Gručenje z omejitvami na podlagi besedil in grafov pri razporejanju akademskih člankov

Tadej Škvorc

Fakulteta za računalništvo in informatiko, Univerza v Ljubljani

2018

Opis problema

- Znanstvene konference so pomembne za širjenje znanosti
- Lahko so zelo velike – problemi z organizacijo
- Predvsem izdelava urnika konference

08:30 – 08:40	Welcome						
08:40 – 09:40	Invited Talk: Susan Athey, **Causal Inference for Policy Evaluation**						
09:40 – 10:20	Break						
10:20 – 11:10	Neural Networks and Deep Learning	Reinforcement Learning	Optimization (Continuous)	Online Learning	Clustering	Bayesian Nonparametric Methods	Matrix Factorization / Neuroscience Applications
11:11 – 11:30	Break						
11:30 – 12:20	Neural Networks and Deep Learning	Reinforcement Learning	Optimization (Continuous)	Online Learning	Clustering	Bayesian Nonparametric Methods	Matrix Factorization / Neuroscience Applications
	Lunch – on your own						
02:00 – 02:50	Neural Networks and Deep Learning	Optimization / Online Learning	Machine Learning Applications	Matrix Factorization and Related Topics	Bandit Problems	Graphical Models	Transfer Learning / Learning Theory
02:51 – 03:10	Break						
03:10 – 04:00	Neural Networks and Deep Learning	Optimization / Online Learning	Machine Learning Applications	Matrix Factorization and Related Topics	Bandit Problems	Graphical Models	Transfer Learning / Learning Theory
04:01 – 04:15	Break						
04:15 – 05:05	Neural Networks and Deep Learning I	Neural Networks and Deep Learning II (Computer Vision)	Approximate Inference	Metric and Manifold Learning / Kernel Methods	Statistical Learning Theory	Structured Prediction / Monte Carlo Methods	Online Learning

Figure: Del urnika konference ICML 2016.

Rešitev

- Aplikacija za podporo organizacije konferenc
- Podobne že obstajajo – EasyChair, OpenConf ...
- Novost – samodejno razvrščanje člankov v urnik

