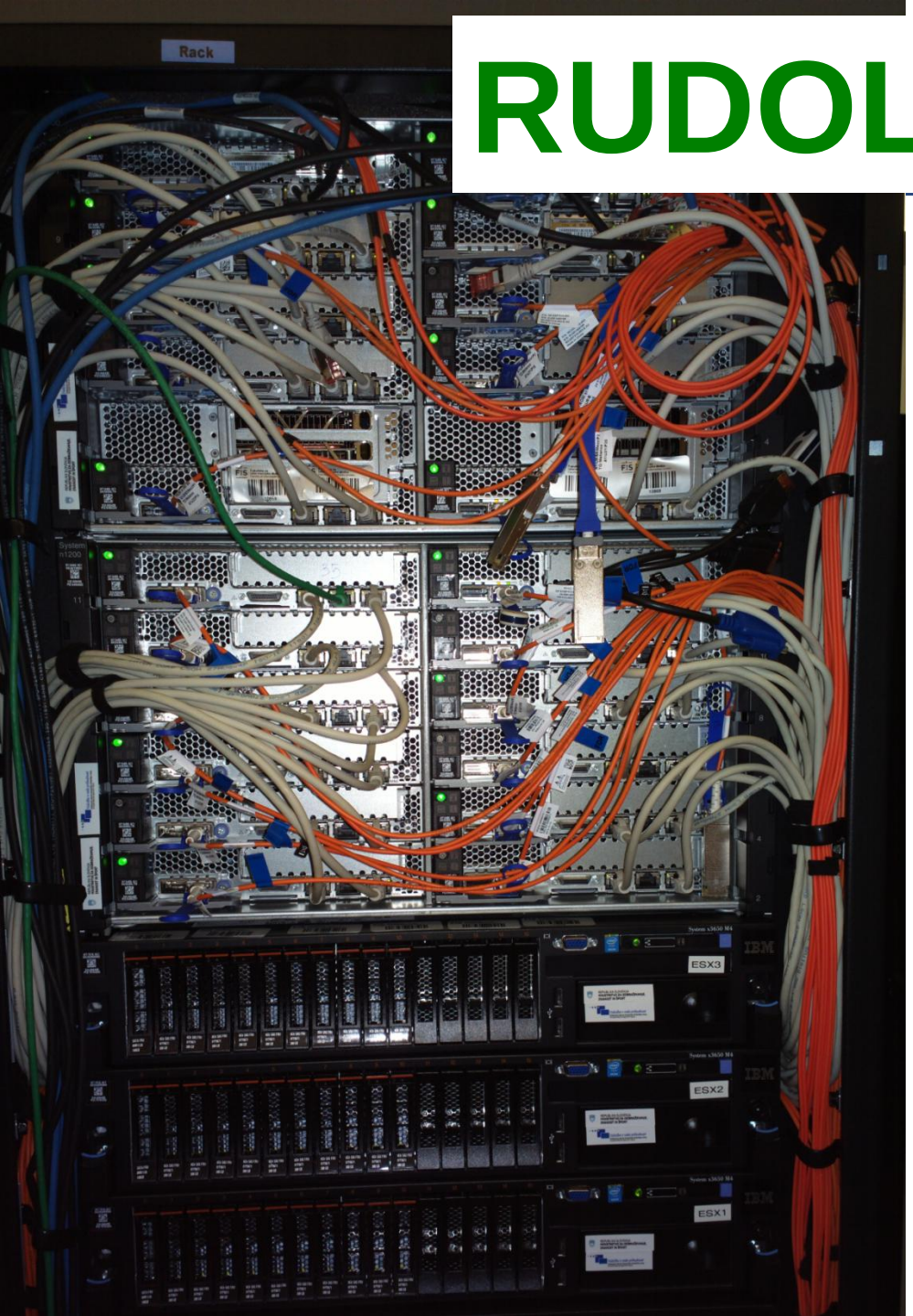
S superračunalnikom nad velike količine besedil in slik

