REPUBLIC OF SLOVENIA
STATISTICAL OFFICE RS

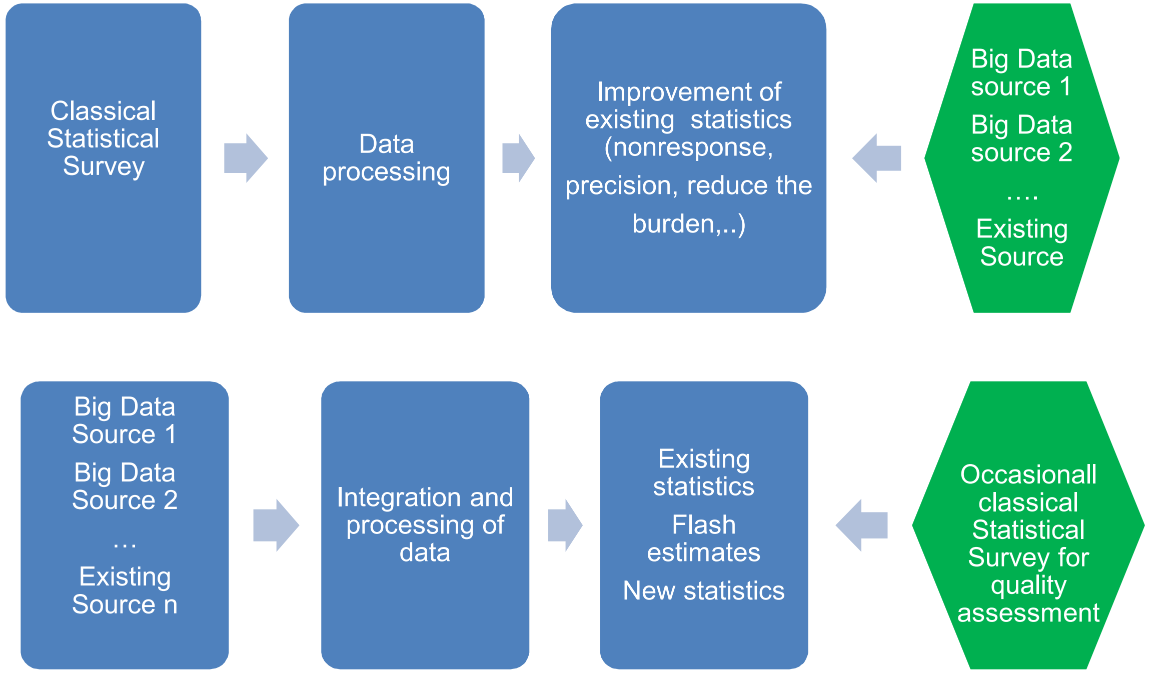
WEB SCRAPING FOR JOB STATISTICS

Boro Nikić

International Conference on Big Data for Official Statistics

Dublin, September 2016

Big data implementation strategy at SURS



Classical Statistical Survey

Data processing

Improvement of existing statistics (nonresponse, precision, reduce the burden,..)

Big Data source 1

Big Data source 2

….

Existing Source

Short term goal

Big Data Source 1

Big Data Source 2

…

Existing Source n

Integration and processing of data

Existing statistics

Flash estimates

New statistics

Occasionall classical Statistical Survey for quality assessment

Long term goal

Current Survey on JV

Observed unit	Enterprise (legal unit)
Probability Sample (4315 units)	Stratified (activity, size),
Administrative part of sample (3633 units)	Enterprise under majority government control
Reference point	Last day of middle month of each quarter
Mode of collection of data	WEB, CATI, Administrative source, Contact center (email, telephone,..)
Statistics	Number of JV ads broken down by activity and size

Current Survey on JV

Observed population: LU of business activities from B to S (sectors)

Reference period: Last working day in the middle month in every quarter.

