

Identifying Web Tables

Supporting a Neglected Type of Content
on the Web

Michael Galkin, Dmitry Mouromtsev, Sören Auer

Outline

- Motivation
- Neglected?
- Pipeline
- Machine Learning
- Evaluation
- Applications

Motivation

<http://commoncrawl.org/> <http://stats.lod2.eu/>



Michael Galkin

Motivation

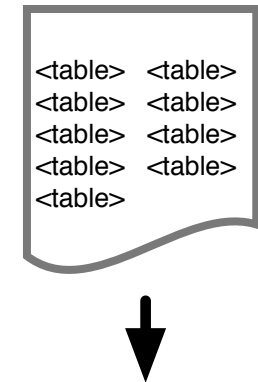
the **DATA**

<http://commoncrawl.org/> <http://stats.lod2.eu/>



Michael Galkin

Motivation



the **DATA**

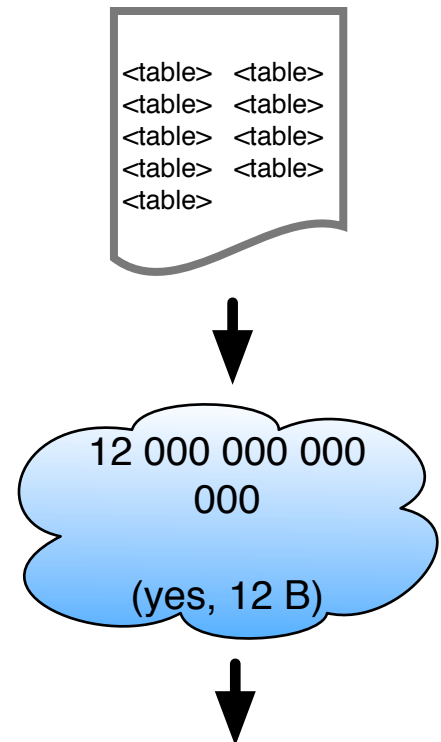
- 9 tables on an average HTML page

<http://commoncrawl.org/> <http://stats.lod2.eu/>

Motivation

the **DATA**

- 9 tables on an average HTML page
- 12 B tables were extracted

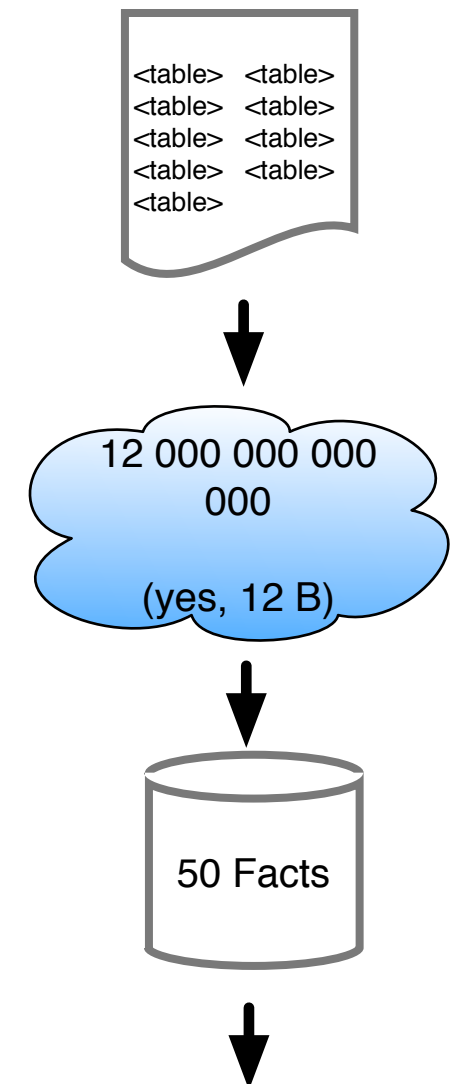


<http://commoncrawl.org/> <http://stats.lod2.eu/>

Motivation

the **DATA**

- 9 tables on an average HTML page
- 12 B tables were extracted
- A table contains ~50 facts

