Integrating Product Data from the Web
Hallo

- **Prof. Dr. Christian Bizer**
- Professor for Information Systems V
- Research Interests:
  - Web Data Integration
  - Data and Web Mining
  - Linked Data Technologies
- Room: B6 - B1.15
- eMail: chris@informatik.uni-mannheim.de
- Consultation: Wednesday, 13:30-14:30
Hallo

- **Anna Primpeli**
- Graduate Research Associate
- Research Interests:
  - Data Extraction
  - Web Data Integration
  - Active Learning
  - Structured Data on the Web
- Room: B6, 26, C 1.04
- eMail: anna@informatik.uni-mannheim.de
Agenda of Today‘s Kickoff Meeting

1. Introduction and Project Goals
2. The Product Data Corpus
3. Organization and Schedule
4. Specific Subtasks
Motivation of the Team Project

The Web is a rich source of product information

- the same product is described by 100s of websites
  - by merchants (offers)
  - the producer (product specs)
  - by consumers (reviews and ratings)

If we can determine which pages describe a specific product **(identity resolution)**, we are able to

- build comprehensive product catalogues and search engines
- conduct global price comparison engines
- understand market structure and consumer preferences
Identity Resolution is the Key Task for Downstream Applications
Features that help us to Distinguish Products on the Web

- **Product Identifiers**
  - GTINs, UPCs, ISBNs, MPN, ...
- **Product Titles**
  - Product name plus selected features
- **Product Descriptions**
  - long free texts
- **Specification Tables and Lists**
  - Detailed features as key/value pairs
- **Product Pictures**

**Before** we can use these features:

- values need to be cleansed and normalized
- we might want to apply information extraction in order to increase the structuredness of data

Das Samsung Galaxy S4 ist der unterhaltsame und hilfreiche Begleiter für Ihr mobiles Leben. Es verbindet Sie mit Ihren Liebsten. Es lässt Sie gemeinsam unvergessliche Momente erleben und festhalten. Es vereinfacht Ihren Alltag.
Difficulty of the Task depends on the Product Category

- **Books**
  - wide adoption of identification schema (ISBNs)
  - problem mostly solved 😊
  - other features like title and author often only used for sanity checks

- **Phones / Computers / Cameras**
  - rather structured descriptions, often including tables/lists
  - different sites often describe same features
  - identity resolution methods for structured data can be applied

- **Cloths / Bags / ....**
  - rather unstructured descriptions, not too many tables/lists
  - only weak agreement of attributes
  - identity resolution / disambiguation methods for texts need to be applied (bag of words methods)
## Identity Resolution Methods

<table>
<thead>
<tr>
<th>Text-oriented</th>
<th>Attribute-oriented</th>
</tr>
</thead>
<tbody>
<tr>
<td>(bag of words)</td>
<td>(weighted matching rules)</td>
</tr>
<tr>
<td>Unsupervised</td>
<td>Supervised</td>
</tr>
<tr>
<td>(TF/IDF+Cosine)</td>
<td>(random forest)</td>
</tr>
<tr>
<td>Symbolic</td>
<td>Sub-symbolic</td>
</tr>
<tr>
<td>(matching rules)</td>
<td>(embeddings)</td>
</tr>
</tbody>
</table>
Project Goals

- Integrate product data from a large number of websites, using:
  - different identity resolution methods
  - different information extraction methods

- Compare performance of methods w.r.t.:
  - product categories (structured vs. semi-structured input)
  - product popularity (head vs. tail products)

- Evaluate the usefulness of weak supervision that is found on the Web for product matching
  - weak supervision = product identifiers such as GTINs
Learning Targets

Improve your technical skills
- Work as a Data Scientist: clean, profile, integrate, classify data, classify record pairs
- Understand the nature of Web Data
- Improve your technical expertise / programming skills

Improve your soft skills
- Work as part of a bigger team on a more complex project
- Organize yourself and assign tasks based on your skills
- Communicate and coordinate your work
2. The Product Data Corpus

1. Semantic Annotations in HTML Pages
2. Web Data Commons Project
3. Web Data Commons – Silver Standard for Large-Scale Product Matching
Semantic Annotation of HTML Pages: Schema.org

- ask site owners since 2011 to annotate data for enriching search results
- 675 Types: Event, Place, Local Business, Product, Review, Person
- Encoding: Microdata, RDFa, JSON-LD
Usage of Schema.org Data @ Google

