Web Mining

Introduction and Course Organization

Prof. Dr. Christian Bizer
Prof. Dr. Simone Ponzetto
Keti Korini
Alexander Brinkmann

FSS 2024
Hallo

Prof. Dr. Christian Bizer
Professor for Information Systems V
Research Interests:
- Information Extract from the Web
- Large-Scale Data Integration
- Data and Web Mining
Room: B6, 26 - B1.15
Consultation: Wednesday 13:30-14:30
eMail: christian.bizer@uni-mannheim.de

will teach the lectures on Web Usage Mining and Web Structure Mining
Hallo

Prof. Dr. Simone Ponzetto

Professor for Information Systems III

Research Interests:
- Natural Language Processing
- Computational Social Science

Room: B6 26, B 1.14

Consultation: Tuesday 13:30-14:30

eMail: ponzetto@uni-mannheim.de

will teach the lectures on Web Content Mining
Hallo

- M. Sc. Wi-Inf. Keti Korini
- Graduate Research Associate
- Research Interests:
  - Schema Matching
  - Table Annotation using Deep Learning
- Room: B6, 26, C 1.03
- eMail: kkorini@uni-mannheim.de

will teach the labs and will supervise the student projects (IE684)
Hallo

M. Sc. Wi-Inf. Alexander Brinkmann
Graduate Research Associate

Research Interests:
- Data Search using Deep Learning
- Semantic Annotations in Web Pages
- Product Data Categorization

Room: B6, 26, C 1.03
eMail: alexander.brinkmann@uni-mannheim.de

will teach the labs and will supervise the student projects (IE684)
Introduction and Course Organization

1. Course Organization

2. The World Wide Web
   1. The Classic Document Web
   2. The Web of Data
   3. Web 2.0 Applications

3. What is Web Mining?
   1. Web Usage Mining
   2. Web Structure Mining
   3. Web Content Mining
1. Course Organization

- **Lecture (IE 671, 3 ECTS)**
  - covers different types of Web Mining methods
  - presents examples of Web Mining applications
  - discusses how to evaluate the learned models

- **Labs**
  - students experiment with the methods using different Python libraries

- **Evaluation**
  - 60 min written exam + registration for exam via Portal2 until 24.April

- **Student Projects (IE 684, 3 ECTS, 50 places)**
  - teams of five students realize a Web Mining project
  - teams may choose their own tasks and data sets
    (in addition, we will propose some suitable data sets and tasks)
  - write a summary about the project, present the project results

- **Evaluation**
  - report + presentation + code + individual registration via Portal2 until 24.April
Course Organization

- **Website**
  - organizational information
  - lecture slides + exercise material

- **Two ILIAS Groups**
  - IE 671 Web Mining [V]
  - IE 684: Web Mining Project [PRO]

both used as forum and mailing list.

- **Time and Location**
  - Lecture: Tuesday, 10:15 to 11:45, B6 A101
  - Labs: Thursday, 13:45 to 15:15, B6, A203
<table>
<thead>
<tr>
<th>Week</th>
<th>Lecture: Tuesday 10:15</th>
<th>Labs: Thursday 13:45</th>
</tr>
</thead>
<tbody>
<tr>
<td>13.02.2024</td>
<td>Introduction and Course Outline</td>
<td></td>
</tr>
<tr>
<td>20.02.2024</td>
<td>Web Usage Mining (Bizer)</td>
<td>Web Usage Mining</td>
</tr>
<tr>
<td>27.02.2024</td>
<td>Web Usage Mining (Bizer)</td>
<td>Web Usage Mining</td>
</tr>
<tr>
<td>05.03.2024</td>
<td>Web Content Mining (Ponzetto)</td>
<td>Web Content Mining</td>
</tr>
<tr>
<td>12.03.2024</td>
<td>Web Content Mining (Ponzetto)</td>
<td>Web Content Mining</td>
</tr>
<tr>
<td>19.03.2024</td>
<td>Web Structure Mining (Bizer)</td>
<td>Web Structure Mining</td>
</tr>
<tr>
<td></td>
<td>- Easter break -</td>
<td></td>
</tr>
<tr>
<td>09.04.2024</td>
<td>Web Structure Mining (Bizer)</td>
<td>Web Structure Mining</td>
</tr>
<tr>
<td>16.04.2024</td>
<td>Introduction to the Student Projects</td>
<td>Preparation of project outlines</td>
</tr>
<tr>
<td>23.04.2024</td>
<td>Feedback on the projects outline</td>
<td>Project work</td>
</tr>
<tr>
<td>30.04.2024</td>
<td>Project work</td>
<td>Coaching</td>
</tr>
<tr>
<td>07.05.2024</td>
<td>Project work</td>
<td>Coaching</td>
</tr>
<tr>
<td>14.05.2024</td>
<td>Project work</td>
<td>Coaching</td>
</tr>
<tr>
<td>21.05.2024</td>
<td>Project presentations</td>
<td>-</td>
</tr>
</tbody>
</table>
Literature

