



S superračunalnikom nad velike količine besedil in slik

izr. prof. dr. Janez Povh
viš. pred. mag. Andrej Dobrovoljc
as. Jože Bučar
Darko Zelenika

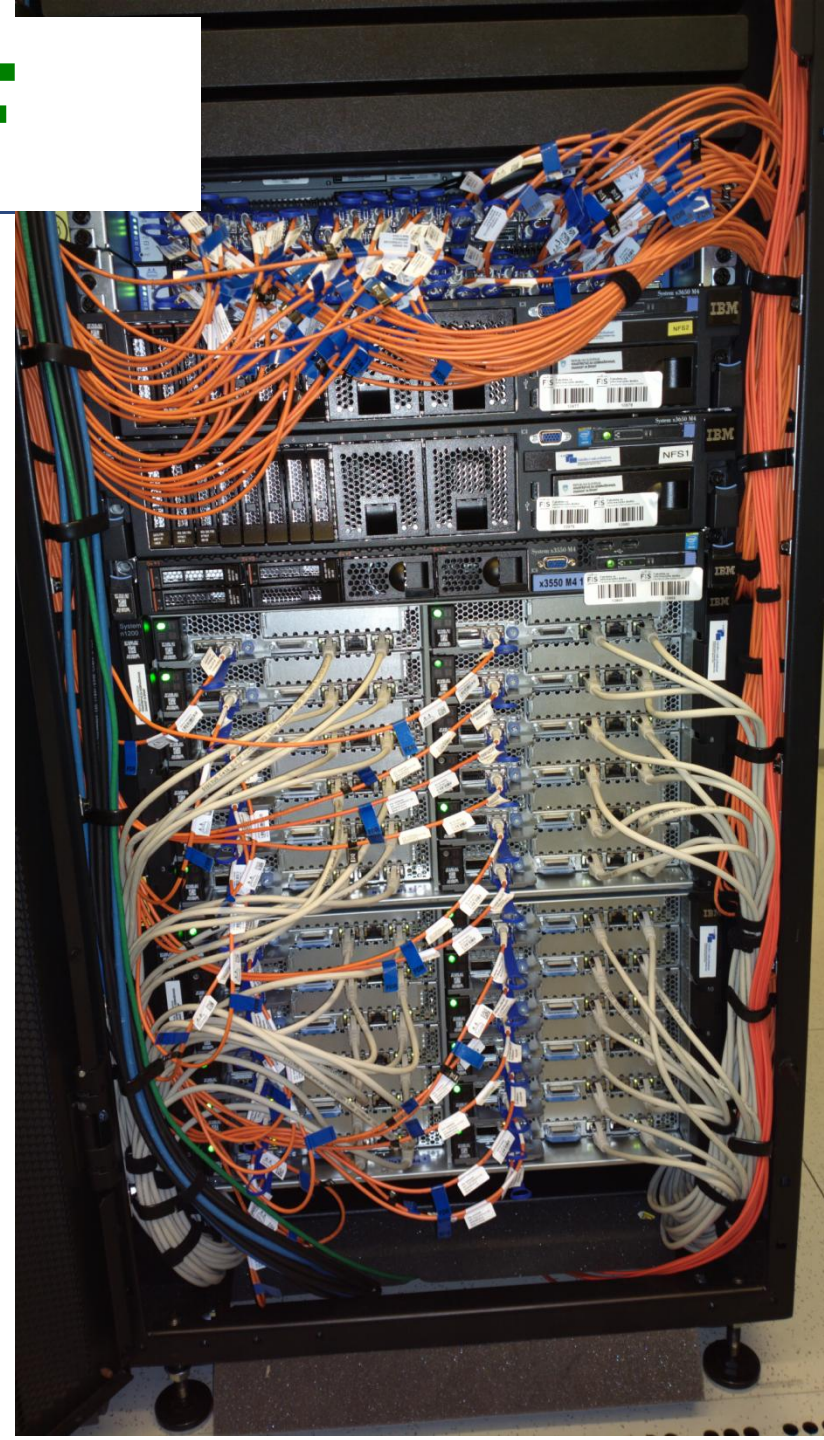
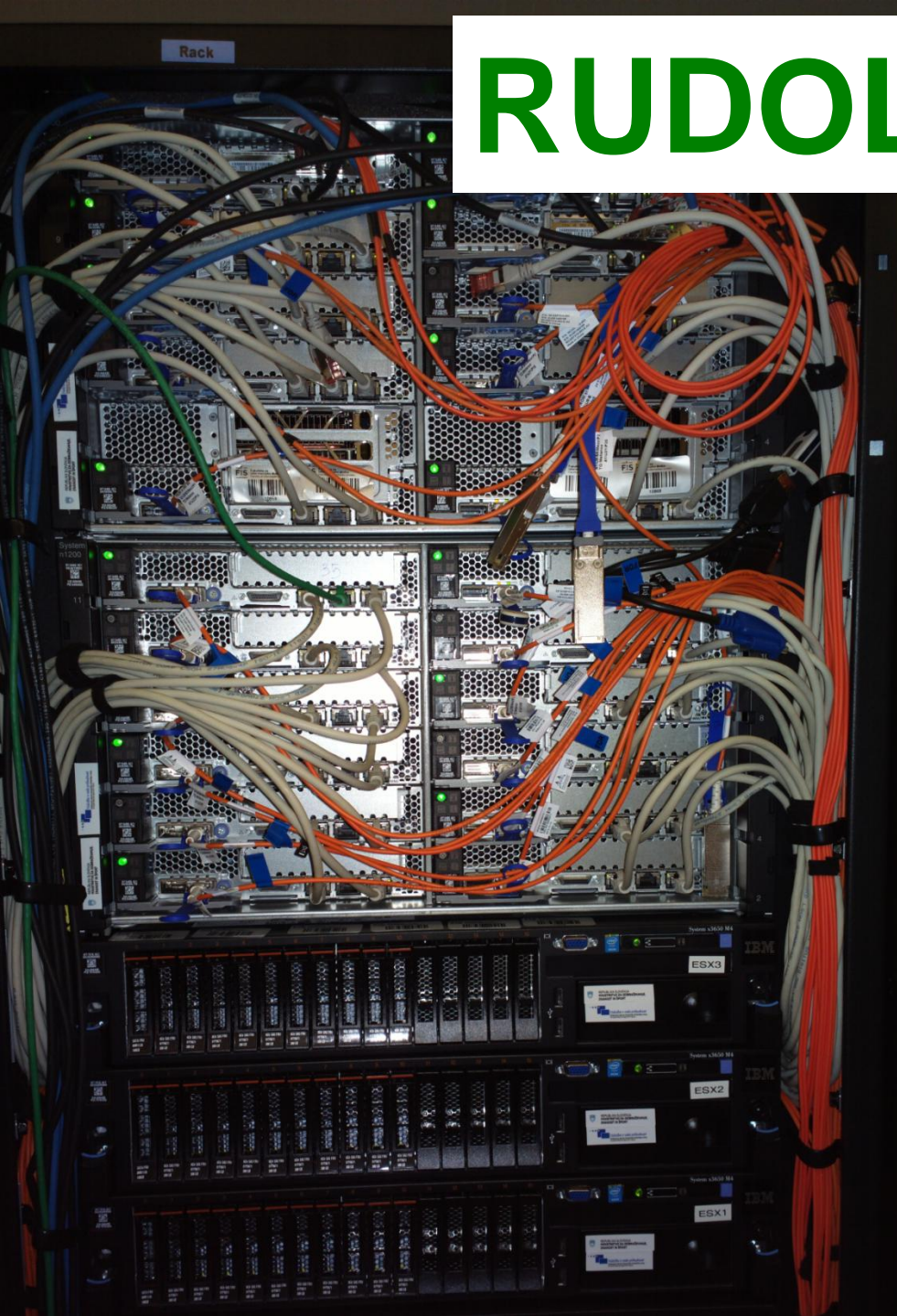