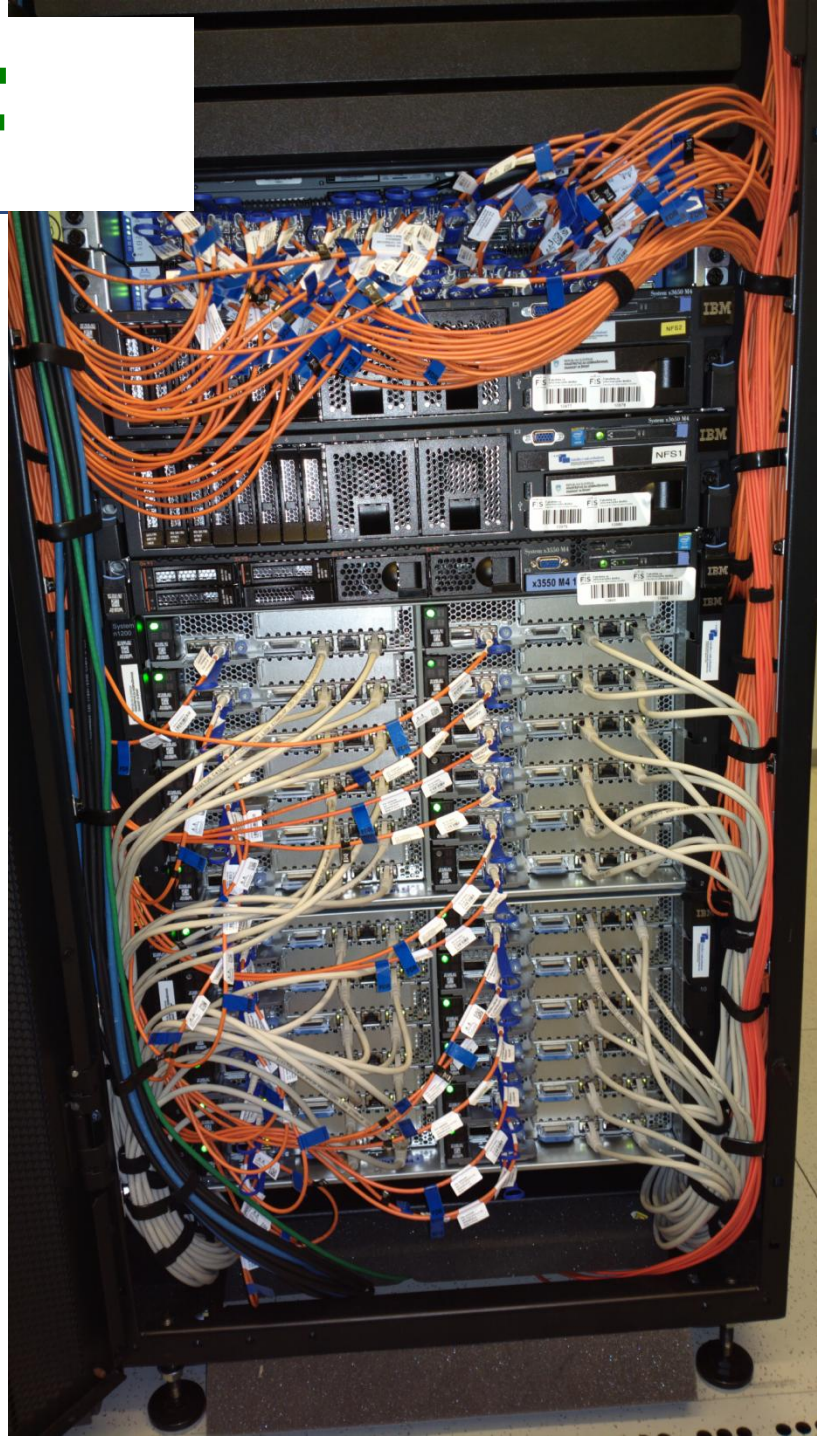
izr. prof. dr. Janez Povh
viš. pred. mag. Andrej Dobrovoljc
as. Jože Bučar
Darko Zelenika