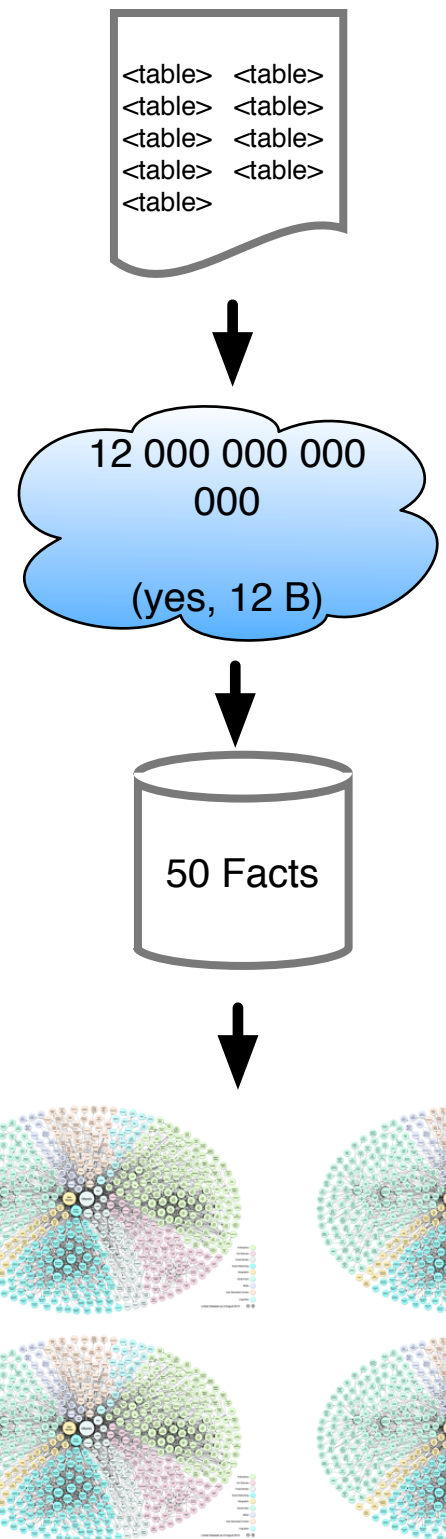


<http://commoncrawl.org/> <http://stats.lod2.eu/>

Motivation

the **DATA**

- 9 tables on an average HTML page
- 12 B tables were extracted
- A table contains ~50 facts
- 600 B facts



<http://commoncrawl.org/> <http://stats.lod2.eu/>

Motivation

Tables are a natural way how people interact with structured data (for humans)

	Thing1	Thing2	Thing3
Param1	val11	val21	val31
Param2	val12	val22	val32
Param3	val13	val23	val33

Motivation

Tables are a natural way how people interact with structured data (for humans)

Thing1 Thing2 Thing3

Param1	val11	val21	val31
Param2	val12	val22	val32
Param3	val13	val23	val33

```
1 <style type="text/css">
2 .tg {border-collapse:collapse;border-spacing:0;}
3 .tg td{font-family:Arial, sans-serif;font-size:14px;padding:10px
4 5px;border-style:solid;border-width:1px;overflow:hidden;word-break:normal;}
5 .tg th{font-family:Arial, sans-serif;font-size:14px;font-weight:normal;padding:10px
6 5px;border-style:solid;border-width:1px;overflow:hidden;word-break:normal;}
7 .tg .tg-s6z2{text-align:center}
8 </style><table class="tg" style="undefined;table-layout: fixed; width: 208px">
9 <colgroup><col style="width: 41px"><col style="width: 55px"><col style="width: 51px">
10 <col style="width: 61px"></colgroup><tr><th class="tg-031e"></th><th class="tg-s6z2">P1</th>
11 <th class="tg-s6z2">P2</th><th class="tg-s6z2">P3</th></tr><tr><td class="tg-031e">Obj1</td>
12 <td class="tg-s6z2">a1</td><td class="tg-s6z2">b2</td><td class="tg-s6z2">c3</td></tr><tr>
13 <td class="tg-031e">Obj2</td><td class="tg-s6z2">d1</td><td class="tg-s6z2">e2</td>
14 <td class="tg-s6z2">f3</td></tr><tr><td class="tg-031e">Obj3</td><td class="tg-s6z2">g1</td>
15 <td class="tg-s6z2">h2</td><td class="tg-s6z2">i3</td></tr><tr><td class="tg-031e">Obj4</td>
16 <td class="tg-s6z2">j1</td><td class="tg-s6z2">k2</td><td class="tg-s6z2">l3</td></tr><tr>
17 <td class="tg-031e">Obj5</td><td class="tg-s6z2">m1</td><td class="tg-s6z2">op</td>
18 <td class="tg-s6z2">xyz</td></tr></table>
```