Data snippets within search results

Data snippets within info boxes
Example: Microdata Annotations in HTML

```html
<div itemtype="http://schema.org/Hotel">
  <span itemprop="name">Vienna Marriott Hotel</span>
  <span itemprop="address" itemscope="" itemtype="http://schema.org/PostalAddress">
    <span itemprop="streetAddress">Parkring 12a</span>
    <span itemprop="addressLocality">Vienna</span>
  </span>
  <div itemprop="aggregateRating" itemscope itemtype="http://schema.org/AggregateRating">
    <span itemprop="ratingValue">4</span> stars-based on
    <span itemprop="reviewCount">250</span> reviews.
  </div>
</div>
```
Product-related schema.org Classes

The Web Data Commons Project

- extracts all Microformat, Microdata, RDFa, JSON-LD data from the Common Crawl
- analyzes and provides the extracted data for download
- statistics about some extraction runs
  - 2017 CC Corpus: 3.1 billion HTML pages → 38.2 billion RDF triples
  - 2016 CC Corpus: 3.1 billion HTML pages → 44.2 billion RDF triples
  - 2014 CC Corpus: 2.0 billion HTML pages → 20.4 billion RDF triples
  - 2012 CC Corpus: 3.0 billion HTML pages → 7.3 billion RDF triples
- uses 100 machines on Amazon EC2
  - approx. 2000 machine/hours
    (100 spot instances of type c3.xlarge) → 350 Euro
- http://www.webdatacommons.org/structureddata/
schema.org Annotations: Most Popular Classes

Development of Selected Classes by #PLDs

http://webdatacommons.org/structureddata/
## Properties used to Describe Products 2017

<table>
<thead>
<tr>
<th>Top 15 Properties</th>
<th>PLDs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#</td>
</tr>
<tr>
<td>schema:Product/name</td>
<td>535,625</td>
</tr>
<tr>
<td>schema:Offer/price</td>
<td>462,444</td>
</tr>
<tr>
<td>schema:Product/offers</td>
<td>462,233</td>
</tr>
<tr>
<td>schema:Offer/priceCurrency</td>
<td>430,556</td>
</tr>
<tr>
<td>schema:Product/image</td>
<td>419,391</td>
</tr>
<tr>
<td>schema:Product/description</td>
<td>377,639</td>
</tr>
<tr>
<td>schema:Offer/availability</td>
<td>337,876</td>
</tr>
<tr>
<td>schema:Product/url</td>
<td>263,720</td>
</tr>
<tr>
<td>schema:AggregateRating/ratingValue</td>
<td>184,004</td>
</tr>
<tr>
<td>schema:Product/sku</td>
<td>126,696</td>
</tr>
<tr>
<td>schema:AggregateRating/reviewCount</td>
<td>112,408</td>
</tr>
<tr>
<td>schema:Product/aggregateRating</td>
<td>101,434</td>
</tr>
<tr>
<td>schema:Product/brand</td>
<td>73,934</td>
</tr>
<tr>
<td>schema:Product/productID</td>
<td>35,211</td>
</tr>
<tr>
<td>schema:Product/manufacturer</td>
<td>21,967</td>
</tr>
</tbody>
</table>

Das Samsung Galaxy S4 ist der unterhaltsame und hilfreiche Begleiter für Ihr mobiles Leben. Es verbindet Sie mit Ihren Liebsten. Es lässt Sie gemeinsam unvergessliche Momente erleben und festhalten. Es vereinfacht Ihren Alltag.

UPC 610214632623
The WDC Silver Standard for Large-Scale Product Matching

- Silver Standard grouping schema.org product annotations by identifier value.
  - all WDC product data is included that
  - provides come sort of product ID
- Initial cleaning steps are performed
- Results clusters of product descriptions from different websites that share identifier values.

Details and Download
http://webdatacommons.org/largescaleproductcorpus/index.html
The WDC Silver Standard for Large-Scale Product Matching

- Distribution of offer entities per ID-cluster

- Distribution of identifier values per ID-cluster

- Distribution of offer entities per TLD

3. Organization and Schedule

**Duration**: 6 months (28.09.2018 – 28.03.2018)

**ECTS**: 12

**Participants**: 8 people

**Type of work**: Team and subgroup based

**Milestones**: 4 project phases

**Evaluation**:

- Individual contribution to the deliverables
- Deliverables: Presentations, final report, code, data
- Every project phase determines 25% of your final grade
Questions and Subtasks