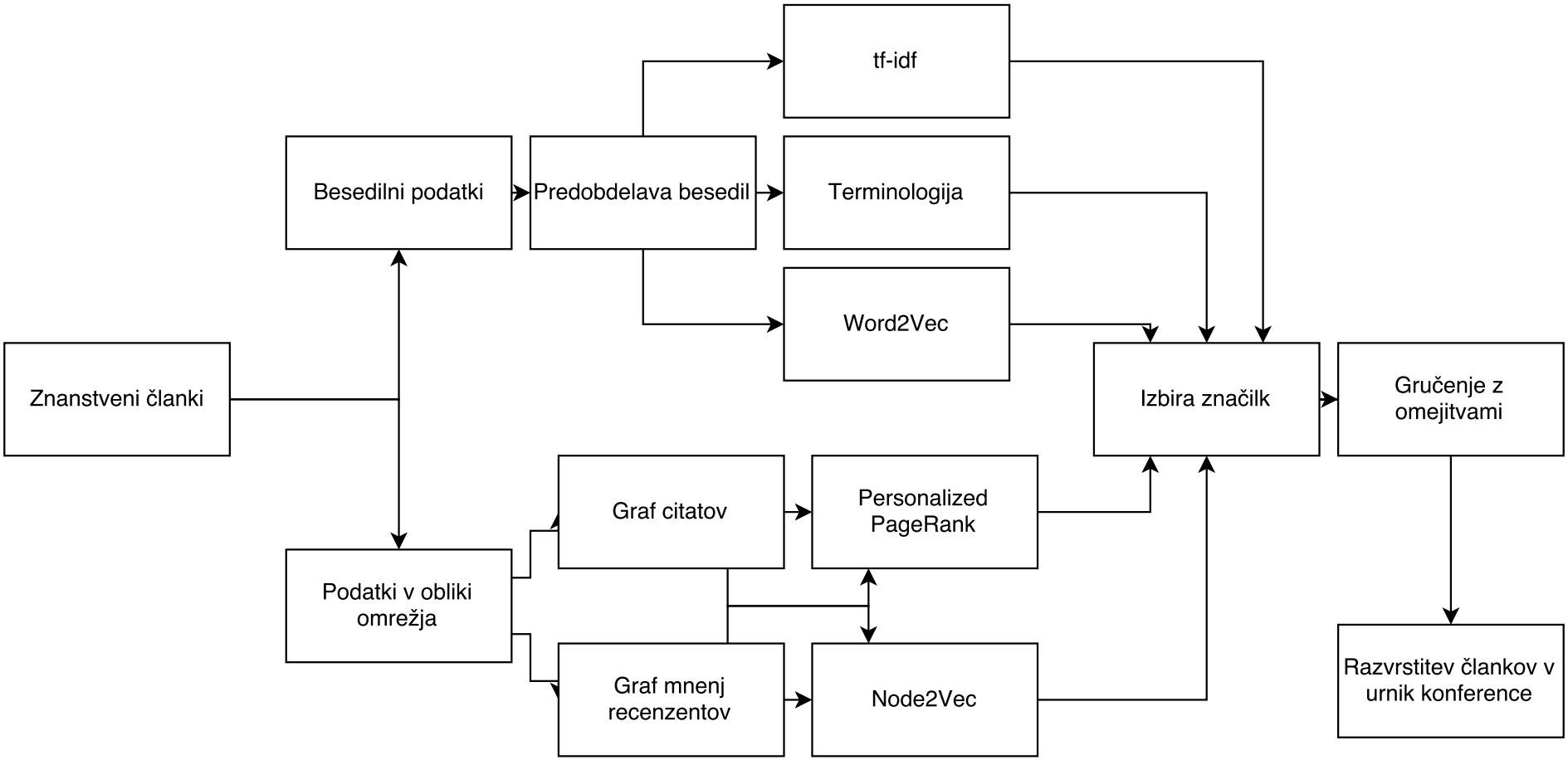
Struktura urnika konference

- Osnovna enota je seja
- Vsaka seja ima svojo tematiko in vsebuje članke s to tematiko
- Hkrati lahko teče več sej
- Samodejno razvrščanje:
  - Uporabnik ročno definira strukturo urnika
  - Razporeditev člankov je samodejna

Glavne metode

- Obdelava naravnega jezika in analiza omrežij za iskanje podobnih člankov
- Gručenje z omejitvami za razporeditev člankov v urnik

Struktura rešitve



Besedilni podatki

- Metode obdelave naravnega jezika
- Predstavitev z vrečo besed, uteženo s tf-idf
- Luščenje terminologije
- Vektorske vložitve besed

Vreča besed in tf-idf

Table: Predstavitev z vrečo besed za stavke *"Žena sedi na stolu in bere časopis"*, *"Mož sedi na stolu in gleda televizijo"* in *"Riba plava v morju"*.

mož	žena	sedi	na	stolu	in	bere	časopis	gleda	televizijo	riba	plava	v	morju
0	1	1	1	1	1	1	1	0	0	0	0	0	0
1	0	1	1	1	1	0	0	1	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0	1	1	1	1

Luščenje terminologije

- Izberi kandidate glede na besedne vrste
- Oceni kandidate glede na:
  - Relativno frekvenco pojavitve kandidata v znanstvenih člankih
  - Relativno frekvenco pojavitve kandidata v korpusu splošnih besedil (COCA)
  - Številom člankov, v katerih se kandidat pojavi
- Odstrani nekoristne terminološke izraze