Veliki podatki na FIŠ

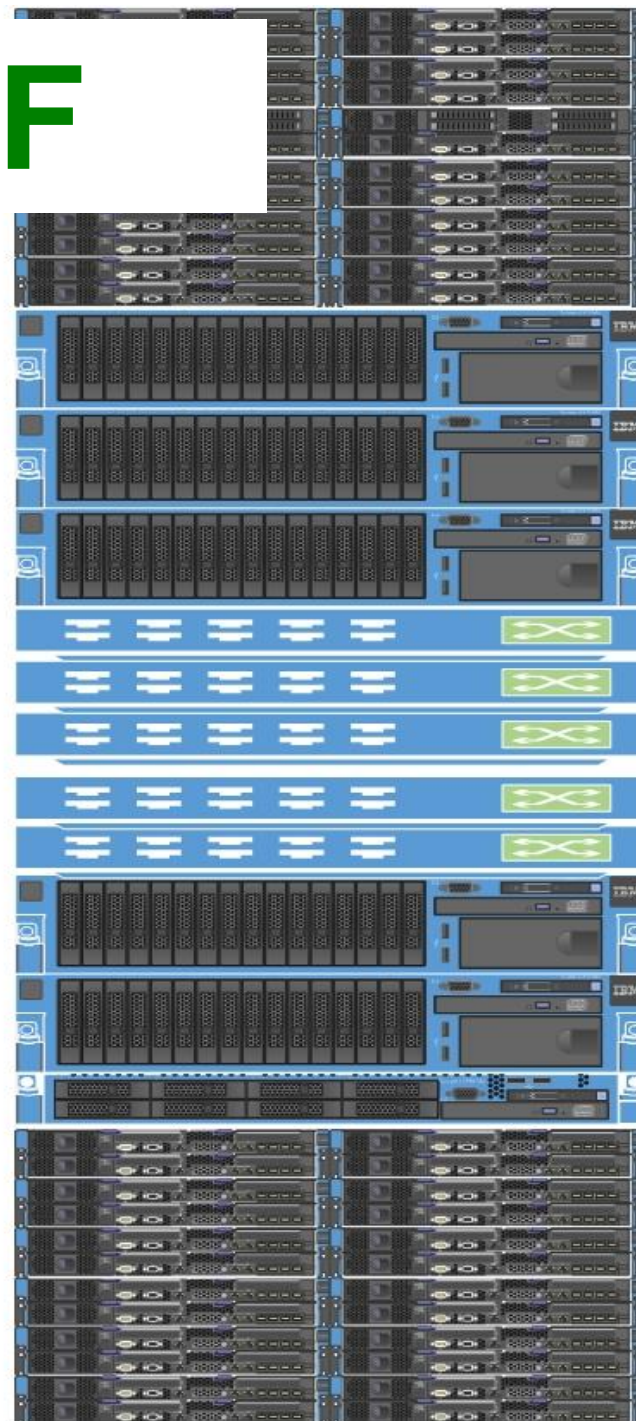
1. Spletna besedila
2. Besedila v obliki slik



RUDOLF



RUDOLF



8x nadgrajeno vozlišče

Tesla
Windows Matlab strežnik

12x navadno vozlišče
1 dodatno za virt ualizacijo
1 Anylogic
1 Matlab ubuntu Biljana

3x Virtualizacijski strežniki
Vmware

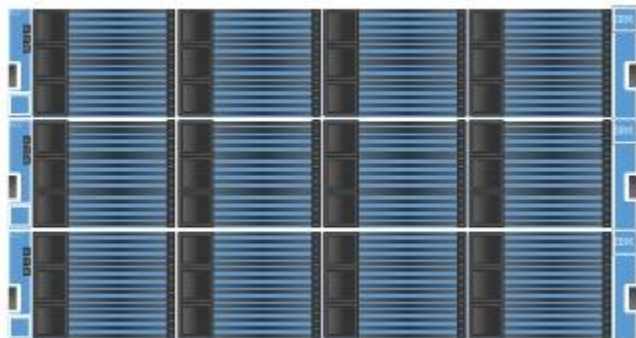
3x mrežno stikalo Cisco

2x inforband stikalo

2x NFS strežnik

Administrativni strežnik

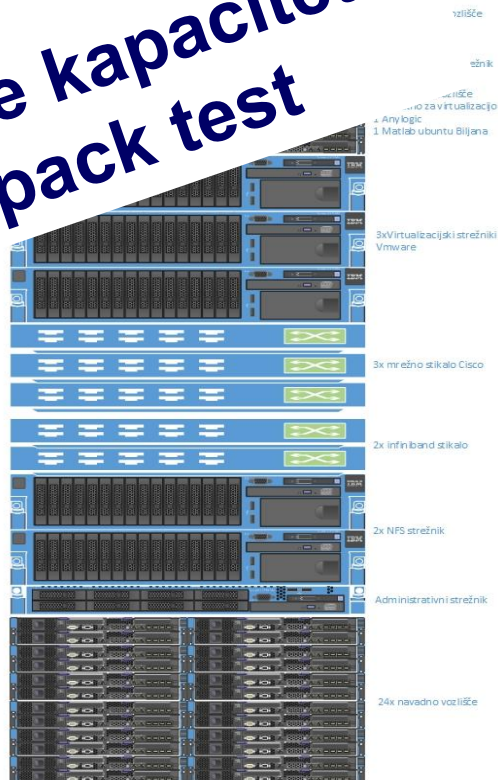
24x navadno vozlišče



IBM V3700 diskovni sistem
32 diskov po 3TB
(102 TB RAW kapaciteta)
4x 7+1 RAID-5 konfiguracija
(72 TB neto kapaciteta)