Neglected?

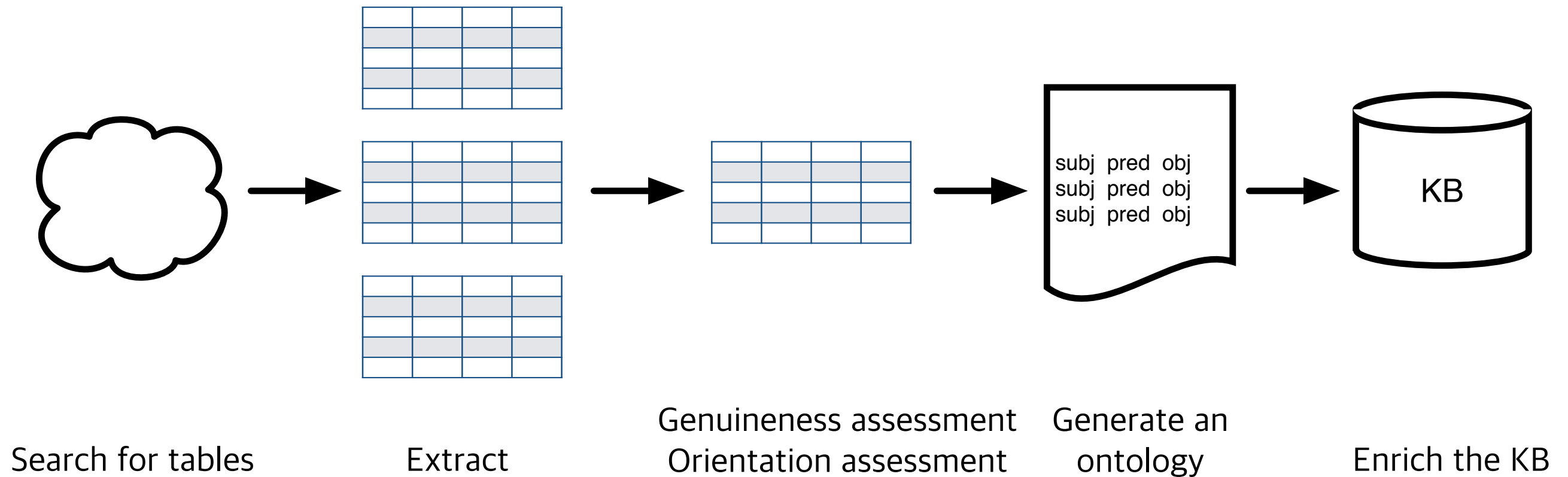
- Processing - search engines index anything except tables
- Annotation - web tables annotation tools are on the edge of extinction
- Retrieval - exists for pictures, audio, video data

Goals

Bring the tables knowledge to the LOD Cloud

- Search & Ingestion
- Explore characteristics of a table
- Extract knowledge & map with the existing knowledge graph

Pipeline



Machine Learning

Did we find a data table or not?