Statistics: Totals of advertised job vacancies

Total number of Job Vacancies on observed population and domains defined by size and activity:

2 Size classes: 1-9 employees; 10+ employees

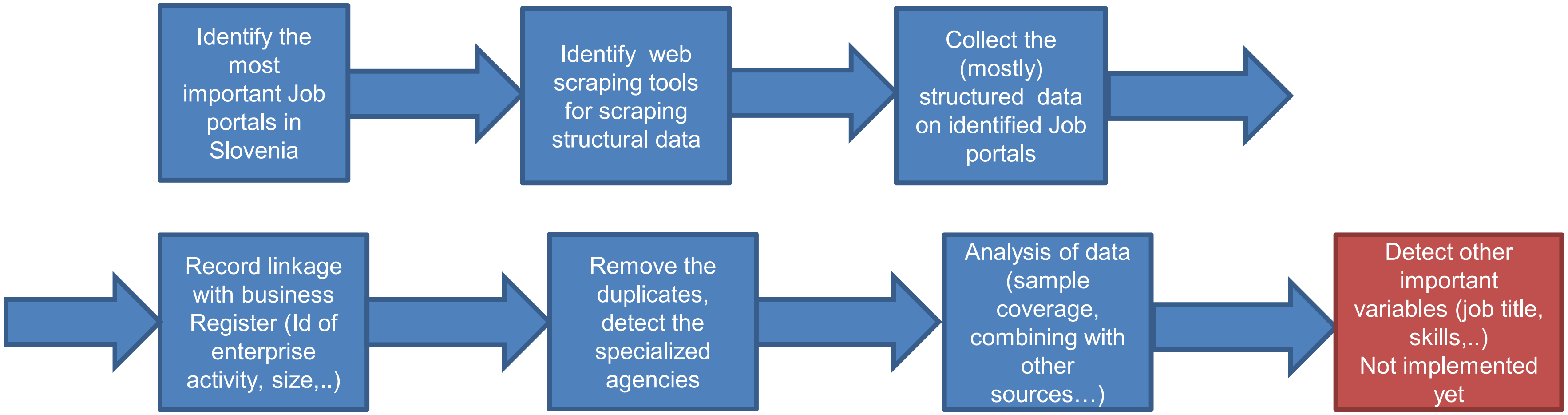
18 activity classes: B-S

Totals of advertised job vacancies on:

- population,
- population broken down by activities,
- units with 10+ employees
- units with 10+ employees broken down by activities

Methodology of scraping of Job portals (1)

Tasks:



Identify the most important Job portals in Slovenia

Identify web scraping tools for scraping structural data

Collect the (mostly) structured data on identified Job portals

Record linkage with business Register (Id of enterprise activity, size,..)

Remove the duplicates, detect the specialized agencies

Analysis of data (sample coverage, combining with other sources…)

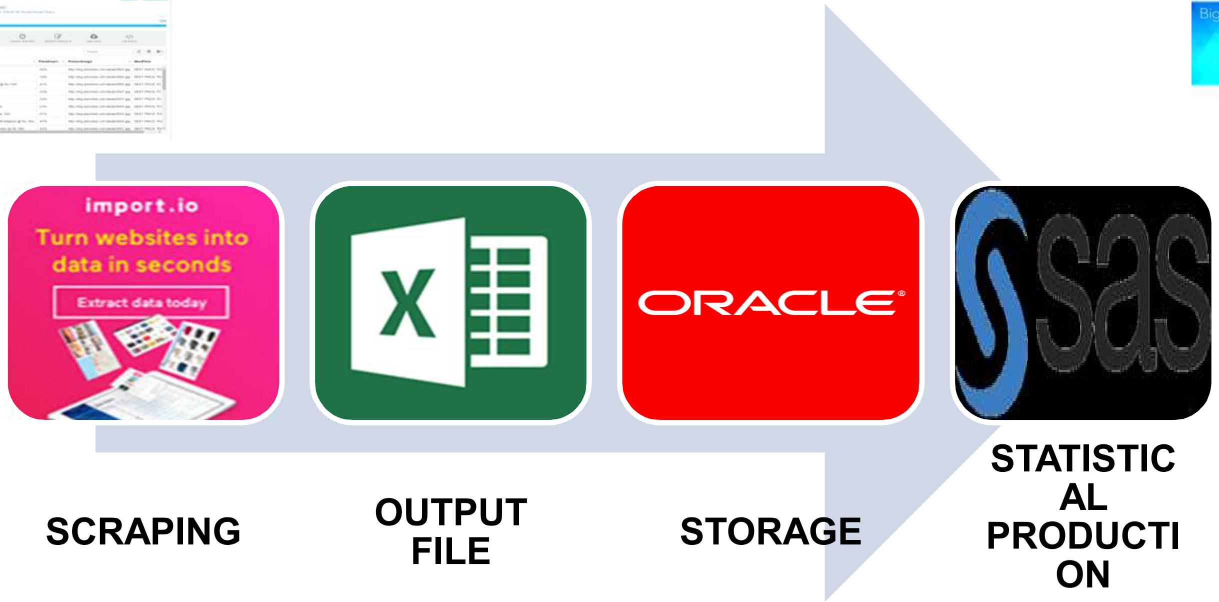
Detect other important variables (job title, skills,..)
Not implemented yet