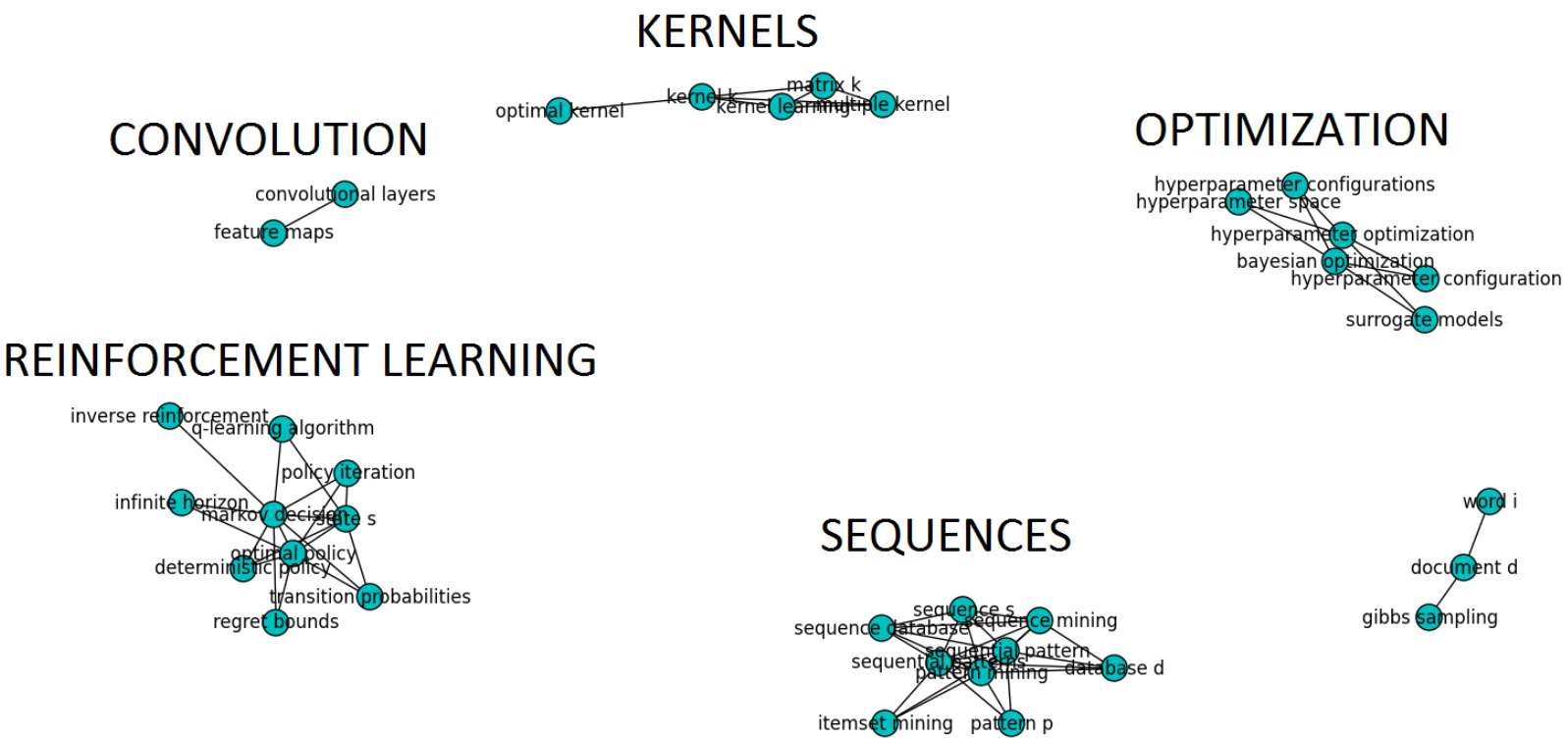
Oblika fraze	Primer fraze
Samostalnik	Kernel
Samostalnik, samostalnik	Machine learning
Pridevnik, samostalnik	Neural network
Prid., prid., sam.	Deep neural network
Prid., sam., sam.	Deep learning theory
Sam., sam., sam.	Column subset selection