Veliki podatki na FIŠ

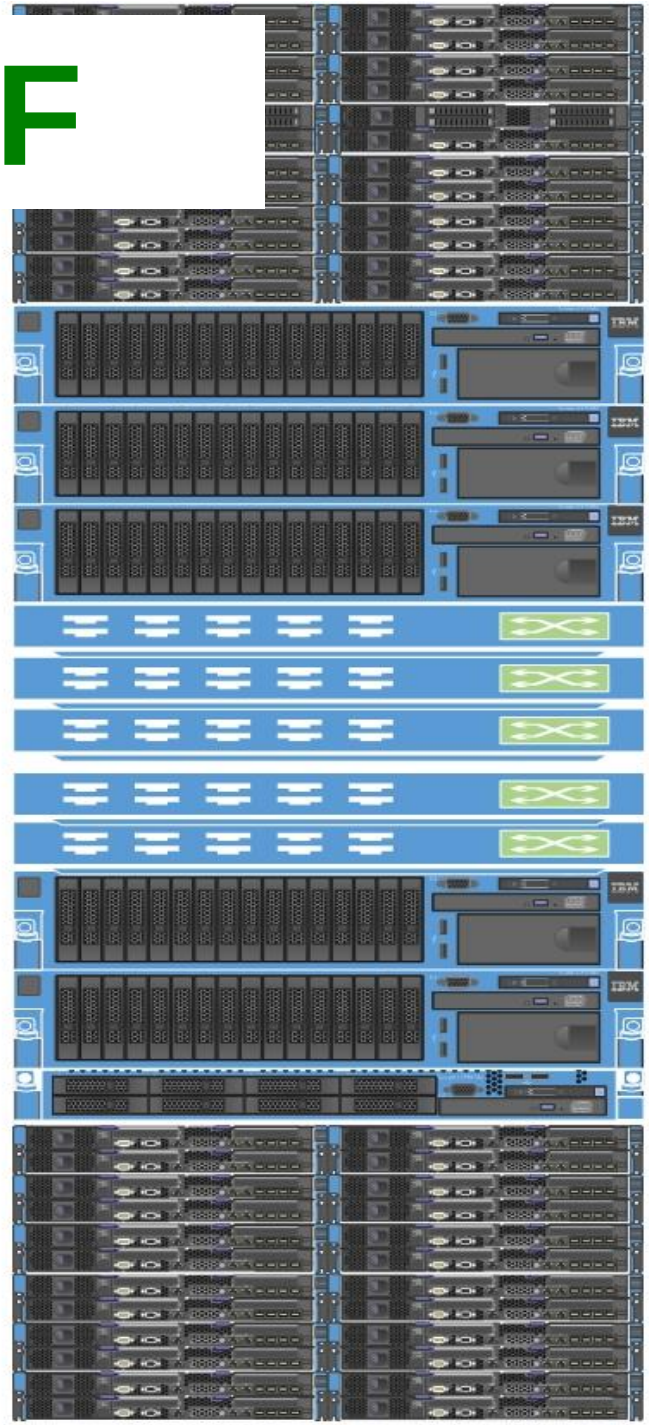
1. Spletna besedila
2. Besedila v obliki slik