Software Libraries

- Within the Labs, we use the following libraries

[Images of software libraries: Surprise, NetworkX, PyG, scikit-learn]

- We assume some fluency in Python and Jupyter ...
Questions?
2. The World Wide Web

The Web is a global decentralized information space build on a set of technical standards for the identification, retrieval and representation of content.

- **Uniform Resource Identifiers (URIs):** Globally unique identification of Web resources.
- **Hypertext Transfer Protokoll (HTTP):** Protocol for interacting with Web resources.
- **Content Formats:** HTML, XML, RDF, …
- **The Web was invented in 1989 at CERN by Tim Berners-Lee**
- **Architectural Principles of the Web**
  
Topology of the Web Today

The Classic Document Web

The Web of Data

Web 2.0 Applications
(accessible via HTML interface and APIs)
2.1 The Classic Document Web

Global information space consisting of **interlinked resources** (HTML pages, images, multimedia).

**The Size of the Web**

- **overall size**: 1 trillion pages on the Web at once
  - announced by Google in 2008
  - [http://googleblog.blogspot.de/2008/07/we-knew-web-was-big.html](http://googleblog.blogspot.de/2008/07/we-knew-web-was-big.html)

- **Indexed Web**: approx. 50 billion pages
  - estimate based on search engine hit counts for popular words
  - Example: The word "the" appears in 67% of all English pages and has 25.2 billion hits on Google
Public Web Corpora

- The non-profit Common Crawl project regularly crawls the Web and publishes large Web corpora
  - size: 2.5 to 3.5 billion pages
  - pay-level-domains: 30 – 35 million
  - release cycle: 1 month

- Public download from Amazon S3
  - size: around 80 Terrabytes compressed

- Public alternative to private corpora owned by Google and Microsoft

- Disadvantages:
  - one order of magnitude smaller than private crawls
  - monthly releases compared to permanent updates
Link Structure of the Web: In-Degree

- The link distribution follows (kind of) a power law
  - A small number of pages is target of many links
  - A large number of pages is target of only a few or no links

- Classic Paper:
  - Broder at al.: Graph Structure in the Web. WWW2000
  - AltaVista crawl with over 200 million pages and 1.5 billion links
  - Conclusion: Log-log scale plot shows power-law
In-Degree Distribution

Broder et al. (2000)
Power law with exponent 2.1
(200 million pages and 1.5 billion links from Altavista crawl 2000)

WDC Hyperlink Graph (2012)
Best power law exponent 2.24
(3 billion pages and 128 billion links from Common Crawl 2012)
Link Structure of the Web: Bow-Tie

Four mayor components (Border at al., WWW2000)

- Central Strongly Connected Component (SCC)
  - pages that can reach one another along directed links
  - about 30% of the Web (normal pages)

- IN Group
  - can reach SCC but cannot be reached from it
  - about 20% (maybe new pages or boring ones)