Vpliv splošnega korpusa

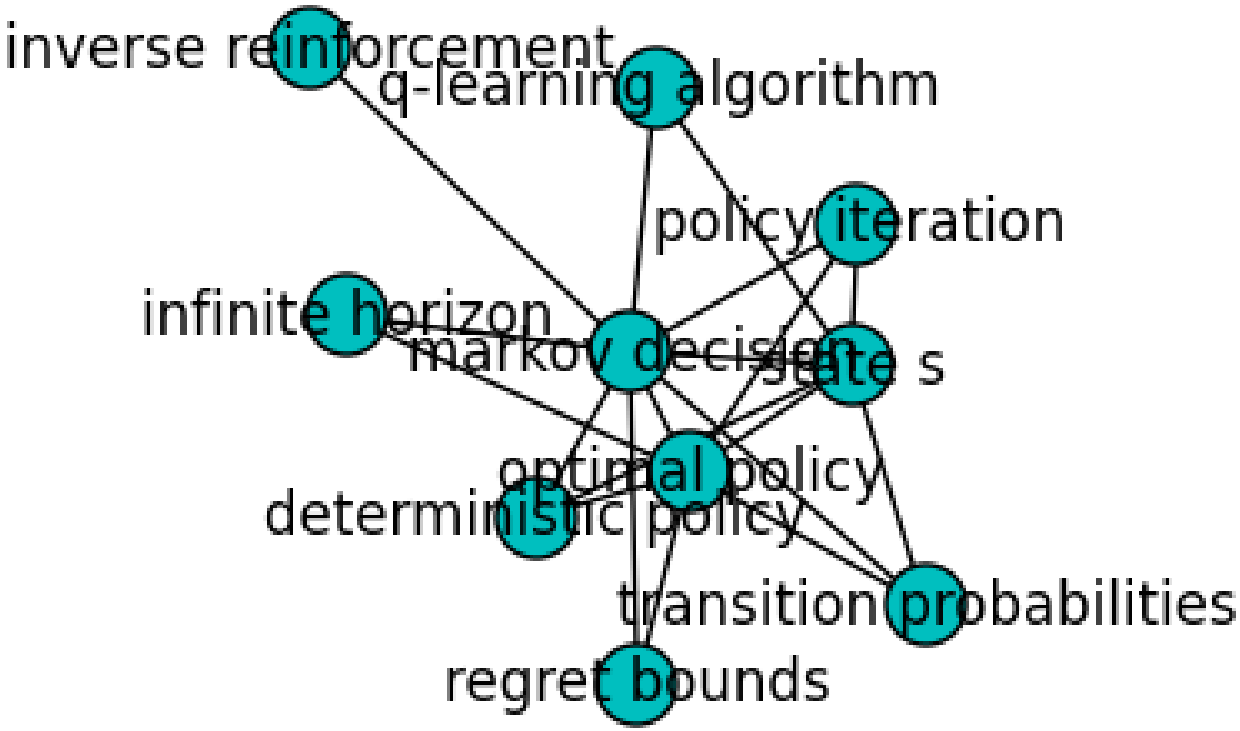
Table: Primeri kandidatov za terminološke izraze, ki se pojavijo v korpusu s splošnim besediščem.

Več kot 100 pojavitev	0 do 100 pojavitev	0 pojavitev
information networks	many contexts	proximal map
detection system	additional challenges	scale-free distribution
fast pace	same row	optimistic policy
natural conditions	rate hop	optimal algorithm
other goals	real student	incoherence requirement
	principle components	normalized weighing