- 736 CPU jeder
- 3584 GB RAM
- 102 TB diskovne kapacitete
- 12.5 Tflops Linpack test





I. Spletna besedila

34

PUBLICATIONS

2k

Views

1,836

Downloads

211

Citations

24.98

Impact Points

View stats

FEATURED PUBLICATIONS

Article: On an extension of Pólya's
Positivstellensatz
Peter J. C. Dickinson, Janez Povh

3

Views

4

Downloads

0

Citations

Edit

Article: Community Structure and the Evolution
of Interdisciplinarity in Slovenia's Scientific
Collaboration Network
Borut Lužar, Zoran Levnajič, Janez Povh, Matjaž Perc

Ra...
(ECB).
odkupovat

ECB namerava do
skupaj znesa 1.140 m.

Bančna luknja načela tudi
pokojninske prihranke
Zadnje upanje - ustavno sodišče

11. marec 2015 ob 15:43,
zadnji poseg: 11. marec 2015 ob 15:52
Ljubljana - MMC RTV SLO/Televizija Slovenija

Izbris delnic in podrejenih obveznic zaradi krpanja
bančne luknje je pokojninske družbe in sklade, ki
dodatno starostno zavarujejo pol milijona ljudi,
oškodoval za 32 milijonov evrov.

Ustavno sodišče je prejelo več zahtev za presojo
ustavnosti določb zakona o bančništvu, ki je razlastil
oziroma "postrigel" lastnike podrejenih obveznic
saniranih bank. Med oškodovanimi se največkrat
omenja sto tisoč posameznikov, ki so ob sanaciji NLB-
ja, NKBM-ja, Abanke, Probanke, Factor banke in
Banke Celje ostali brez skoraj 600 milijonov evrov
vrednih delnic in podrejenih obveznic.



I. Spletna besedila

- Na spletu je veliko besedil
- Besedila so nestrukturirana
- Različni jeziki
- Slovnična pomanjkljivost
- Vsega ni mogoče prebrati

Ugotavljanje sentimenta v spletnih besedilih



VIR: http://www.dataweave.in/img/xmen_sentiment.png

**Sentiment = čustvo, stališče pisca do
izbrane teme.**

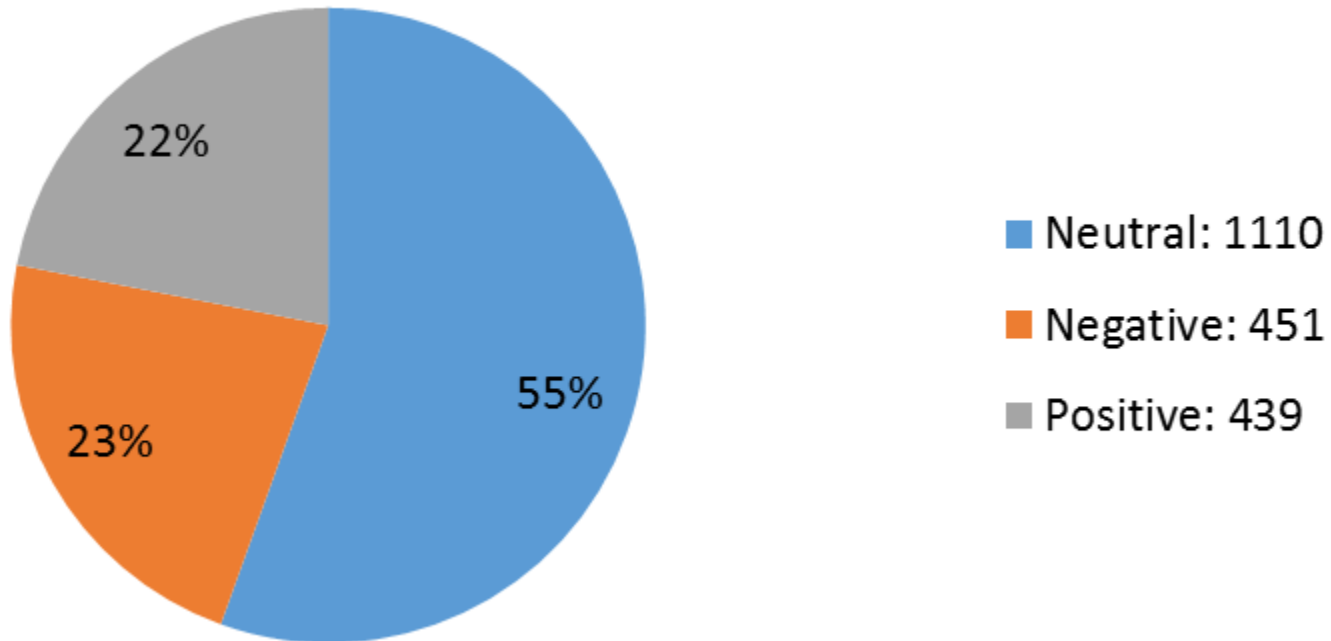
Zakaj analiza sentimenta?