- OUT Group
  - can be reached from SCC but cannot reach it
  - about 20% (maybe company pages that don’t link)

- Tendrils
  - cannot reach SCC and cannot be reached by it
  - about 20%

- Unconnected
  - about 10%

Probability of path between nodes is 24%
A strongly connected component (SCC) in a directed graph is a subset of the nodes such that:

1. every node in the subset has a path to every other node
2. the subset is not part of some larger set with the property that every node can reach every other.
Size of Central Strongly Connected Component

Largest SCC

- Broder, 2000: 27.7%
- WDC, 2012: 51.3%

→ Factor 1.8 larger
→ Also, factor 4.9 more links/page
→ The Web has become denser
2.2. The Web of Data

More and more Websites
- semantically *markup the content* of their HTML pages
- publish *structured data* in addition to HTML

Markup Formats:  
- Microdata
- RDFa

Structured Data Formats:  
- JSON-LD
- Linked Data

Data is crawl-able using generic crawlers in contrast to Web APIs
Microdata

- format for annotating structured data within webpages
- proposed in 2009 by WHATWG as part of HTML5 work

```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>
```
JSON-LD

- used for embedding data into the HEAD of HTML pages
- putting data in HEAD is recommended by Google as it is empirically less error prone than annotations in BODY

```html
<script type="application/ld+json">
{
  "@context": "http://schema.org",
  "@type": "Product",
  "description": "Has six preset cooking ....",
  "name": "Kenmore White 17\\" Microwave",
  "offers": {
    "@type": "Offer",
    "availability": "http://schema.org/InStock",
    "price": "55.00",
    "priceCurrency": "USD"
  }
}
</script>
```
schema.org

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

Data snippets within search results

Local businesses on maps

Data snippets within info boxes
Usage of Schema.org Data @ Google

https://developers.google.com/search/docs/guides/search-gallery
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
  - 2023 CC Corpus: 3.4 billion HTML pages → 97.6 billion RDF triples
  - 2019 CC Corpus: 2.4 billion HTML pages → 44.2 billion RDF triples
  - 2014 CC Corpus: 2.0 billion HTML pages → 20.4 billion RDF triples
  - 2010 CC Corpus: 2.8 billion HTML pages → 5.1 billion RDF triples
- uses 100 machines on Amazon EC2
  - approx. 3500 machine/hours → 1000 Euro
- http://webdatacommons.org/structureddata/
Overall Adoption 2023

1.7 billion HTML pages out of the 3.4 billion pages provide semantic annotations (50.6%).

15 million pay-level-domains (PLDs) out of the 34 million pay-level-domains covered by the crawl provide semantic annotations (42.9%).

http://webdatacommons.org/structureddata/2023-12/stats/stats.html
# Frequently used Schema.org Classes

<table>
<thead>
<tr>
<th>Class</th>
<th># Websites (PLDs)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Microdata</td>
</tr>
<tr>
<td>schema:WebPage</td>
<td>1,124,583</td>
</tr>
<tr>
<td>schema:Product</td>
<td>812,205</td>
</tr>
<tr>
<td>schema:Offer</td>
<td>676,899</td>
</tr>
<tr>
<td>schema:BreadcrumbList</td>
<td>621,344</td>
</tr>
<tr>
<td>schema:Article</td>
<td>612,361</td>
</tr>
<tr>
<td>schema:Organization</td>
<td>510,069</td>
</tr>
<tr>
<td>schema:PostalAddress</td>
<td>502,615</td>
</tr>
<tr>
<td>schema:ImageObject</td>
<td>360,875</td>
</tr>
<tr>
<td>schema:Blog</td>
<td>337,843</td>
</tr>
<tr>
<td>schema:Person</td>
<td>324,349</td>
</tr>
<tr>
<td>schema:LocalBusiness</td>
<td>294,390</td>
</tr>
<tr>
<td>schema:AggregateRating</td>
<td>258,078</td>
</tr>
<tr>
<td>schema:Review</td>
<td>124,022</td>
</tr>
<tr>
<td>schema:Place</td>
<td>92,127</td>
</tr>
<tr>
<td>schema:Event</td>
<td>88,130</td>
</tr>
</tbody>
</table>

http://webdatacommons.org/structureddata/2018-12/
### Adoption by Travel Websites