IT tools involved in scraping Job portals

Data Scraping Studio

SAS Contextual Analytics

Big Data Matters
Improving Machine Learning with SAS Contextual Analysis
An On Demand Webinar Series



SCRAPING

OUTPUT
FILE

STORAGE

STATISTIC
AL
PRODUCTI
ON

Methodology of scraping of enterprise websites (2)

There are around 30 Job portals in Slovenia. Two of the most important ones cover more then 95% JV ads.

MojeDelo.com
Izberi prihodnost

MOJAzaposlitev.SI

Since May 2016 weakly collection of data from those two portals.

Structure of the scraped data

Position	**Enterprise**	**Location**	**Date**
Pizzopek m/ž	Trummer osebni servis d.o.o.	Maribor	Objavljeno: 15.04.2016
Vodja kuhinje m/ž	Trummer osebni servis d.o.o.	Maribor	Objavljeno: 15.04.2016
Knjigovodja m/ž	SPORTINA Bled d.o.o.	Lesce	Objavljeno: 15.04.2016
Asistent vodji produktov m/ž v Mariboru	Trenkwalder kadrovske storitve d.o.o.	Maribor	Objavljeno: 15.04.2016

Record linkage with BR

	31.5.		16.8.		may		june		june1		july	
	Number	Rate	Number	Rate	Number	Rate	Number	Rate	Number	Rate	Number	Rate
1	1400	**72%**	1173	**76%**	2109	**73%**	2802	**73%**	2662	**73%**	2192	**74%**
2	365	**19%**	222	**14%**	488	**17%**	678	**18%**	645	**18%**	506	**17%**
3	12	**1%**	22	**1%**	20	**1%**	34	**1%**	34	**1%**	31	**1%**
4	14	**1%**	18	**1%**	32	**1%**	41	**1%**	39	**1%**	37	**1%**
5	2	**0%**	0	**0%**	2	**0%**	2	**0%**	2	**0%**	0	**0%**
6	6	**0%**	23	**1%**	13	**0%**	15	**0%**	15	**0%**	8	**0%**
7	74	**4%**	52	**3%**	94	**3%**	117	**3%**	117	**3%**	91	**3%**
missing	84	**4%**	42	**3%**	128	**4%**	153	**4%**	154	**4%**	108	**4%**
total	1957		1552		2886		3842		3668		2973	
Agencies	558	**28,51**	487	**31,38**	749	**25,95**	999	**26,00**	943	**25,71**	775	**26,07**

•**Legend:**

•**1** - merging by unique short name of enterprise

•**2** - merging by unique compete name of enterprise

•**3** - merging by non-unique short name and location of enterprise

•**4** - Record linkage by using distance function (short name of enterprise)

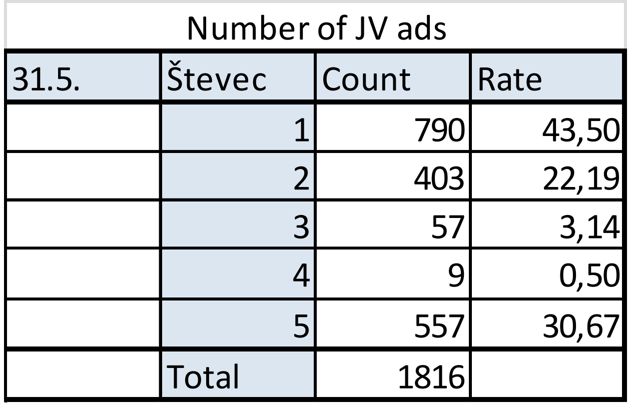
5 -- Merging by non-unique complete name and location of enterprise

•**6** - Record linkage by using distance function (complete name of enterprise)