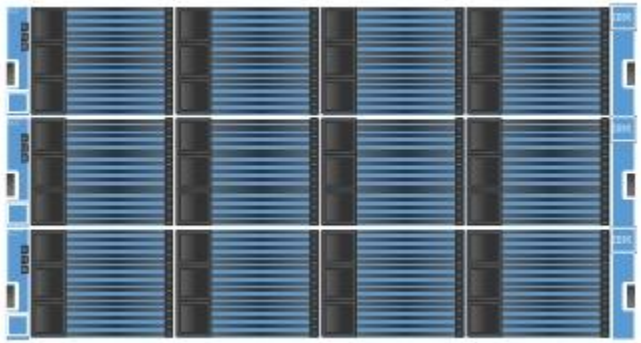
RUDOLF





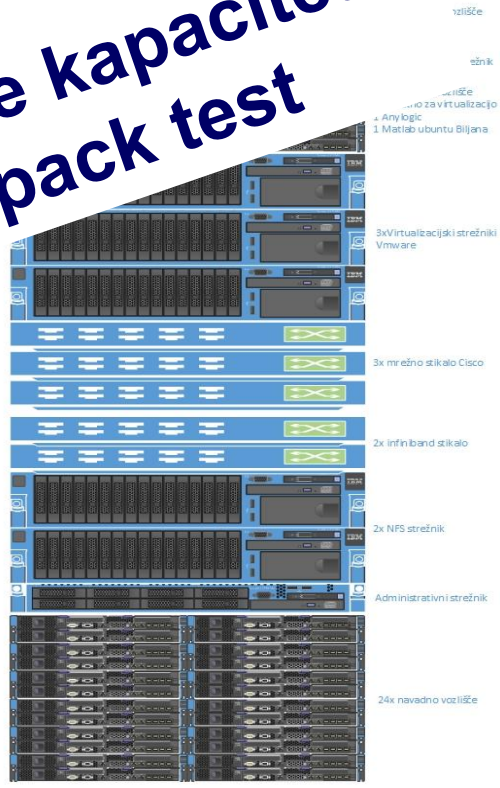
RUDOLF


8x nadgrajeno vozlišče
Tesla
Windows Matlab strežnik
12x navadno vozlišče
1 dodatno za virtualizacijo
1 Anylogic
1 Matlab ubuntu Biljana
3xVirtualizacijski strežniki
Vmware
3x mrežno stikalo Cisco
2x infiniband stikalo
2x NFS strežnik
Administrativni strežnik
24x navadno vozlišče