Omrežje terminologije



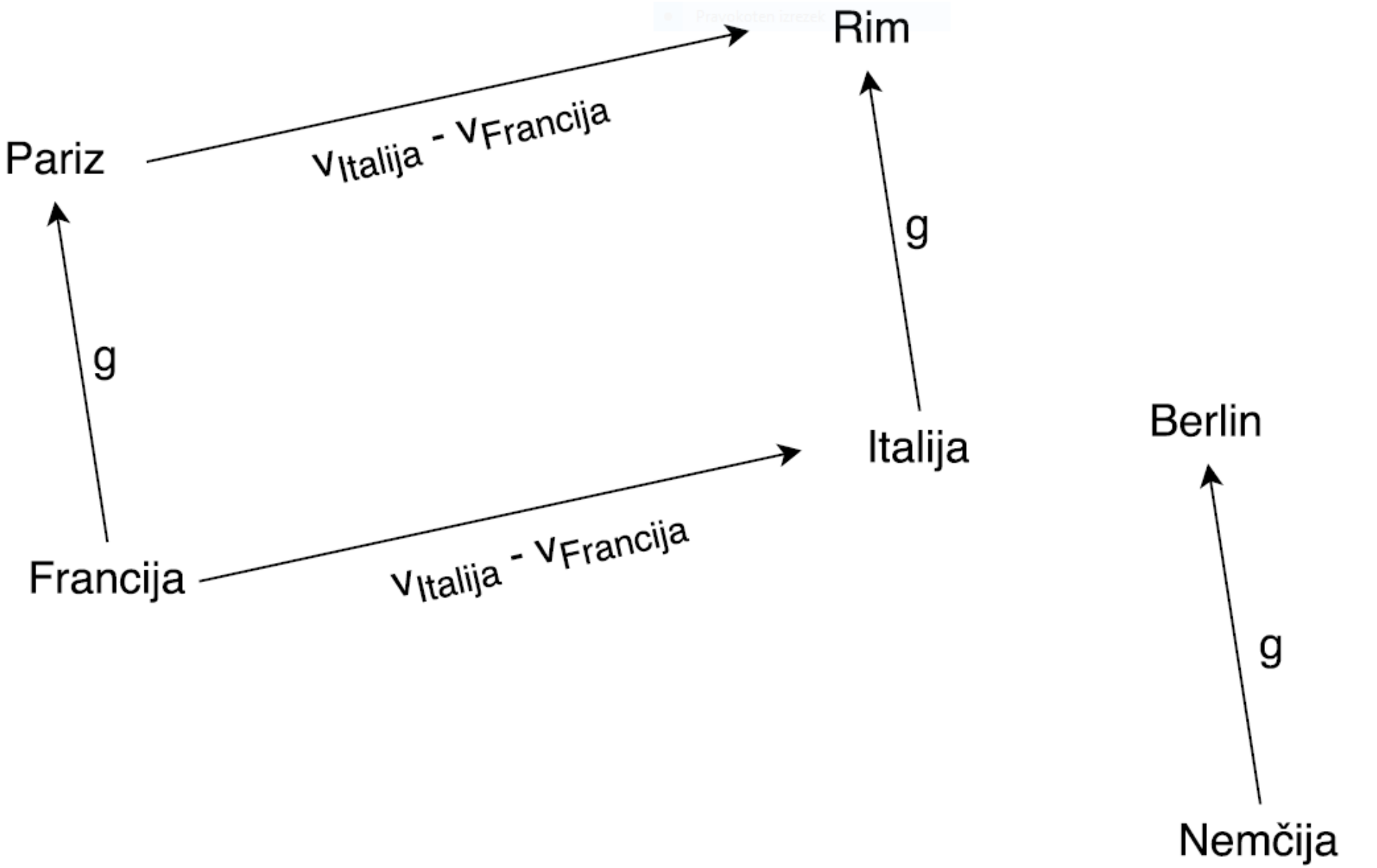
REINFORCEMENT LEARNING



Vektorske vložitve besed

- Besede med seboj niso neodvisne
- Želimo metodo, ki jih ustrezno preslika v vektorje glede na odvisnosti
- To storijo vektorske vložitve
- Metodi CBOW in Skip-grams (implementirani kot word2vec)

Vektorske vložitve besed



Uporaba vektorskih vložitev besed

- Povprečje vseh vektorjev kot vektor dokumenta
- Metoda doc2vec

Podatki iz omrežij

- 2 tipa omrežij:
  - Omrežje bibliografske povezanosti
  - Omrežje mnenj recenzentov
- Tudi omrežje terminologije
- Potrebno pretvoriti v vektorsko obliko