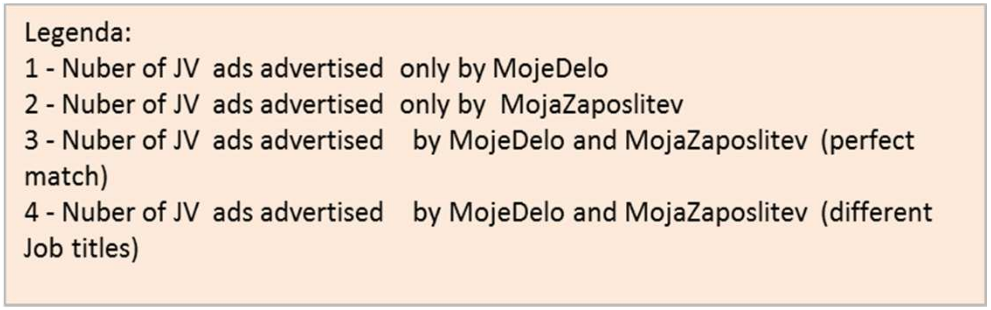
7 - Manual (agencies, biger enterprises)

Duplicates

Key for merging: name of enterprise, job title, location

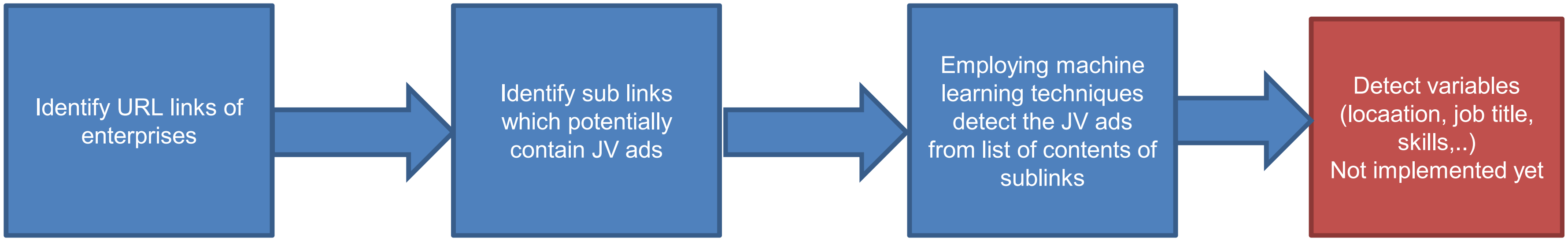


Number of JV ads			
31.5.	Števec	Count	Rate
	1	790	43,50
	2	403	22,19
	3	57	3,14
	4	9	0,50
	5	557	30,67
	Total	1816	



Legenda:
1 - Nuber of JV ads advertised only by MojeDelo
2 - Nuber of JV ads advertised only by MojaZaposlitev
3 - Nuber of JV ads advertised by MojeDelo and MojaZaposlitev (perfect match)
4 - Nuber of JV ads advertised by MojeDelo and MojaZaposlitev (different Job titles)

Methodology of scraping of enterprise websites (1)

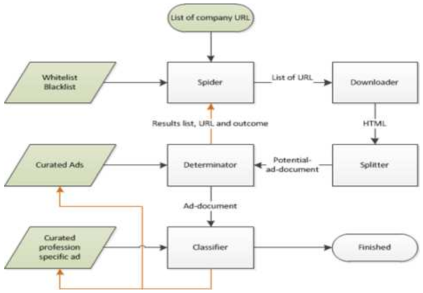


Identify URL links of enterprises

Identify sub links which potentially contain JV ads

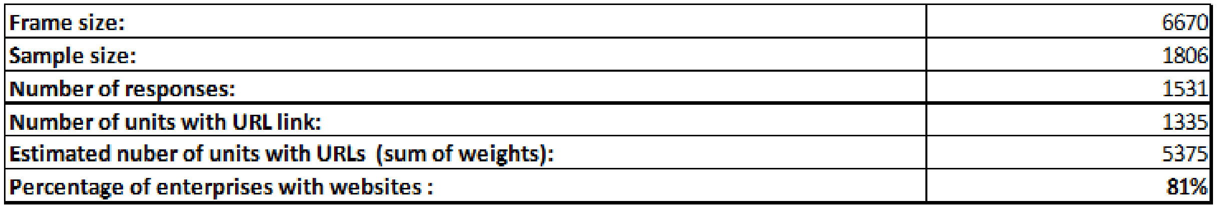
Employing machine learning techniques detect the JV ads from list of contents of sublinks

Detect variables (locaation, job title, skills,..)
Not implemented yet



Methodology of scraping of enterprise websites (2)

