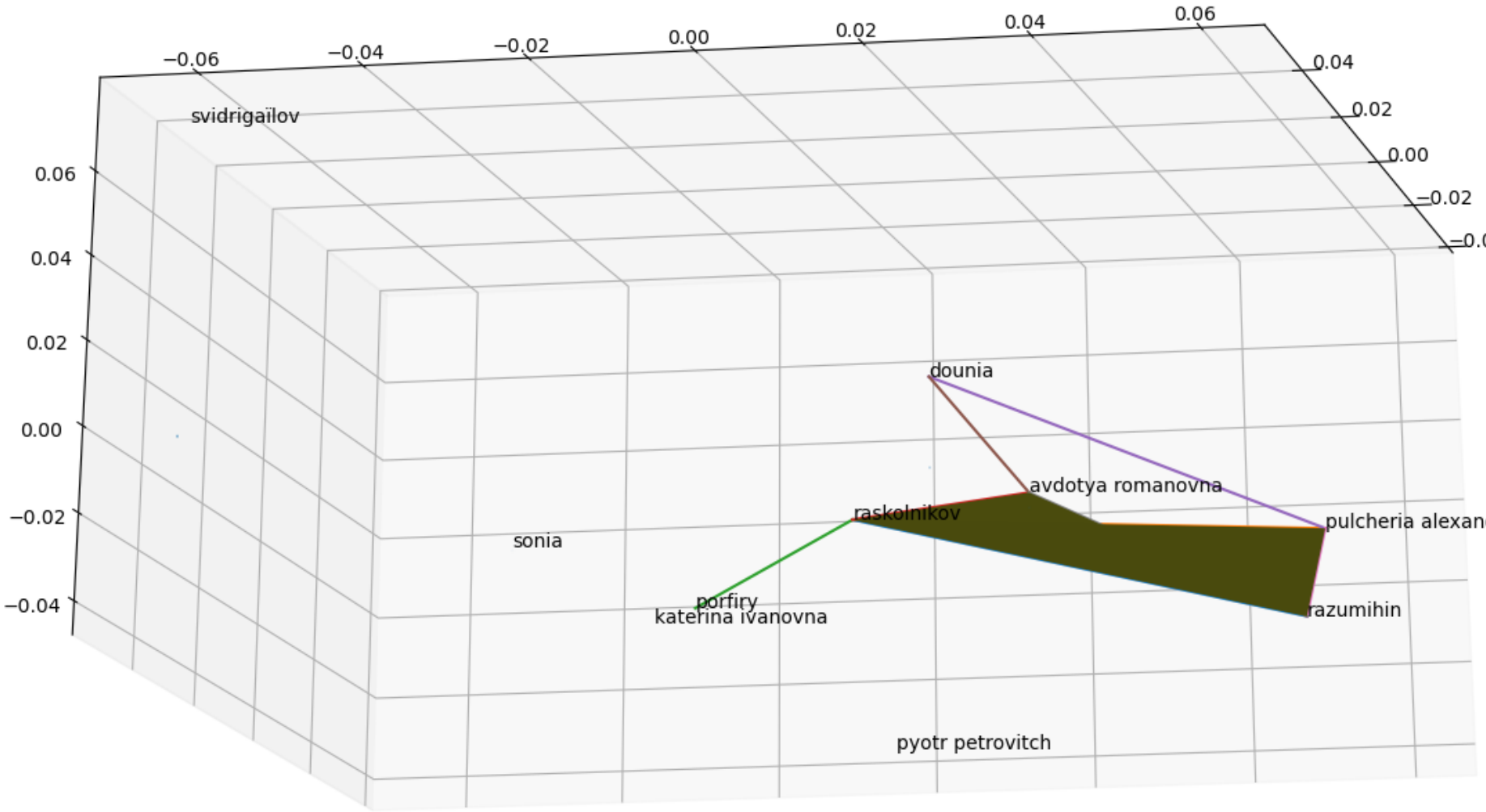
- Prave se Ripsovi dijagrami od matrica distanci između likova
- Pravi se matrica rastojanja između knjiga:

$$Distance_t(X, Y) = WD\{PD_t^0(X), PD_t^0(Y)\} + WD\{PD_t^1(X), PD_t^1(Y)\}$$

$$Distance(X, Y) = \{ \sum_{t \in \{-0.1, 0, 0.1\}} Distance_t(X, Y)^2 \}^{\frac{1}{2}}$$

- $PD_t^0(X)$ - komponente u Ripsovom dijagramu za knjigu X
- $PD_t^1(X)$ - rupe u Ripsovom dijagramu za knjigu X