Detekcija skupnosti

- 2 metodi:
  - Personalized PageRank
  - Node2vec

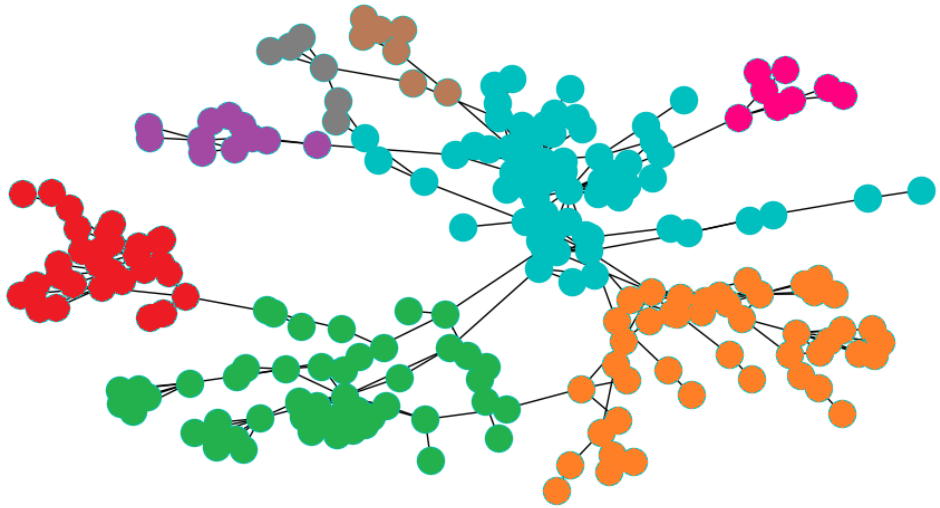


Figure: Primer skupnosti v omrežju

Izbira značilk

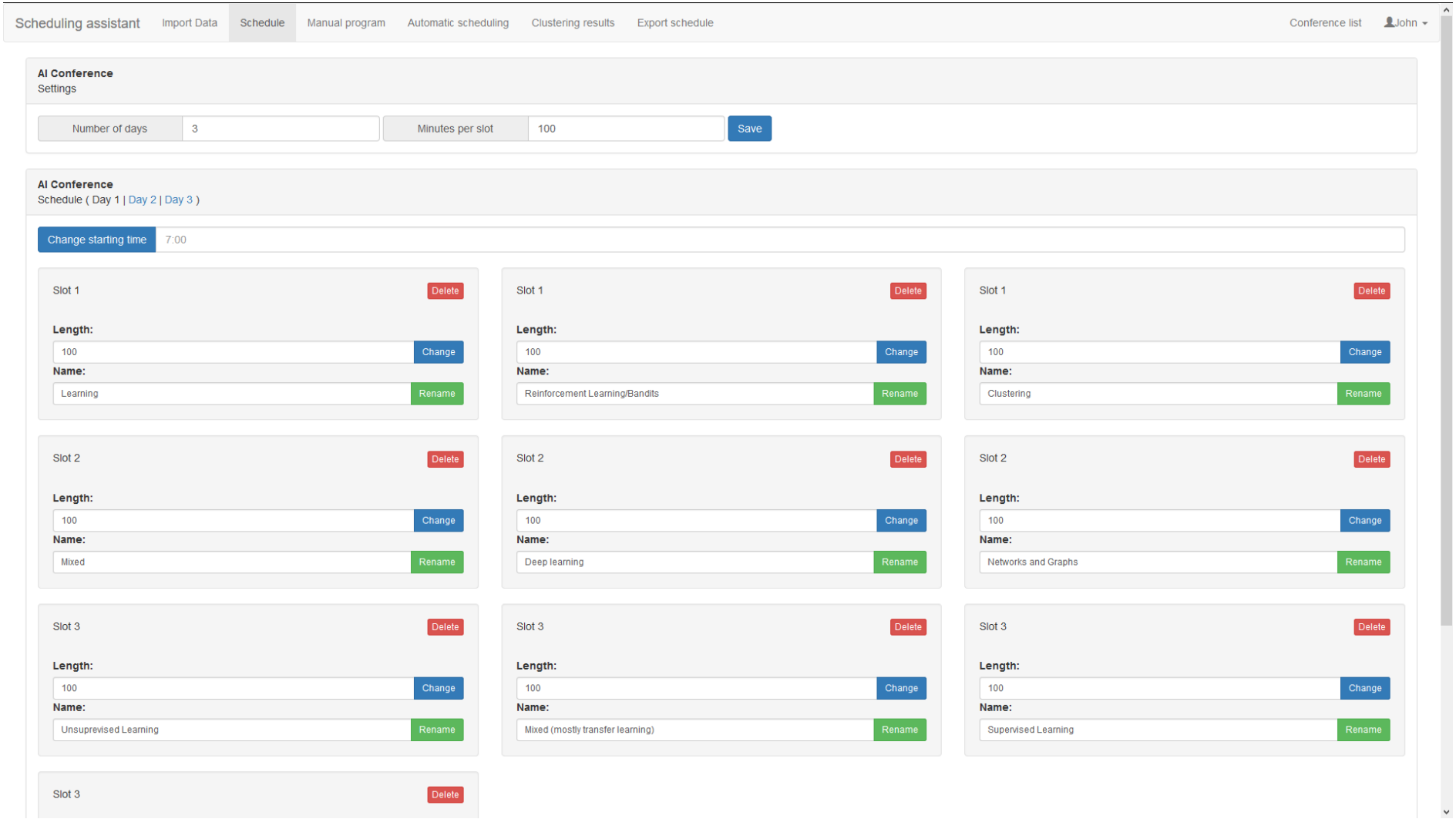
- Metoda ANOVA in ReliefF
- ANOVA je hitra, dobri rezultati
- ReliefF zazna odvisnosti med značilkami, počasen, ni izboljšal rezultatov