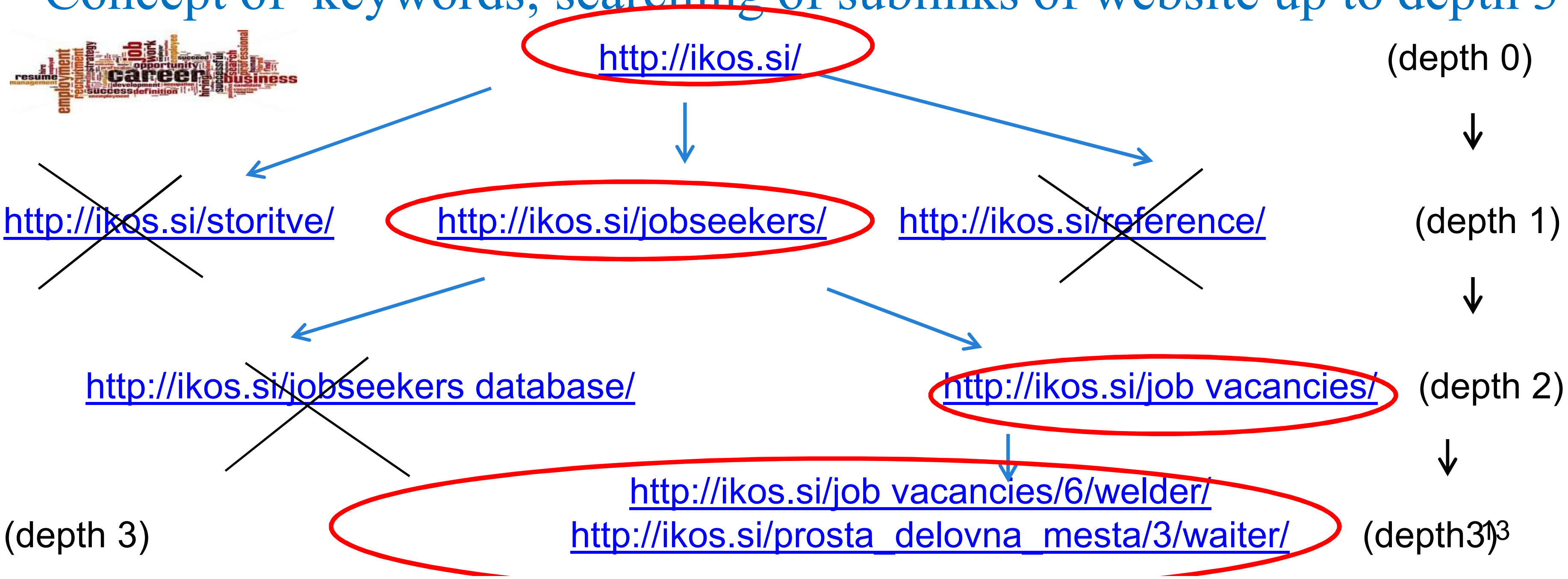
- 2486 URL links (out of 8942 units in sample) of enterprises
- Sources: ICT Survey, Business Register, Phonebook...
- Excluded: international enterprises, Webshops, media, job portals, specialized agencies
- Pilot on units in ICT Survey



Frame size:	6670
Sample size:	1806
Number of responses:	1531
Number of units with URL link:	1335
Estimated nuber of units with URLs (sum of weights):	5375
Percentage of enterprises with websites :	**81%**

Methodology of scraping of enterprise websites (3)

Concept of keywords; searching of sublinks of website up to depth 3



http://ikos.si/ (depth 0)

http://ikos.si/storitve/ http://ikos.si/jobseekers/ http://ikos.si/reference/ (depth 1)

http://ikos.si/jobseekers database/ http://ikos.si/job vacancies/ (depth 2)

http://ikos.si/job vacancies/6/welder/
http://ikos.si/prosta_delovna_mesta/3/waiter/

(depth 3) (depth3)

Methodology of scraping of enterprise websites (3)