- Genuineness assessment
a table contains some data, not html formatting

Официальный сайт
Администрации Санкт-Петербурга

📄 📁

» [Всё для удобства](#)

Губернатор

Власть

Государственные услуги

Правозащитный центр

Земельные участки

Странички информации

[Справочник](#) / [Вопросы](#) / [Вопросы читателей, поступающие в службу](#) / [Вопросы по вопросам транспортной инфраструктуры](#) / [Справка](#) / [Справочник](#) / [Справочник](#) / [Справочник](#)

Адренная программа текущего ремонта дорог на 2012 год

Наименование	Вторичный участок, кв. м	Вторичный участок, кв. м	Всего (тысяч м)
Адвертительная дорога	171,87	71,38	343,25
Адвертительная дорога от Пискаревского пр. до Пискаревского пр.	7,54	0,00	7,54
Волковская пр. от наб. р. Мойки до Садовой ул.	1,77	3,10	4,87
пер. Гривина	6,22	2,68	8,90
Загородный пр. от Ботанической ул. до Московского пр.	25,94	6,81	32,75
Земельный пр. от Загородного пр. до ул. Кассаткина	14,36	5,35	19,71
Кассаткина пр. от пер. Гривина до Фонарного пер.	6,17	0,00	6,17
3-й Красноварский ул.	5,69	3,85	9,53
11-й Красноварский ул.	4,26	4,17	8,43
13-й Красноварский ул. от Лейбштетского пер. до Ботанической пр.	2,89	1,85	4,83
Лейбштетский пр. от наб. Фонтанки до наб. Обводного канала	16,53	3,56	20,09
наб. Обводного канала от Подъёмного пер. до Даровского пр.	0,00	16,67	16,67
Пискаревский пр. от Загородного пр. до наб. Обводного канала	12,40	0,00	12,40
наб. р. Гривина от наб. р. Мойки до Пискаревского пр.	5,95	0,32	5,97

Власть

Комитет, управление, инспекция и служба
Комитет по развитию транспортной инфраструктуры Санкт-Петербурга

Официальная информация
Структура ИТУВ

Административное управление
Основные подразделения

Документы

Техническая документация

Службы

Различные информационные ресурсы

Наши

Основные сведения

Дорожные материалы

Объекты инженерной инфраструктуры и реконструкции

Адвертительная программа текущего ремонта дорог на 2012 год

Перевод адресов текущего ремонта, выполняемого на платном обслуживании

Перевод адресов платного текущего ремонта, выполняемого в рамках платного обслуживания

Дорожные фонды

Адреса перемещаемых автомобилей в Санкт-Петербурге

График работы инспекции в течение 2012 года

Организованное движение транспорта

Государственные услуги

Решение дорог 2014
Информация о программах мероприятий "Восстановление"

both are
represented via
<table>

[About Us](#) | [Contact Us](#) | [Help](#) | [Sitemap](#) | [Research Help](#)

my Account
[Register](#)

username

☐ Remember login?
[Forgotten password?](#)

- [Scripts & Text](#)
- [Direct Online](#)
- [AP Instant Chat](#)
- [News Guide](#)
- [Email Alerts](#)

Media Products

- News
- Middle East Extra
- Direct
- Entertainment
- Horizons
- Technology
- sntv
- Archive
- GraphicsBank
- Photostream
- Images
- International Text
- DataStream
- Headline
- Global Media Services
- Events
- Customised Production
- Global Facilities
- Assignments

MONDAY, 18 August 2014 9:25 GMT

ASSOCIATED PRESS

Delivering breaking global news, sport, entertainment, technology and

[Find out more](#)

AP IS NOW HD FOR VIDEO NEWS

News, Sports, Entertainment, Lifestyles and Technology

... access breaking news as it happens

AP Direct is the **award-winning premium LIVE video news service**, which offers incoming LIVE video news from AP's fixed and mobile links from around the world enabling you to access breaking news and scheduled events in real-time.