Gručenje z omejitvami

- Modifikacija metode k-voditeljev
- Upošteva velikosti sej urnika
- 2 koraka:
  - Inicializacija: podobno kot metoda k-voditeljev, vendar upošteva velikosti gruč
  - Optimizacija: zamenjuj pare primerov tako, da se zmanjša vsota razdalj do centroidov

Spletna aplikacija

- Uporabnik se prijavi
- Izbere/ustvari konferenco
- Definira strukturo urnika
- Uvozi članke
- Aplikacija razvrsti članke v urnik


Scheduling assistant
Import Data
Schedule
Manual program
Automatic scheduling
Clustering results
Export schedule
Conference list
John
AI Conference
Settings
Number of days
3
Minutes per slot
100
Save
AI Conference
Schedule (Day 1 | Day 2 | Day 3)
Change starting time
7:00
Slot 1
Delete
Length:
100
Change
Name:
Learning
Rename
Slot 1
Delete
Length:
100
Change
Name:
Reinforcement Learning/Bandits
Rename
Slot 1
Delete
Length:
100
Change
Name:
Clustering
Rename
Slot 2
Delete
Length:
100
Change
Name:
Mixed
Rename
Slot 2
Delete
Length:
100
Change
Name:
Deep learning
Rename
Slot 2
Delete
Length:
100
Change
Name:
Networks and Graphs
Rename
Slot 3
Delete
Length:
100
Change
Name:
Unsuprevised Learning
Rename
Slot 3
Delete
Length:
100
Change
Name:
Mixed (mostly transfer learning)
Rename
Slot 3
Delete
Length:
100
Change
Name:
Supervised Learning
Rename
Slot 3
Delete

Evalvacija

- Na testni podatkovni bazi ročno označenih člankov
- Na člankih konference ECML-PKDD 2017

Konferenca	Število člankov
ECML-PKDD 2015	63
ECML-PKDD 2016	90
ICML 2016	262
SIGKKD, AISTATS, NIPS	200
Skupaj	615

Table: Članki uporabljeni za testiranje metode

Testiranje na testni podatkovni bazi

- Klasifikacija z 10-kratnim prečnim preverjanjem
- Logistična regresija, naključna drevesa in metoda podpornih vektorjev

Uporabljene značilke	Klasifikacijska točnost
Vreča besed s tf-idf	0.461
Omrežje bibliografske povezanosti	0.405
Terminologija	0.407
doc2vec	0.083
Graf recenzentov z node2vec	0.290
Vse značilke	0.490
Najboljših 250	0.539
Najboljših 1000	0.565
Najboljših 1500	**0.573**
Najboljših 5000	0.553