<http://www.had.si/>

Spletna podpora

Zakaj analiza sentimenta?



Spletno vzdušje

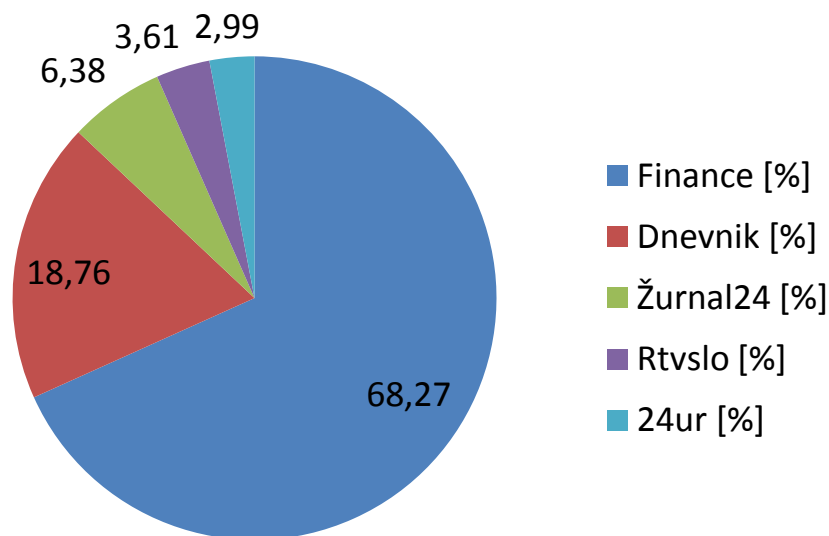


Analiza sentimenta s strojnim učenjem

- potrebujemo učno in testno množico (označeni podatki)
- izberemo značilke (features) za opis besedil
- izračunamo napovedni model
- Ovrednotimo model (prečno preverjanje)
- Uporabimo model na neoznačenih podatkih

Izdelava označene množice

- iz 5 virov smo zbrali 289.782 besedil (1.9.2007-31.12.2013) o politiki, gospodarstvu in financah
- slučajno smo izbrali 5x2000 besedil
- angažirali smo 6 označevalcev



Izdelava označene množice

Ročno označevanje

Znanje je bila vedno najboljša naložba.	<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input checked="" type="radio"/> 4 <input type="radio"/> 5	<input type="radio"/> 1
To meni tudi Mojca Hergouth Koletič, ki vodi jezikovni center.	<input type="radio"/> 1 <input type="radio"/> 2 <input checked="" type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5	<input type="radio"/> 2 <input type="radio"/> 3
Pravi, da ljudje, kljub krizi, prepoznajo vrednost znanja tujih jezikov.	<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input checked="" type="radio"/> 4 <input type="radio"/> 5	<input type="radio"/> 4 <input type="radio"/> 5

Statistični podatki o korpusu:	
Dokumenti:	10.526
Odstavki:	194.381
Stavki:	365.009
Kategorije:	3 (poz, neu, neg)
Enolične besede:	132.763



Podatkovni model

- **12 različnih klasifikacijskih metod**
- **5760 kombinacij nastavitev**
- **10-kratno prečno preverjanje**
- **100-kratna ponovitev prečnega preverjanja**



Rezultati analize sentimenta

- **Naivni Bayes:**
 - Napovedna točnost: 92.2 %
 - Občutljivost (TPP): 94.5 %, Specifičnost (TNR): 87.1 %
 - Preciznost (PPV): 94.2 %, NPV: 87.7 %
- **Metoda podpornih vektorjev:**
 - Napovedna točnost: 82.2 %
 - Občutljivost (TPP): 95,1 %, Specifičnost (TNR): 53,1 %,
 - Preciznost (PPV): 81,9 %, NPV: 83,0 %
- **Logistična regresija (časovno zelo potratno):**
 - napovedna točnost > 95 %



Kje potrebujemo Rudolfa?