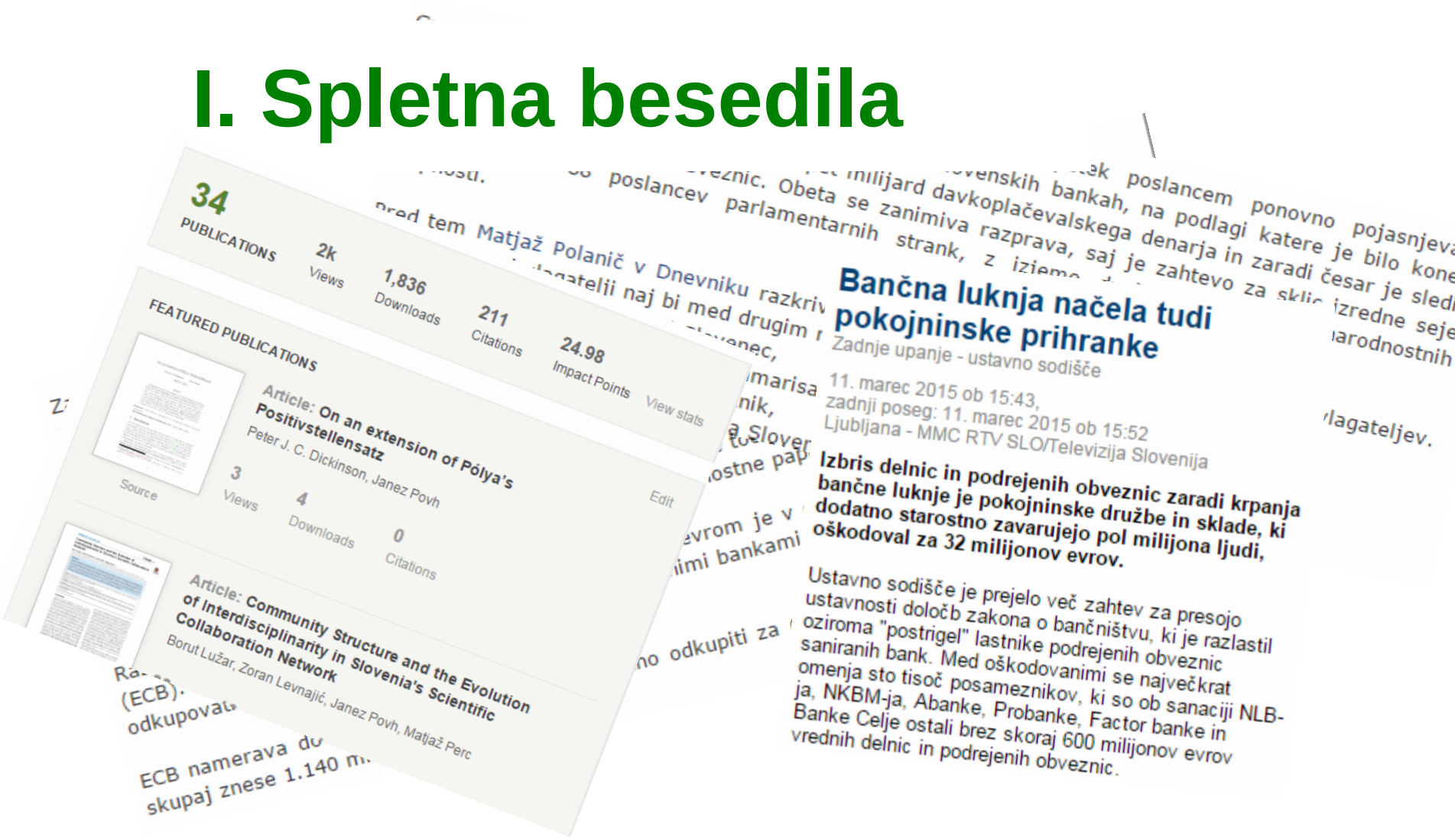


IBM V3700 diskovni si
32 diskov po 3TB
(102 TB RAW kapacite
4x 7+1 RAID-5 konfigu
(72 TB neto kapaciteta

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