<table>
<thead>
<tr>
<th>Top 15 Travel Websites</th>
<th>schema:Hotel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Booking.com</td>
<td>Yes</td>
</tr>
<tr>
<td>TripAdvisor</td>
<td>Yes</td>
</tr>
<tr>
<td>Expedia</td>
<td>Yes</td>
</tr>
<tr>
<td>Agoda</td>
<td>Yes</td>
</tr>
<tr>
<td>Hotels.com</td>
<td>Yes</td>
</tr>
<tr>
<td>Kayak</td>
<td>Yes</td>
</tr>
<tr>
<td>Priceline</td>
<td>Yes</td>
</tr>
<tr>
<td>Travelocity</td>
<td>Yes</td>
</tr>
<tr>
<td>Orbitz</td>
<td>Yes</td>
</tr>
<tr>
<td>ChoiceHotels</td>
<td>Yes</td>
</tr>
<tr>
<td>HolidayCheck</td>
<td>Yes</td>
</tr>
<tr>
<td>ChoiceHotels</td>
<td>Yes</td>
</tr>
<tr>
<td>InterContinental Hotels Group</td>
<td>Yes</td>
</tr>
<tr>
<td>Marriott International</td>
<td>Yes</td>
</tr>
<tr>
<td>Global Hyatt Corp.</td>
<td>No</td>
</tr>
</tbody>
</table>

Adoption: 93 %
Hands-on: How to get the Data?

- N-Quads: http://www.webdatacommons.org/structureddata/
- JSON: http://webdatacommons.org/structureddata/schemaorgtables/

Class-Specific Subsets of the Schema.org Data

<table>
<thead>
<tr>
<th>Class Name</th>
<th>Total Number of</th>
<th>Top Classes (Entity Count)</th>
<th>Total File Size</th>
<th>Quad File</th>
</tr>
</thead>
</table>

- Only tip of the iceberg, as each website is only partly crawled.
3. Alternative Approach: Linked Data

Extend the Web with a single global data graph
- by using RDF to publish structured data on the Web
- by setting links between data items within different data sources
Entities are identified with HTTP URIs

```
http://richard.cyganiak.de/foaf.rdf#cygri
http://dbpedia.org/resource/Berlin
```

HTTP URIs take the role of global primary keys.
URIs can be linked to navigate the data space

By following RDF links applications can

- navigate the global data graph
- discover new data sources
The Linked Open Data Cloud

1,239 datasets connected by 16,147 sets of RDF links (as of March 2019)

https://lod-cloud.net/
Uptake in the Libraries Community

**Institutions publishing Linked Data**
- Library of Congress (subject headings and catalog)
- German National Library (PND dataset and subject headings)
- Swedish National Library (Libris catalog)
- Hungarian National Library (OPAC and digital library)
- Europeana Digital Library (catalog)
- Springer Nature (publications, researchers, projects)

**Goals:**
1. Interconnect resources between repositories (by topic, by location, by historical period, by ...)
2. Integrate library catalogs on global scale
Uptake in the Life Science Domain

■ Goals:

1. Connect life science datasets in order to support
   - biological knowledge discovery
   - drug development
2. Reuse results of previous integration efforts

■ Projects

BIO2RDF

Open PHACTS
2. Web 2.0 Applications and Web APIs

- A multitude of Web-based applications has sprung up which enable users to share information.