- **Izračun značilik**
- **Izvajanju različnih metod pri različnih nastavitvah**
- **Vzporedno računanje modelov pri prečnem preverjanje**
- **Analiza sentimenta v velikih količinah besedil**
- **Uporabljamo R, Matlab, Weka, lastno kodo**

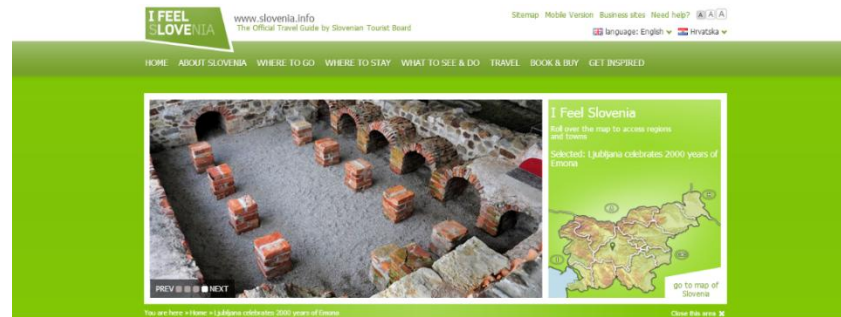


Uporaba sentimenta

- **Merjenje spletnega ugleda/podpore**
- **Merjenje odziva na spletne novice**
- **Zaznavanje sentimenta v besedilnih tokovih (twitter, novice) in koreliranje z drugimi časovnimi vrstami.**
- **Izdelava napovedi (volitve, referendumi, gospodarska aktivnost)**



II. Grafično predstavljena besedila



55
Like
Share
0
Tweets
1
8+1

Ljubljana celebrates 2000 years of Emona

By First · Add to travel planner

This year, Ljubljana celebrates the 2000th anniversary of Emona, a Roman city that once stood on the site of one of the core areas of Ljubljana city centre. Join the celebrations and experience the atmosphere of the time of the Roman Empire!

The programme of events in celebration of the anniversary of Emona includes several interesting museum exhibitions, two Emona-themed city tours led by guides dressed in Roman costumes, and a number of other events, the major highlight being the three-day event Ave Emona!, set to be held in the central Kongressni trg square from 22 to 24 August.

Exhibitions marking the 2000th anniversary of Emona

The City Museum of Ljubljana hosts a major Emona-themed exhibition titled Emona: A City of the Empire, which presents the role of the colony of Emona within the Roman Empire. This summer will also see the opening of the National Museum of Slovenia's new permanent exhibition entitled Stories from the Meeting Point of Different Worlds, which will bring together various archaeological finds from Roman times discovered across Slovenian territory.

Emona-themed events

This year, Ljubljana will host so many Emona-themed events that it is impossible to mention them all. The main highlight will be the living history event Ave Emona!, set to be held over three days in August. The event, during which Ljubljana's Kongressni trg square will be transported back to Roman times, will feature performances by members of different Slovenian and international historical societies acting the roles of Roman legionaries, senators, Vestal Virgins, craftsmen, etc. Those attending the event will be able to taste Roman dishes and get souvenirs from a Roman market.

Roman sights of Ljubljana

In the centre of Ljubljana, the layout of several streets and squares is based on the layout of their Roman predecessors and a number of Roman remains have actually been preserved to the present day. Apart from two archaeological parks (Emona House and Early Christian Centre), you can see one of the longest Roman city walls in the part of Europe, the remains of a Roman road preserved in the basement rooms of the City Museum of Ljubljana, an interesting interactive museum exhibition titled Emona and much more.

Experiencing Roman Emona – guided tours and trips

Experience the atmosphere of Emona. Join one of the two special Emona-themed city tours also give you an opportunity to put on Roman robes, explore the traces of Roman history in the streets and squares of Ljubljana in the company of a professional guide and a Roman legionary, and learn about Roman culture, customs, and daily life through a street-performance. Evening tours are held by torchlight. Roman lunch or dinner can be arranged for organised groups as part of the tour.

Should you be interested in visiting not only Emona, but also some other major Roman sights of Slovenia, join a full-day trip called Via Slovenica, which also includes the remains of the Roman cities of Celea and Poetovna.