I. Spletna besedila



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

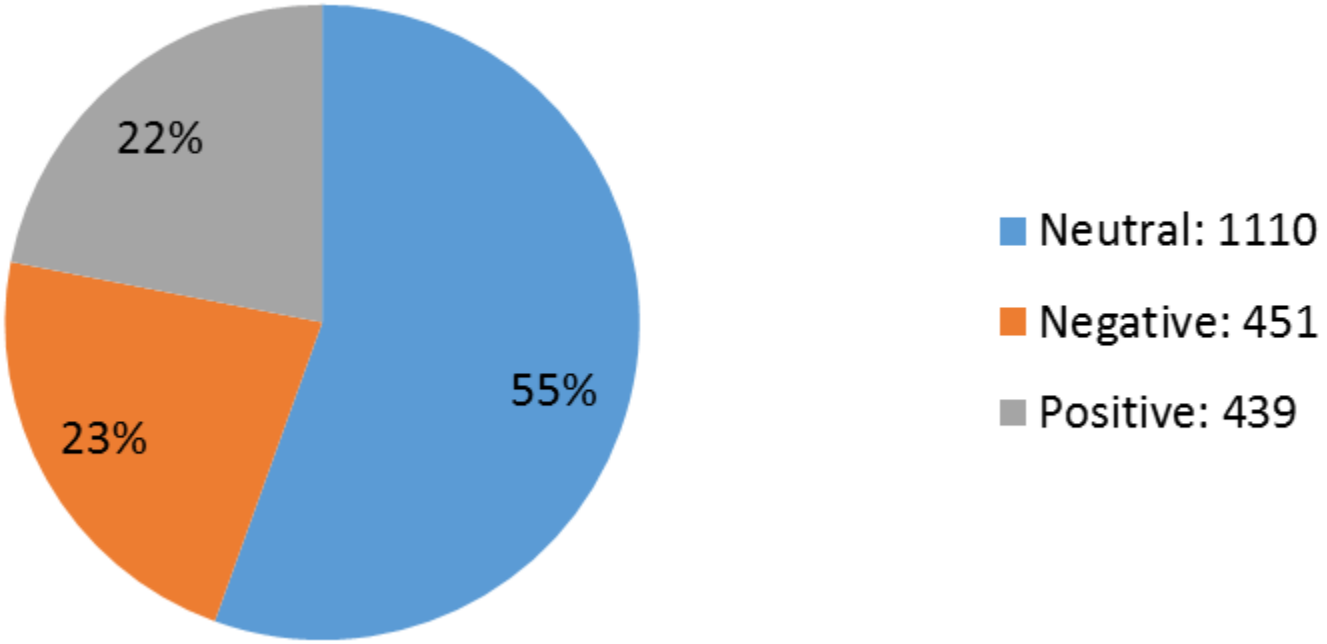
Sentiment = čustvo, stališče pisca do izbrane teme.