Horizons: World War I

From 18 to 22 August 2014, Horizons will feed content looking back at World War I and how it changed the world

Global Media Services

Current/Upcoming Events

- July and August 14 **BROOK**
Middle East Ceasefire *MULTIPLE LOCATIONS*
- Ongoing **BROOK**
Iraq Unrest *MULTIPLE LOCATIONS*
- 14 to 18 August 14 **BROOK**
Pope Francis to Visit South Korea

[View All Events](#)

Broadcast Schedule

Now: CCTV News Content 0915
Next: Prime News - Americas

[View Broadcast Schedule](#)

Planned Coverage

- 48 August 2014
Vince Neil, Nikki Stixx and Justin Moore Talk Covers
Vince Neil, Nikki Stixx and Justin Moore talk about country covers of Motley Crue classics. (Entertainment Americas Early bulletin)
- 48 August 2014
DJ Cassidy Talks Style and Influence
American producer DJ

[View All](#)

Looking for LIVE breaking news video content? Sign-up for a trial >

News

Machine Learning

Which data table did we find?

- Orientation assessment

it is essential to distinguish concepts from their properties

	Header 1	Header 2	Header 3
Obj1			
Obj2			
Obj3			

horizontal orientation

	Obj1	Obj2	Obj3
Header 1			
Header 2			
Header 3			

vertical orientation

Machine Learning

String Metric → **Heuristics** → **ML**

Levenshtein

Horizontal
cell similarity

Naive Bayes

Jaro-Winkler

kNN

n-grams

Vertical cell
similarity

J48

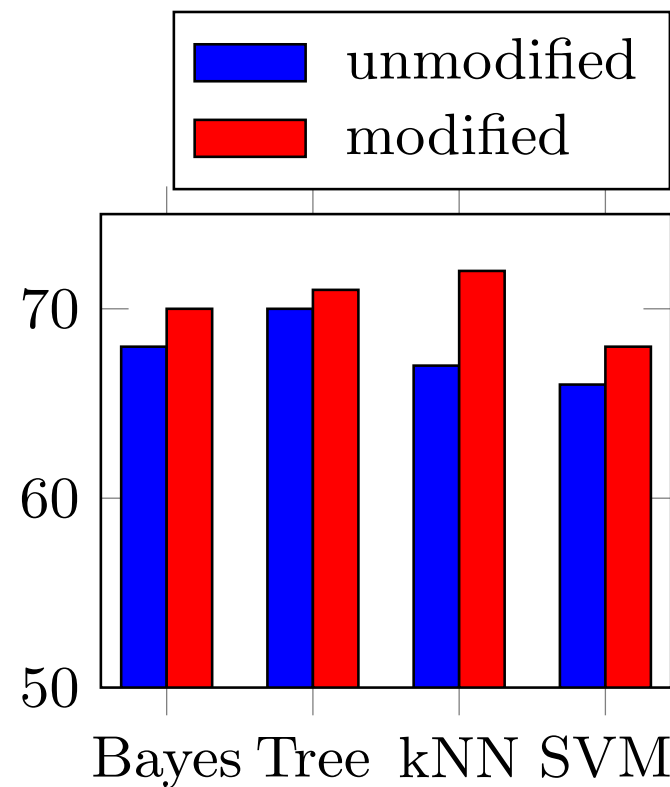
Improvements

1. All numbers are the same
2. Fixed similarity for long sentences