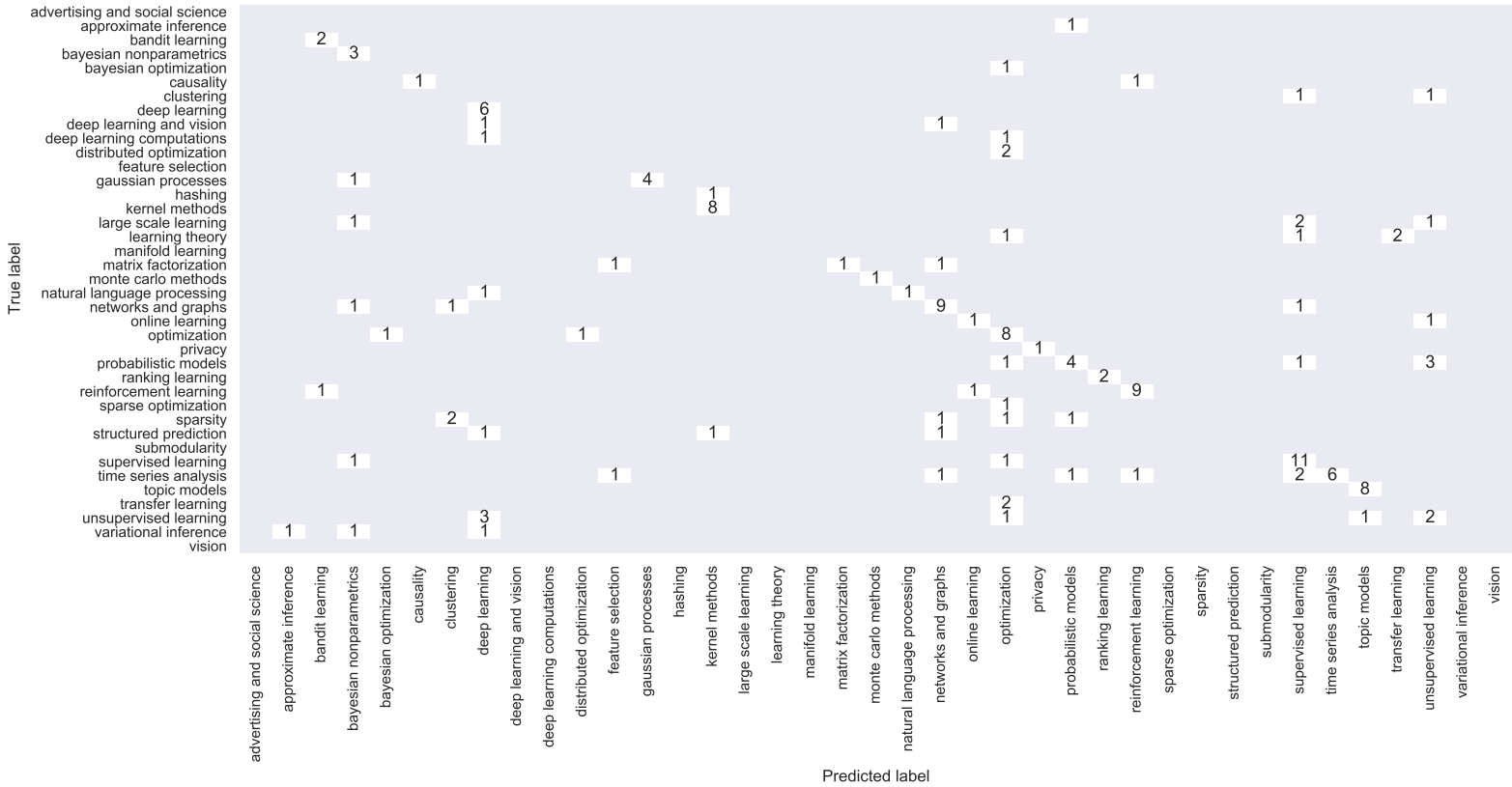
Table: Doseženi rezultati pri klasifikaciji na 39 področij člankov.

Top-N klasifikacija

N	Klasifikacijska točnost
1	0.573
2	0.638
3	0.717
4	0.750
5	0.783
6	0.803
7	0.822
8	0.836
9	0.862
10	0.882

Table: Doseženi rezultati s top-N klasifikacijo.

Matrike klasifikacije



Testiranje na člankih konference ECML-PKDD 2017

- Preverili delovanje gručenja z omejitvami
- Razvrstili 136 člankov v 29 sej
- 17 sej je vsebovalo večino člankov iz istega področja
- 12 sej je vsebovalo mešane članke

Zaključek

- Uspešno izdelali metodo za samodejno razvrščanje člankov v urnik konference
- Nadaljnjo delo:
  - Dodatne metode luščenja značilk
  - Testiranje na različnih področjih
  - Dodatni pristopi gručenja z omejitvami – CSCLP in K-MedoidsSC