Zakaj analiza sentimenta?

http://www.had.si/

Spletna podpora

Zakaj analiza sentimenta?



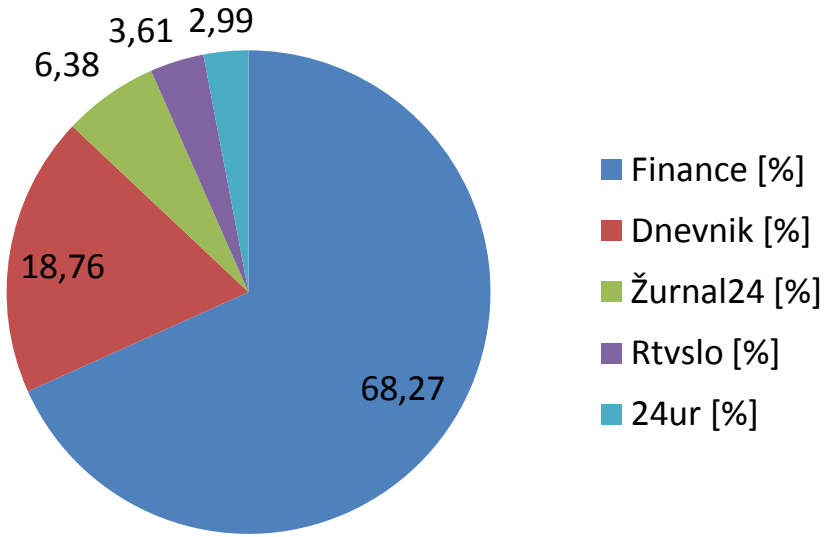
Spletno vzdušje

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

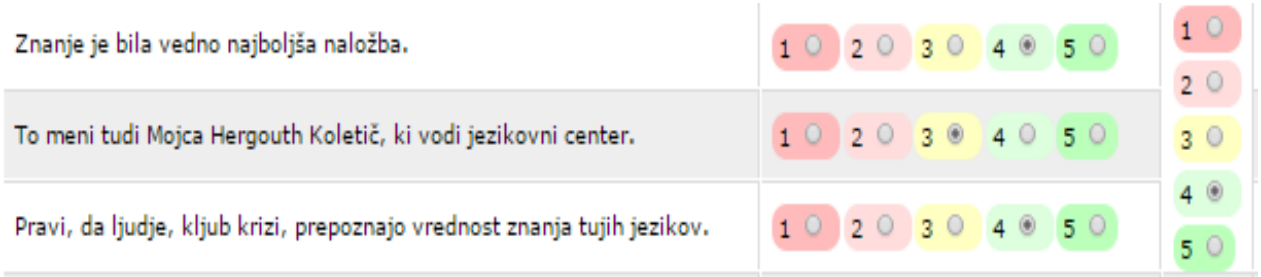
Izdelava označene množice

- 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



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 %**

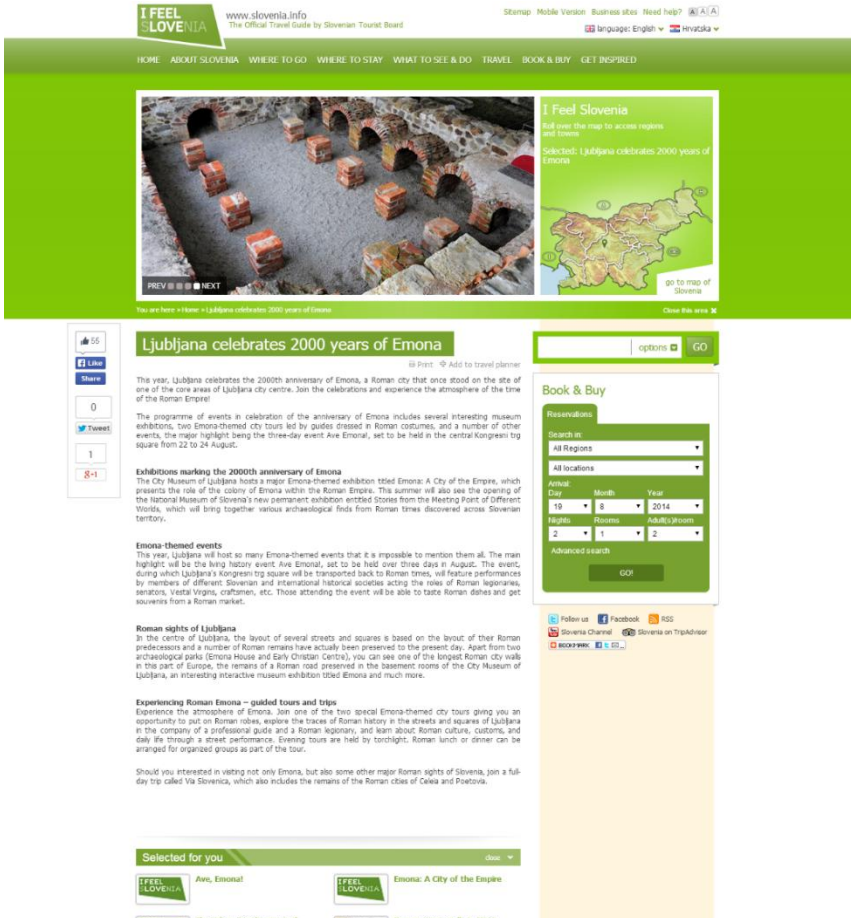
Kje potrebujemo Rudolfa?