Simplicijalni kompleks likova iz Zločina i Kazne



Dijagrami

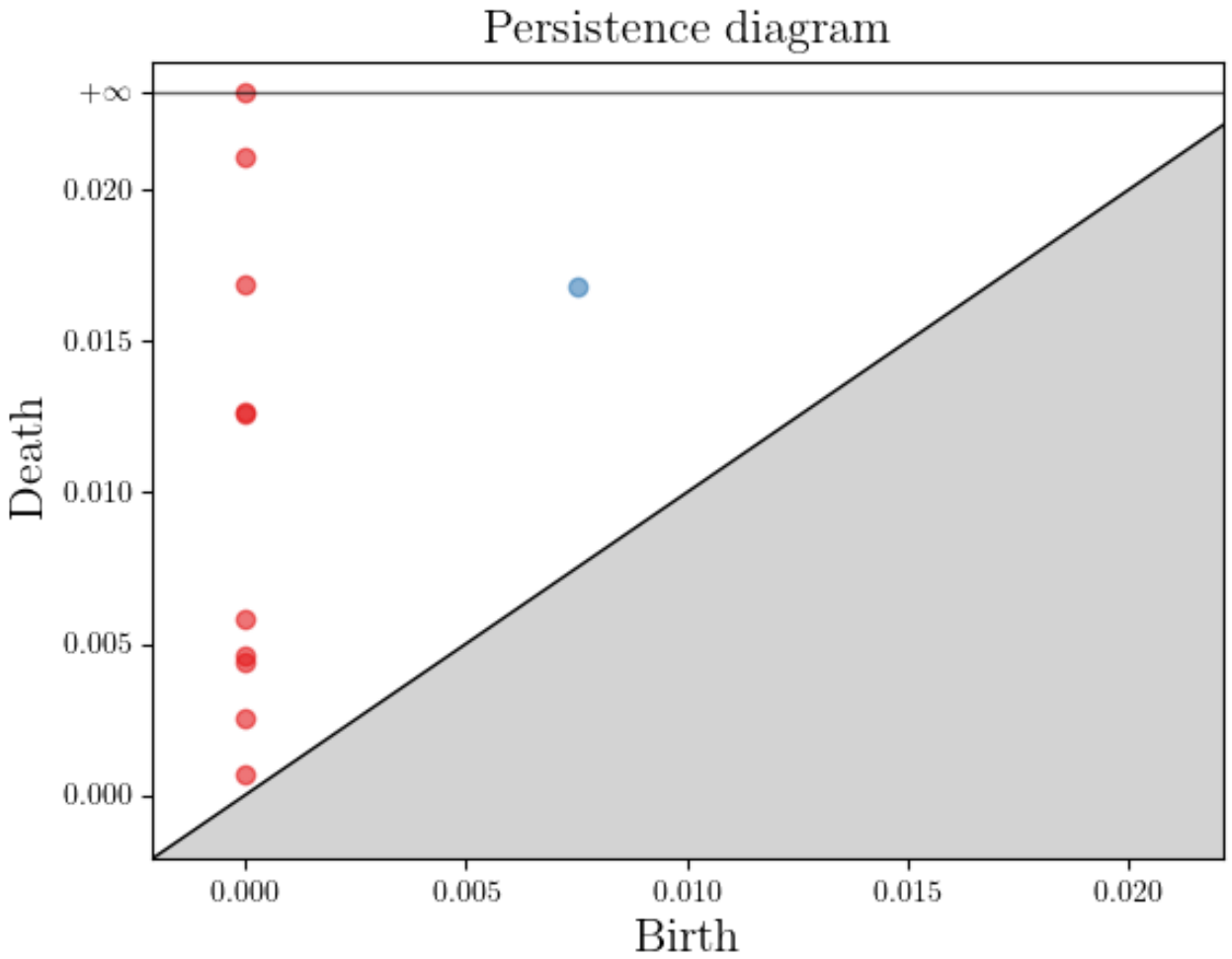


Figure: Fyodor Dostoyevski Crime and Punishment, $t = 0$

Predikcije autora

- Binarni KNN sa kros-validacijom
  - Za svaka dva autora radi se KNN nad njihovim knjigama
  - Jednak broj knjiga za oba autora
  - Radi se 200 puta i računa prosečna preciznost

Razlike u implementaciji

- Originalna implementacija
  - Koreference
  - Imena sa više od dve reči
  - Normalizacija pozicija pojavljivanja likova dužinom knjige
- Implementacija seminarskog
  - Normalizacija pozicija pojavljivanja likova dužinom vektora
  - Normalizacja dijagrama dužinom knjige
  - Foldovi su 5-10; K za KNN je 3-5

Rezultati sa 8 foldova i 4 komšije

	Charles Dickens	Emilie Zola	Fyodor Dos-toyevski	Jane Austen	Mark Twain	Walter Scott
Charles Dickens	1	0.81	0.56	0.91	0.63	0.6
Emilie Zola	0.81	1	0.69	0.55	0.8	0.8
Fyodor Dostoyevski	0.56	0.69	1	0.65	0.49	0.6
Jane Austen	0.91	0.55	0.65	1	0.67	0.92
Mark Twain	0.63	0.8	0.49	0.67	1	0.5
Walter Scott	0.60	0.8	0.6	0.92	0.5	1

Normalizacija dužinom vektora; 7 foldova i 4 komšija

	Charles Dickens	Emilie Zola	Fyodor Dos-toyevski	Jane Austen	Mark Twain	Walter Scott
Charles Dickens	1	0.66	0.41	0.83	0.72	0.51
Emilie Zola	0.66	1	0.6	0.66	0.74	0.82
Fyodor Dostoyevski	0.41	0.6	1	0.88	0.43	0.57
Jane Austen	0.83	0.66	0.88	1	0.75	0.98
Mark Twain	0.72	0.74	0.43	0.75	1	0.57
Walter Scott	0.51	0.82	0.57	0.98	0.57	1

Normalizacija dužinom vektora i knjige; 7 foldova i 3 komšija

	Charles Dickens	Emilie Zola	Fyodor Dos-toyevski	Jane Austen	Mark Twain	Walter Scott
Charles Dickens	1	0.69	0.43	0.72	0.52	0.62
Emilie Zola	0.69	1	0.61	0.79	0.84	0.78
Fyodor Dostoyevski	0.43	0.61	1	0.5	0.53	0.68
Jane Austen	0.72	0.79	0.5	1	0.89	0.98
Mark Twain	0.52	0.84	0.53	0.89	1	0.84
Walter Scott	0.62	0.78	0.68	0.98	0.84	1

Literatura

- Topological Signature of 19th Century Novelists: Persistent Homology in Text Mining, Shafie Gholizadeh, Armin Seyeditabari and Wlodek Zadrozny, https://s3.amazonaws.com/cdn.ayasdi.com/wp-content/uploads/2018/10/19143440/BDCC-02-00033.pdf