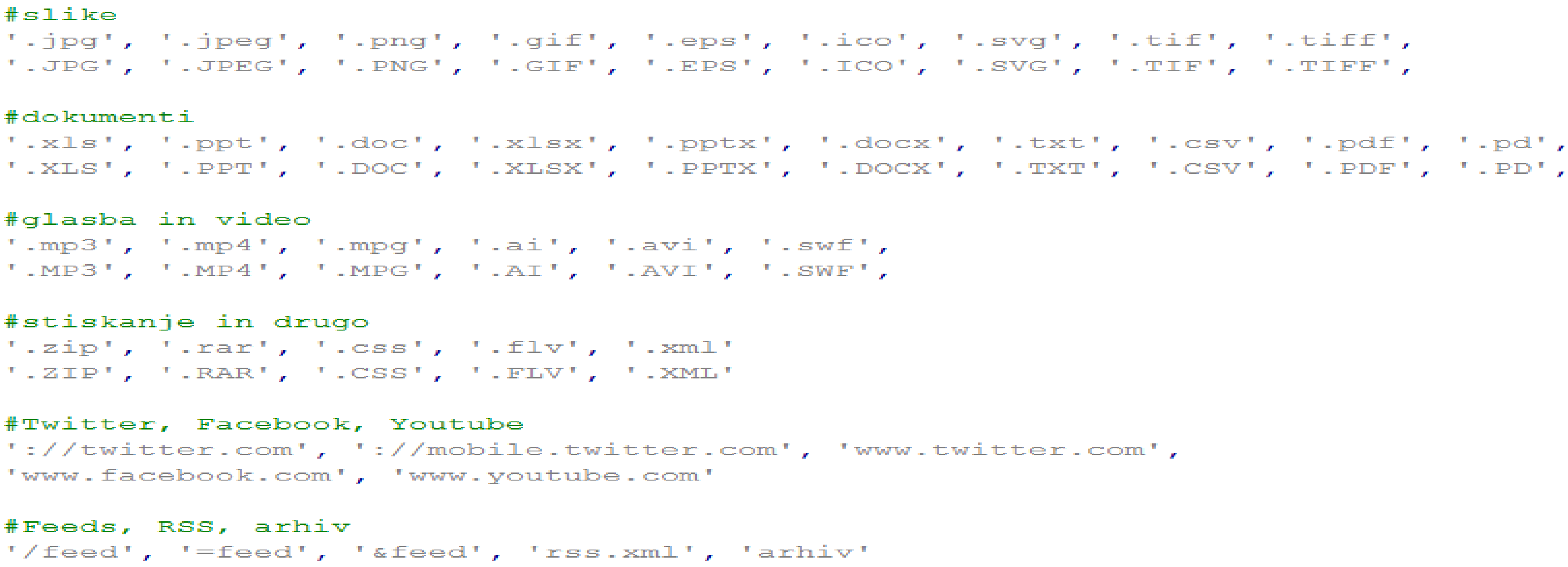
For each webpage the HTML source of the page is checked

- If the link begin with ftp ali sftp, it is excluded
- Domain must remain the same, otherwise it is excluded

  https://mercator.si/ ~~http://www.mercator-ip.si/job/~~

- Links containing images, pdf or word documents, video, facebook or twitter accounts.. are excluded