Selected for you



Ave, Emona!



Emona: A City of the Empire

options GO

Book & Buy

Reservations

Search in:

All Regions

All Locations

Arrive:

City	Month	Year
LJ	8	2014

Flights:

Rooms	Adults/Room
2	2

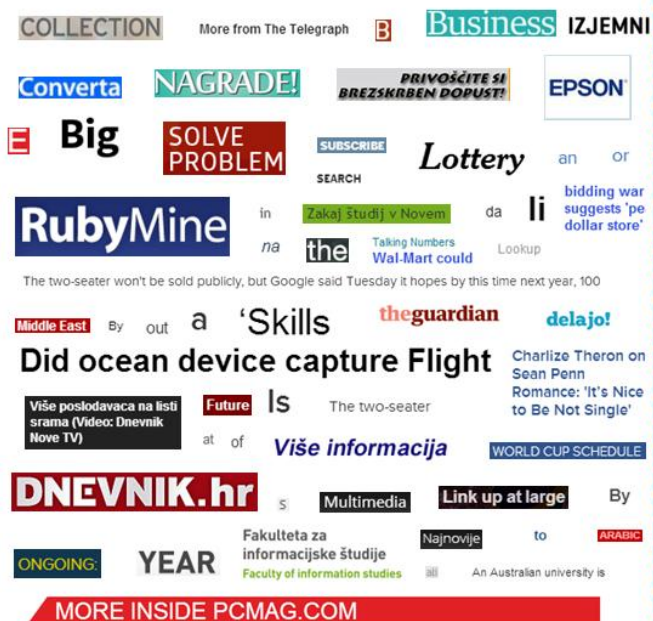
Advanced search

GO

Follow us Facebook RSS
Slovenia Channel Slovenia on TripAdvisor
Slovenia

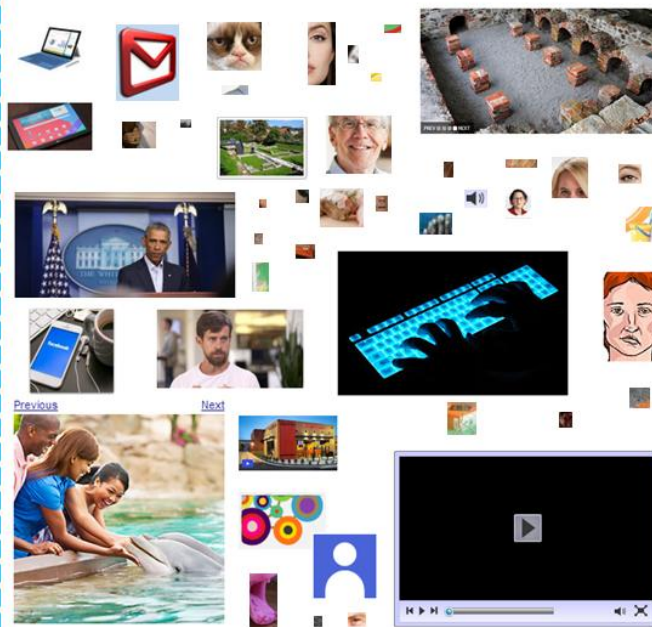
II. Grafično predstavljena besedila

Text base



COLLECTION More from The Telegraph **B** Business IZJEMNI
 Converta NAGRADE! PRIVOŠČITE SI BREZSKRBNEN DOPUST! EPSON
 Big SOLVE PROBLEM SUBSCRIBE Lottery an or
 SEARCH
 RubyMine in Zakaj študij v Novem da li bidding war suggests 'pe dollar store'
 na the Taking Numbers Wal-Mart could Lookup
 The two-seater won't be sold publicly, but Google said Tuesday it hopes by this time next year, 100
 Middle East By out a 'Skills theguardian delajo!
 Did ocean device capture Flight Charlie Theron on Sean Penn Romance: 'It's Nice to Be Not Single'
 Više posodavca na listi srma (Video: Dnevnik Nove TV) Future Is The two-seater
 at of Više informacija WORLD CUP SCHEDULE
 DNEVNIK.hr s Multimedia Link up at large By
 ONGOING YEAR Fakulteta za informacijske študije Najnovije to ARABIC
 Faculty of information studies An Australian university is
 MORE INSIDE PCMAG.COM

Non-Text base



A collection of non-text-based digital content including:
 - A laptop and a mobile phone.
 - A red envelope icon.
 - A cat's face.
 - A woman's face.
 - A screenshot of a building interior with orange cones.
 - A woman's face in a profile view.
 - A screenshot of a green landscape.
 - A man's face.
 - A screenshot of a man speaking at a podium with an American flag.
 - A smartphone displaying a blue screen.
 - A man's face in a video frame.
 - A glowing blue keyboard graphic.
 - A cartoon drawing of a woman's face.
 - A video player interface at the bottom showing a play button and a progress bar.

