Integrating Product Data from the Web
Main Steps of the Project

1. Data Collection
2. IR with BoW
3. Feature Extraction
4. IR with Enhanced Features
5. Comparison of Results
## Detailed Schedule

<table>
<thead>
<tr>
<th>Date</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Friday, 28.09.2018</strong></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>1&lt;sup&gt;st&lt;/sup&gt; Deliverable: 20 minutes presentation, data</td>
</tr>
<tr>
<td><strong>Phase 2 (in subgroups): IR with BoW + Gold Standard Creation</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>2&lt;sup&gt;nd&lt;/sup&gt; Deliverable: 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>3&lt;sup&gt;rd&lt;/sup&gt; Deliverable: 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>4&lt;sup&gt;th&lt;/sup&gt; Deliverable: 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

- Product offer dataset as basis for evaluating different identity resolution methods in the following phases.

- Expected profile of your data set
  - >=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 dataset
  - Same format as WDC Gold Standard for Product Matching and Product Feature Extraction
  - http://www.webdatacommons.org/productcorpus/
Phase 1: Checklist

What you should have by now?

- Offer pages with identifiers (annotated or not)
- For every page you have
  - the schema.org descriptive properties (title, name, description)
  - the whole html content
- Groups of offers with the same identifiers (ID-clusters)
- Groups of offers with the same identifiers after normalization (ID-clusters)
  e.g.: id1: 123-45, id2:12_345 → id:12345
- Profiling results of your offers corpus and the ID-clusters
Phase 2: Identity Resolution with BoW Models + Gold Standard Creation

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 forrests, deep learning

Evaluate on the WDC Silver Standard using a category-specific gold standard.
How to Build an Good Gold Standard?

- You need **ground truth (gold standard)** for your evaluation.
- To create a gold standard, manually label a set of record pairs as **matches** or **non-matches** including **corner cases**
- Rule of thumb for creating an **interesting** gold standard with **acceptable** manual effort:
  1. match records using several simple matching techniques (similar to multi-pass blocking) and
  2. reuse existing information about matches (e.g. ISBN or GTIN numbers that exist in multiple sources)
  3. manually verify a fair amount of the resulting pairs (e.g. 500 pairs) including
     1. matching record pairs (randomly chosen, 20% of GS)
     2. corner cases (30% of GS)
     3. non-matching record pairs (randomly chosen, 50% of GS)
How Select Pairs of Offers for the Gold Standard?

1. Sort ID-Clusters within each product category by amount of offers.
2. Randomly choose 100 large clusters and 50 mid-size clusters.
3. Calculate the Jaccard-Similarity of the product titles within each cluster and sort pairs by similarity.
4. Calculate the Jaccard-Similarity of a product title from each cluster and the product titles of all offers within other clusters of the same category.
5. Manually verify the following pairs for your gold standard:
   1. Normal positives: Randomly choose 1 matching pair from the top-3 similarity quantiles of each cluster.
   2. Corner cases: Randomly choose 1 pair from the fourth similarity quantile of each cluster.
   3. Negatives: Randomly choose 3 pairs top-2 similarity quantiles of the out of cluster pairs.

Rather similar records that are not a match
Rather different records that are a match
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?
Questions?
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?