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

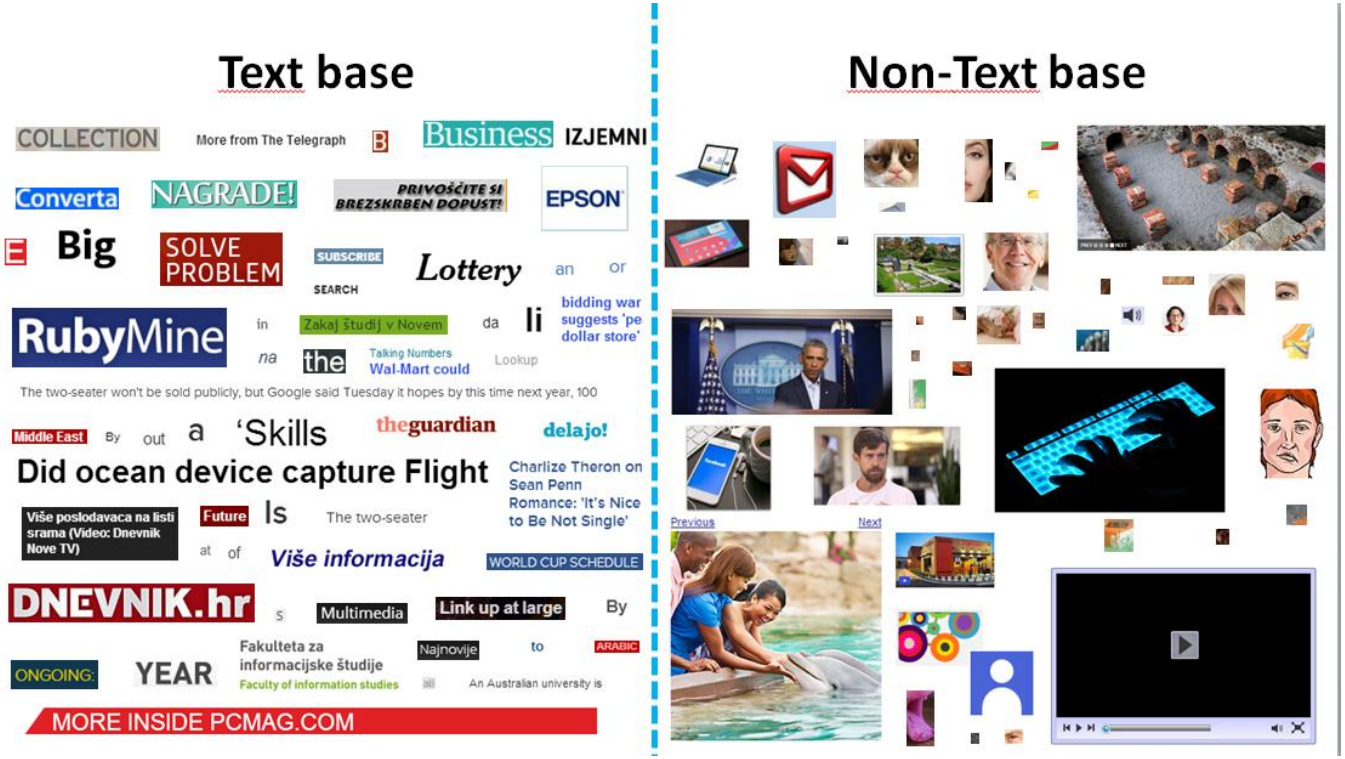
Uporaba sentimenta

- **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



II. Grafično predstavljena besedila



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 Štupar, 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:

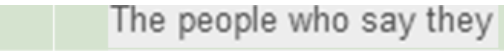
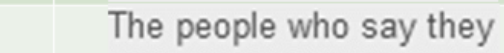
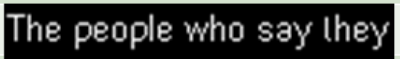
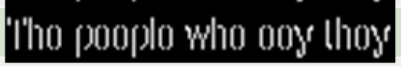
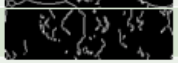
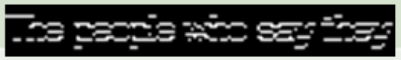
- **Z metodami konvolucije zaznavamo vertikalne in horizontalne robove**

Kategorizacija slik

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

Rudolf pomaga

Reševanje

-	Color Image		
-	Gray image		
-	Binary Image		
1	Number of CC	19	22
2	Std of heights of CC	1.5044	13.7063
3	Std of length of extracted vertical lines	2.2081	10.1460
4	Std of lenght of extracted horizontal lines	1.0166	19.6548
-	Skeleton Image		
-	Vertical lines of Skeleton image - convolution		
5	Std of heights of CC	2.1508	4.7849
6	Std of length of extracted vertical lines	3.2922	2.0808
-	Vertical lines of Binary image - convolution		
7	Std of heights of CC	2.0923	9.0183
8	Std of length of extracted vertical lines	3.3343	2.5036
-	Horizontal lines of Binary image - convolution		
9	Std of width of CC	5.7945	21.2756
10	Std of length of extracted horizontal lines	2.2245	13.1401
11	Std of distance image	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 krepitve 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.«