II. Grafično predstavljena besedila

Podeljene Nagrade GZS za gospodarske in podjetniške dosežke za leto 2014



4.3.2015 - GZS je danes že 47. podelila Nagrade za gospodarske in podjetniške dosežke kot priznanje tistim gospodarstvenikom, ki že vrsto let uspešno vodijo svoja podjetja. Letošnjih nagrajencev je 8, iz 7 podjetij, prihajajo pa iz šestih regij in šestih panog. Med njimi sta tudi Anton Konda in Jože Stupar, direktor in tehnični direktor podjetja KEKO - OPREMA d.o.o. Žužemberk, člana GZDBK.



Raziskovalni problem

- **Segmentacija slik (določitev območij z besedilom in območij pravih slik)**
- **Prepoznavanje besedilnih slik.**



Reševanje

Segmentacija:




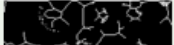
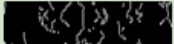
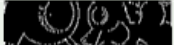

- Z metodami konvolucije zaznavamo vertikalne in horizontalne robove

Kategorizacija slik

- Izdelava učne in testne množico (označeni podatki)
- Definicija in izbira značilik (features) za opis slik
- Izračun napovednega modela
- Ovrednotenje modela (prečno preverjanje)
- Uporaba na neoznačenih podatkih

Rudolf pomaga

Reševanje

-	<u>Color Image</u>	The people who say they	
-	<u>Gray image</u>	The people who say they	
-	<u>Binary Image</u>	The people who say they	
1	<u>Number of CC</u>	19	22
2	<u>Std of heights of CC</u>	1.5044	13.7063
3	<u>Std of length of extracted vertical lines</u>	2.2081	10.1460
4	<u>Std of length of extracted horizontal lines</u>	1.0166	19.6548
-	<u>Skeleton Image</u>	The people who say they	
-	<u>Vertical lines of Skeleton image - convolution</u>	The people who say they	
5	<u>Std of heights of CC</u>	2.1508	4.7849
6	<u>Std of length of extracted vertical lines</u>	3.2922	2.0808
-	<u>Vertical lines of Binary image - convolution</u>	The people who say they	
7	<u>Std of heights of CC</u>	2.0923	9.0183
8	<u>Std of length of extracted vertical lines</u>	3.3343	2.5036
-	<u>Horizontal lines of Binary image - convolution</u>	The people who say they	
9	<u>Std of width of CC</u>	5.7945	21.2756
10	<u>Std of length of extracted horizontal lines</u>	2.2245	13.1401
11	<u>Std of distance image</u>	0	3.7169



Rezultati kategorizacije slik

- **Naivni Bayes:**
 - Napovedna točnost: 78,6 %
 - Občutljivost (TPP): 78,6 %,
 - Preciznost (PPV): 81,3 %
- **Metoda podpornih vektorjev (Gauss):**
 - Napovedna točnost: 96,2 %
 - Občutljivost (TPP): 96,1 %
 - Preciznost (PPV): 96,1 %
- **Slučajni gozd:**
 - Napovedna točnost: 96,4 %
 - Občutljivost (TPP): 96,1 %
 - Preciznost (PPV): 96,0 %



Uporaba

- **Programska rešitev za segmentacijo slik (iskanje blokov na spletni strani)**
- **Prepoznavanje besedilnih blokov (na spletnih straneh, na skeniranih dokumentih)**
- **Pridobivanje besedila z OCR**
- **Pozornost slovenščini in hrvaščini;**



Obvladovanje velikih podatkov je velik izziv

S superračunalnikom je lažje.



Zahvala

- Delo je bilo opravljeno v okviru operacije **Kreativno Jedro: Simulacije** (<http://fis.unm.si>)
- »Operacijo delno financira Evropska unija, in sicer iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa krepitev regionalnih razvojnih potencialov za obdobje 2007-2013, 1. razvojne prioritete: Konkurenčnost podjetij in raziskovalna odličnost, prednostne usmeritve 1.1: Izboljšanje konkurenčnih sposobnosti podjetij in raziskovalna odličnost.«