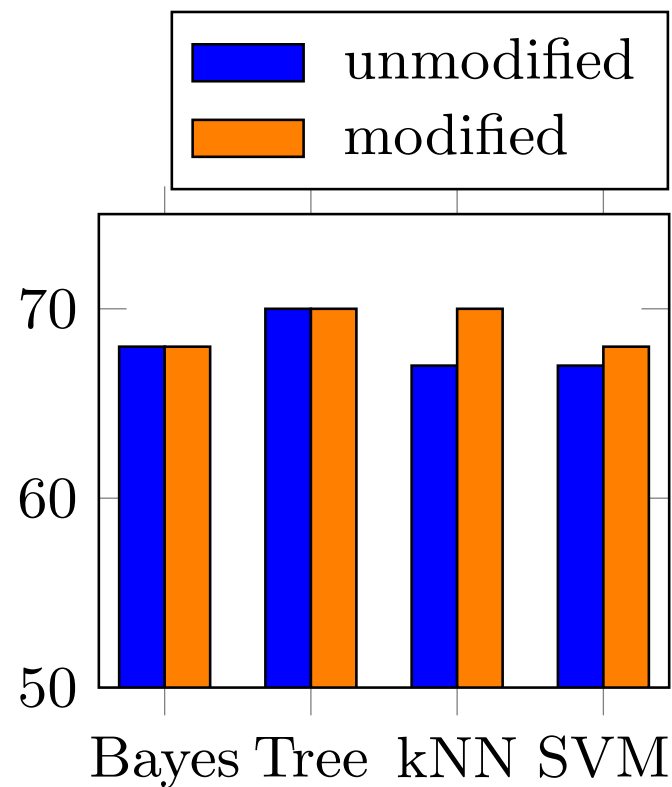
Max/avg

SVM

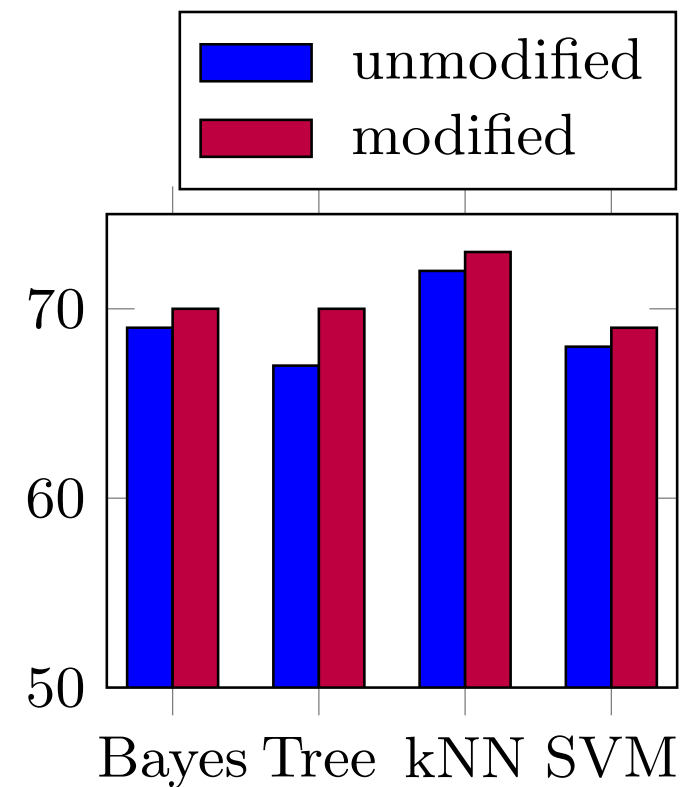
Evaluation



(a) Levenshtein



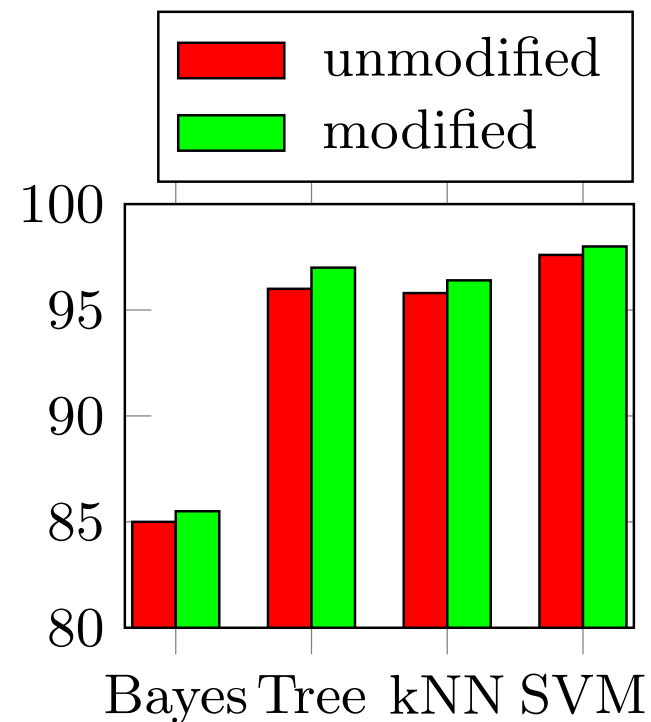
(b) Jaro-Winkler



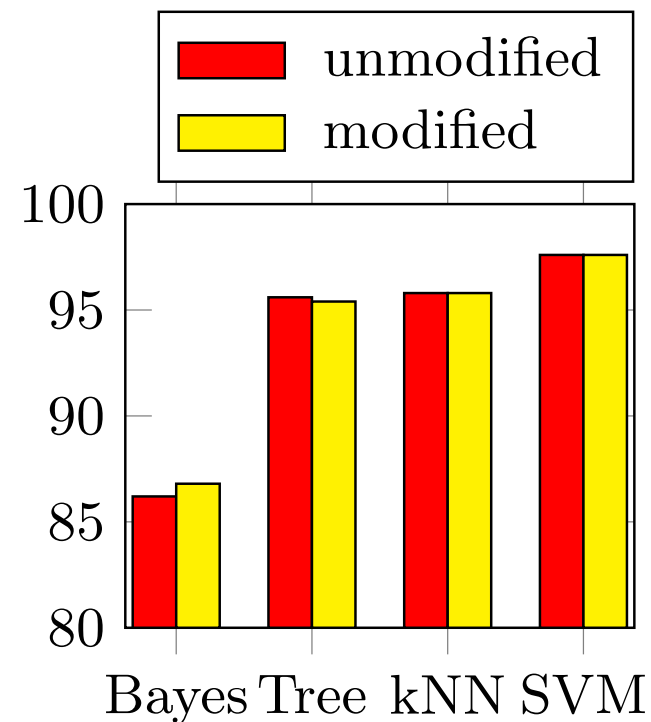
(c) n-grams

Genuineness assessment, F-Measure

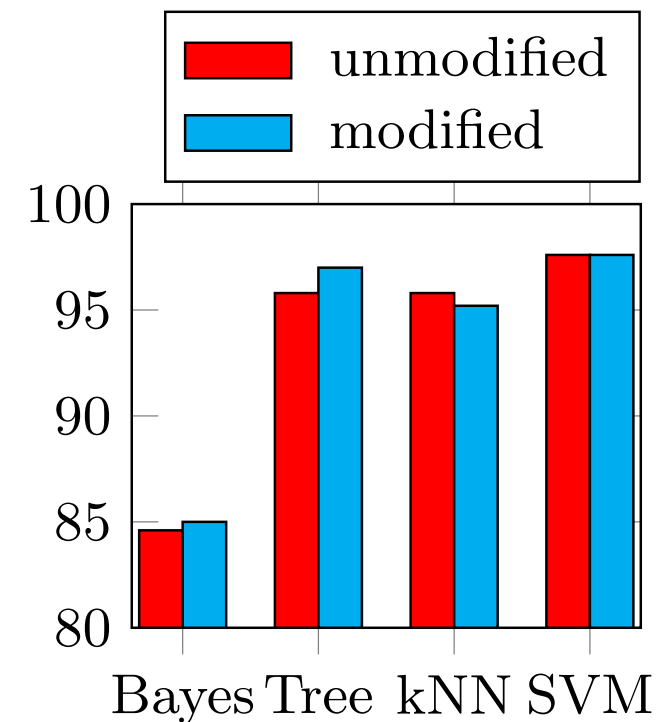
Evaluation



(a) Levenshtein



(b) Jaro-Winkler



(c) n-grams

Orientation assessment, Precision

Applications

- Enrich enterprise or web Knowledge Graphs with billions of facts from the most neglected content type
- Business domain - refine openly published tables with financial KPIs for the data analysis;
Maintain connection with the Enterprise Knowledge Graph;
Use semantic technologies for business processes re-engineering.