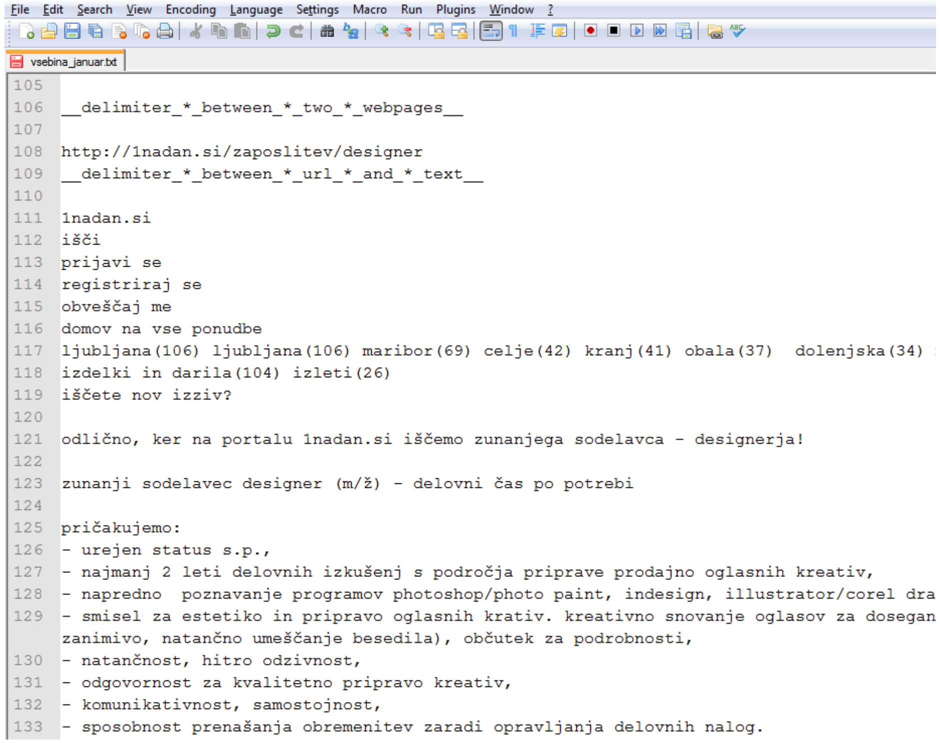
Methodology of scraping of enterprise websites (4)



```
#slike
'.jpg', '.jpeg', '.png', '.gif', '.eps', '.ico', '.svg', '.tif', '.tiff',
'.JPG', '.JPEG', '.PNG', '.GIF', '.EPS', '.ICO', '.SVG', '.TIF', '.TIFF',

#dokumenti
'.xls', '.ppt', '.doc', '.xlsx', '.pptx', '.docx', '.txt', '.csv', '.pdf', '.pd',
'.XLS', '.PPT', '.DOC', '.XLSX', '.PPTX', '.DOCX', '.TXT', '.CSV', '.PDF', '.PD',

#glasba in video
'.mp3', '.mp4', '.mpg', '.ai', '.avi', '.swf',
'.MP3', '.MP4', '.MPG', '.AI', '.AVI', '.SWF',

#stiskanje in drugo
'.zip', '.rar', '.css', '.flv', '.xml'
'.ZIP', '.RAR', '.CSS', '.FLV', '.XML'

#Twitter, Facebook, Youtube
'://twitter.com', '://mobile.twitter.com', 'www.twitter.com',
'www.facebook.com', 'www.youtube.com'

#Feeds, RSS, arhiv
'/feed', '=feed', '&feed', 'rss.xml', 'arhiv'
```

Methodology of scraping of enterprise websites (5)

Only text is scraped from every sub-link with potentially JV ads



```

__delimiter_*_between_*_two_*_webpages__

http://1nadan.si/zaposlitev/designer
__delimiter_*_between_*_url_*_and_*_text__

1nadan.si
išči
prijavi se
registriraj se
obveščaj me
domov na vse ponudbe
ljubljana(106) ljubljana(106) maribor(69) celje(42) kranj(41) obala(37)  dolenjska(34)
izdelki in darila(104) izleti(26)
iščete nov izziv?

odlično, ker na portalu 1nadan.si iščemo zunanjega sodelavca - designerja!

zunanji sodelavec designer (m/ž) - delovni čas po potrebi

pričakujemo:
- urejen status s.p.,
- najmanj 2 leti delovnih izkušenj s področja priprave prodajno oglasnih kreativ,
- napredno  poznavanje programov photoshop/photo paint, indesign, illustrator/corel dra
- smisel za estetiko in pripravo oglasnih krativ. kreativno snovanje oglasov za dosegan
zanimivo, natančno umeščanje besedila), občutek za podrobnosti,
- natančnost, hitro odzivnost,
- odgovornost za kvalitetno pripravo kreativ,
- komunikativnost, samostojnost,
- sposobnost prenašanja obremenitev zaradi opravljanja delovnih nalog.
```

Methodology of scraping of enterprise websites (6)

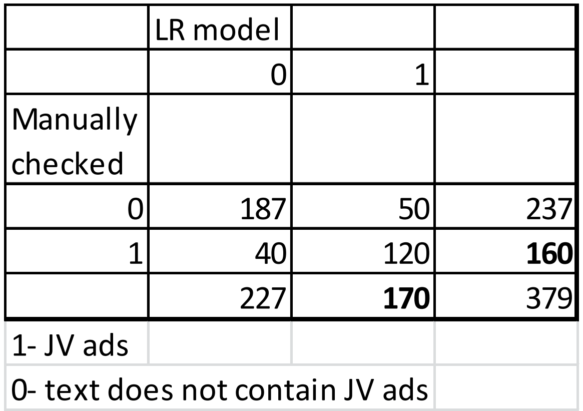
For detection an JV ads machine learning techniques are employed

- Hand-defined list of 1000 examples in February (200 Ads and 800 non Ads)
- Logistic regression model is used

Results; 397 potentially sub-links (397 enterprises) are detected. Manually checked: 160 JV ads (484 JV)

Methodology of scraping of enterprise websites (6)