1. Which two product categories should we use? → Corpus Profiling, Data Selection
2. Is the corpus for these categories dense and clean enough? → Corpus Profiling, Additional Crawling
3. How do supervised and unsupervised IR methods perform? → Basic IR
4. Which features should be extracted to help IR? → Feature Extraction
5. How do IR methods perform given cleaner features? → IR with Feature Extraction
6. How do your different methods compare? → Comparison of Results
Main Steps of the Project

- Data Collection
- IR with BoW
- Feature Extraction
- IR with Enhanced Features
- Comparison of Results
## Detailed Schedule

<table>
<thead>
<tr>
<th>Date</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>Friday, 28.09.2018</td>
<td>Kickoff meeting (today)</td>
</tr>
<tr>
<td><strong>Phase 1 (all members): Data Collection</strong></td>
<td></td>
</tr>
<tr>
<td>Friday, 12.10.2018</td>
<td>Meet Anna and report current results</td>
</tr>
<tr>
<td>Friday, 12.10.2018</td>
<td>Drop-out deadline: Dropping out after this date will result in failing the team project</td>
</tr>
<tr>
<td>Friday, 26.10.2018</td>
<td><strong>1st Deliverable:</strong> 20 minutes presentation, data</td>
</tr>
<tr>
<td><strong>Phase 2 (in subgroups): IR with BoW</strong></td>
<td></td>
</tr>
<tr>
<td>Friday, 16.11.2018</td>
<td>Meet Anna and report current results</td>
</tr>
<tr>
<td>Friday, 30.11.2018</td>
<td><strong>2nd Deliverable:</strong> 20 minutes presentation from each subgroup, code &amp; data</td>
</tr>
<tr>
<td><strong>Phase 3 (in subgroups): Feature Extraction</strong></td>
<td></td>
</tr>
<tr>
<td>Friday, 11.01.2019</td>
<td>Meet Anna and report current results</td>
</tr>
<tr>
<td>Friday, 25.01.2019</td>
<td><strong>3rd Deliverable:</strong> 20 minutes presentation from each subgroup, code &amp; data</td>
</tr>
<tr>
<td><strong>Phase 4 (in subgroups): IR with Features</strong></td>
<td></td>
</tr>
<tr>
<td>Friday, 01.03.2019</td>
<td>Meet Anna and report current results</td>
</tr>
<tr>
<td>Monday, 25.03.2019</td>
<td><strong>4th Deliverable:</strong> 15-20 pages overall report, code &amp; data</td>
</tr>
<tr>
<td>Friday, 29.03.2019</td>
<td>Overall presentation 30 min + Feedback</td>
</tr>
</tbody>
</table>
Phase 1: Data Collection

Participants: All team members
Deliverables: 20 minutes presentation, data, report who did what

Tasks

1. Decide on two product categories
   • Select 2 non-similar product categories, one structured, one less structured e.g. laptops and shoes (NOT phones, headphones, TVs)

2. Decide on a set of products
   • Collect a set of products from each category together with respective IDs (>100 products/category)

3. Create your subcorpus
   • Identify the relevant ID-clusters from WDC Large-Scale GS using the product identifiers
   • Profile the data / perform additional cleansing steps / maybe crawl additional data
   • Report detailed statistics about the initial and final subcorpus (cluster sizes, feature frequency)
Expected Result of Phase 1

- Clean gold standard as basis for evaluating different identity resolution methods in the following phases.

- Expected profile of your gold standard
  - >=2 categories
  - >100 products per category
  - >10 and median 20 pages from different PLDs per product
    - majority of PLDs should be .com/ co.uk
  - All pages should contain schema:title, schema:description, and a product ID (not necessarily annotated).
  - One category: Rather structured product descriptions containing detailed specification tables/lists
  - Other category: Less structured descriptions, not necessarily containing tables/lists.

- Expected format of the Gold Standard
  - Same format as WDC Gold Standard for Product Matching and Product Feature Extraction
  - [http://www.webdatacommons.org/productcorpus/](http://www.webdatacommons.org/productcorpus/)
Phase 2: Identity Resolution with BoW Models

**Duration:** 26.10.2018 – 30.11.2018

**Participants:** 2 subgroups of 4 persons each

**Deliverables:**
- 20 minutes presentation from each subgroup, data & code, report who did what

**Tasks**

Use BoW models with different input data and apply:

**Subgroup 1: Unsupervised IR methods**
- TF/IDF+cosine, embeddings, domain-specific heuristics

**Subgroup 2: Supervised IR methods**
- word weights, decision trees, random forests, deep learning

Evaluate on the WDC Silver Standard
Phases 3: Feature Extraction

Duration: 30.11.2018 – 25.01.2018

Participants: 2 subgroups of 4 persons each

Deliverables:
- 20 minutes presentation from each subgroup, data & code, report who did what

Tasks
Apply advanced feature extraction methods

Subgroup 1: Closed Feature Extraction methods
- dictionary based using auxiliary data for product properties and values

Subgroup 2: Open Feature Extraction methods
- Exploit HTML tables and HTML lists
- Perform schema matching on extracted data

Evaluate against WDC Gold Standards for Product Feature Extraction and Product Matching or manually generated custom gold standard.
Phases 4: Identity Resolution with Enhanced Features

Duration: 25.01.2018 – 25.03.2018

Participants: 2 subgroups of 4 persons each

Deliverables:
- 30 minutes presentation
- 15-20 pages overall report, data & code, report who did what

Tasks
Apply IR methods that exploit the enhanced features
Compare your results
Formal Requirements & Consultation

- **Deliverables**
  - The final report should be 15-20 pages single column
    - including appendixes, not including the bibliography
  - Presentations: PDF version of slides must be send until deadline
  - Every deliverable should be accompanied with an excel sheet stating which team member conducted which subtask.

- **Final grade**
  - 25% for every phase, individual grade / not per team
  - Late submission: reduction of grade by 0.3 per day

- **Consultation**
  - Send one e-mail per team stating your questions to Anna, she answers questions or meets with you
  - Chris does second level support and gives feedback at presentations
How to Structure Your Deliverables?

1. Problem definition
2. Profiling of your subcorpus
3. Methodology
4. Evaluation Results
5. Error Analysis
6. Conclusion

Please provide comprehensive statistics that allow us to understand what is happening.

Accompany your deliverables with the code and data you used

! The phase deadlines apply for the submission of your code and data as well
Submission of Deliverables

Presentation Slides

– Send slide until the submission deadline.
– The exact time of the presentation will be determined case by case.

Team and Subgroup Reports

– Send one e-mail per team or subteam until the deadline date according to the schedule.

Data and Code

– Add your data and code in a zipped folder and send (URL) via e-mail.

Member to subtask report

– Send one excel sheet per team explaining who did what together with the deliverables.

All deliverables should be sent to Chris & Anna!
References: Identity Resolution in General

Lecture Slides
- Bizer: Web Data Integration – Chapter: Identity Resolution, 2017 (see lecture archive)

Book Chapters
- Peter Christen: Data Matching. Springer 2012.

Papers
References: Identity Resolution for Product Data


References: Deep Learning for Identity Resolution


Related Work for Feature Extraction


- Zheng, Mukherjee, Dong: OpenTag: Open Attribute Value Extraction from Product Profiles. KDD, 2018.


Potentially Useful Software

- **Identity Resolution**
  - Winter Framework: [https://github.com/olehmberg/winter](https://github.com/olehmberg/winter)
  - Silk Framework: [https://github.com/silk-framework/silk](https://github.com/silk-framework/silk)
  - DeepMatcher: [https://github.com/anhaidgroup/deepmatcher](https://github.com/anhaidgroup/deepmatcher)

- **Information Extraction**
  - Specification table classifier (template project): [https://github.com/petrovskip/wstl-extractor](https://github.com/petrovskip/wstl-extractor)
  - Feature extraction gold standard tool: [https://github.com/aprimpeli/LabellingTool](https://github.com/aprimpeli/LabellingTool)
  - Stanford NLP: [https://nlp.stanford.edu/software/](https://nlp.stanford.edu/software/)

- **Crawling**
  - Scrapy: [https://scrapy.org/](https://scrapy.org/)
The Project Team

1. Michael, Anne Katrin
2. Ly, Duc Tai
3. Le, Phuong Anh
4. Zhang, Shenghan
5. Yeu, Se Won
6. Amedani, Jurgen
7. Shkrepa, Lerida
8. Erazo Guevara, Maria Alejandra

- A Short Round of Introductions
  - What are you studying? Which semester?
  - Which DWS courses did you already attend?
  - What are your programming and data wrangling skills?
Project Infrastructure?

- **Shared Document Space**
  - for todo lists, brainstorming documents
  - Google Docs? Wiki?

- **ILIAS Group**
  - mailing to all participants
  - for sharing files

- **Code Repository**
  - GitHub?

- **Data Repository**
  - Google Drive? Dropbox?

- **Anything else?**