Results: May, 2016



	LR model		
	0	1	
Manually checked			
0	187	50	237
1	40	120	**160**
	227	**170**	379
1- JV ads			
0- text does not contain JV ads			

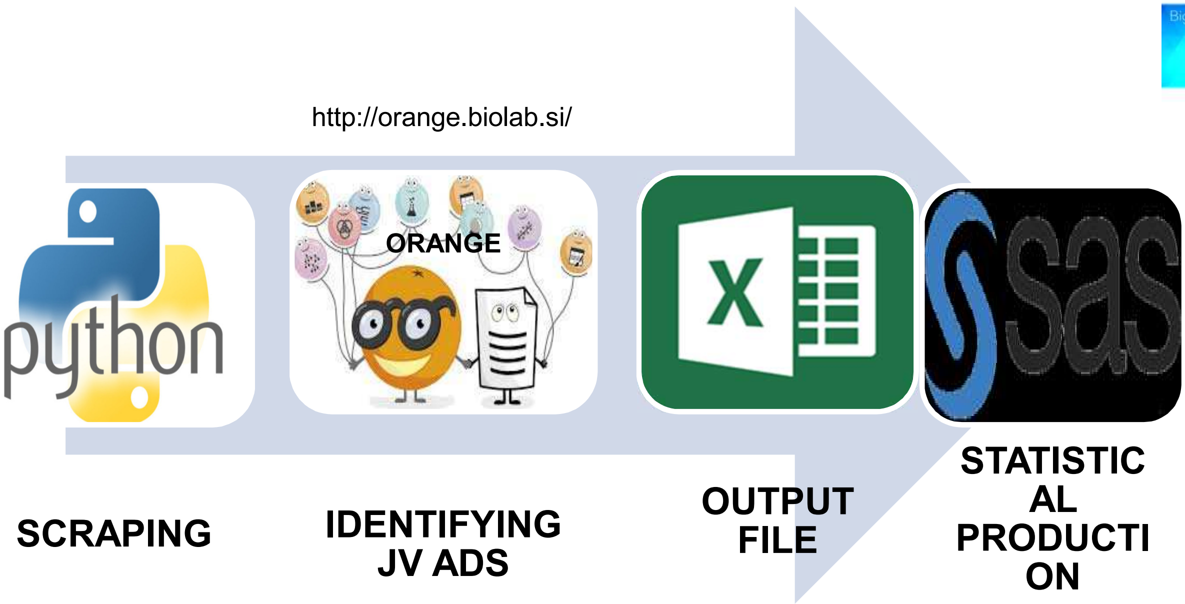
IT tools involved in scraping enterprise websites

SAS Contextual Analytics

Big Data Matters

Improving Machine Learning with SAS Contextual Analysis

An On Demand Webinar Series



http://orange.biolab.si/

SCRAPING

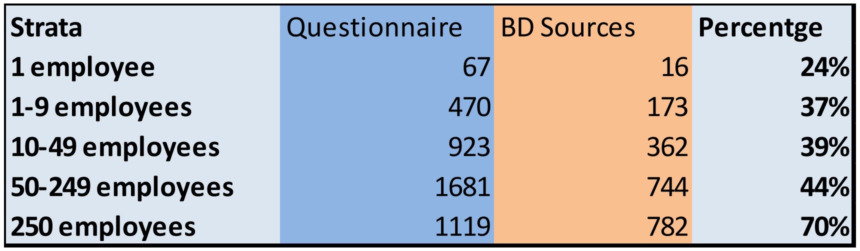
IDENTIFYING JV ADS

OUTPUT FILE

STATISTICAL PRODUCTION

Coverage: Sample vs. scraped&admin data

	Reported data	Scraped& admin data	Job portals	Enterprise websites
Number of JV ads	4312	2321	1073	262
Percentage	100%	54%	25%	6%



Strata	Questionnaire	BD Sources	Percentge
1 employee	67	16	24%
1-9 employees	470	173	37%
10-49 employees	923	362	39%
50-249 employees	1681	744	44%
250 employees	1119	782	70%

Foreseeing activities

- Identify URLs of enterprises
- Improve the scalability of application for identifying sub URL links (Sandbox)
- Text mining (unstructured data)
- Special treatment of specialized JV agencies (Adecco, Hill international…)
- Identify other channels which are used for advertisements of JV (additional questions in questionnaire)
- Include the scraped data in phases of collecting

and processing sample data

Thank you for your attention!

boro.nikic@gov.si