- These applications
  - collect large amounts of data using proprietary schemata.
  - form separate data spaces that are only partly accessible from the Web via:
    1. HTML interfaces
    2. Web APIs
Example: Facebook

- **Users (September 2018)**
  - 2.3 billion monthly active users
  - including 1 billion mobile users

- **740 billion friend connections**

- **4 million likes every minute**

- **250 billion photos uploaded**

- **Data Volume**
  - 4 Petabyte of new data generated every day
  - over 300 Petabyte in Facebook's data warehouse

https://www.brandwatch.com/blog/facebook-statistics/
http://www.technologyreview.com/featuredstory/428150/what-facebook-knows/
Web APIs

- allow to access the data programmatically
- example of using the Twitter API (now X API):

```python
import pandas as pd

# Search tweets
dict_ = {'user': [], 'date': [], 'text': [], 'favorite_count': []}
for status in python_tweets.search(**query)['statuses']:
    dict_['user'].append(status['user']['screen_name'])
    dict_['date'].append(status['created_at'])
    dict_['text'].append(status['text'])
    dict_['favorite_count'].append(status['favorite_count'])

# Structure data in a pandas DataFrame for easier manipulation
def = pd.DataFrame(dict_)
def.sort_values(by='favorite_count', inplace=True, ascending=False)
def.head(5)
```
Web APIs

ProgrammableWeb API Catalog

- lists over 24,000 Web APIs
- lists over 6,800 mashups
- catalog maintained until 10-2022
  (alternatives below*)

APIs usually provide only limited access

- restricted to specific queries (canned queries)
- restricted by amount of queries / number of results
- (try to) prevent crawling

Web APIs versus Linked Data or HTML Embedded Data
3. What is Web Mining?

Definition

Non-trivial extraction of implicit, previously unknown and potentially useful information from
- Web content,
- Web structure
- Web usage data.

Recurring Challenges

1. huge amount of available data \(\rightarrow\) requires sampling or cloud computing
2. semi-structured nature of data \(\rightarrow\) mix of data and text mining techniques
3. heterogeneity of data \(\rightarrow\) data integration and cleansing is a challenge
4. distributed nature of data \(\rightarrow\) often requires large-scale crawling or relying on pre-crawled web corpora
Web Mining is a Multi-Disciplinary Field

- Draws ideas and techniques from
  - Machine Learning
  - Database Systems
  - Natural Language Processing
  - Social Network Analysis

- Sub-Fields
  1. Web Usage Mining
  2. Web Structure Mining
  3. Web Content Mining
3.1 Web Usage Mining

- **Definition**

  Discovery of patterns in click-streams and associated data collected or generated as a result of user interactions with one or more web sites or Web 2.0 applications.

- **Typical Sources of Data**
  1. web server access logs
  2. e-commerce and product-oriented user events (e.g., shopping cart changes, ad or product click-throughs, purchases)
  3. user events on social network sites (e.g., likes, posts, comments)

- **Associated Data**
  1. page attributes, page content, site structure
  2. additional domain knowledge and demographic data
  3. user profiles or user ratings
Web Usage Data: An Endless Sea

THE INTERNET IN 2023 EVERY MINUTE

Web Server Logs

Get Traffic Analysis

Provide Access to

Google Analytics

Universität Mannheim – Bizer/Ponzetto/Korini/Brinkmann: Web Mining – FSS2024 (Version: 13.2.2024) – Slide 47
The Web Usage Mining Process

Data Preparation Phase

- Web & Application Server Logs
- Data Preprocessing
  - Data Cleaning
  - Pageview Identification
  - Sessionization
  - Data Integration
  - Data Transformation

Pattern Discovery Phase

- Aggregate User models
- Pattern Analysis
  - Pattern Filtering
  - Aggregation
  - Characterization

Usage Mining

- Transaction Clustering
- Pageview Clustering
- Correlation Analysis
- Association Rule Mining
- Sequential Pattern Pattern Mining
Example Application: Product Recommendation

- Boundless: Upgrade Your Brain, Optimize Your Body & Defy Aging
- Leadership Strategy and Tactics: Field Manual
- Running Rewired: Reinvent Your Run for Stability, Strength, and... 
- Training for the Uphill Athlete: A Manual for Mountain Runners and...
- Born to Run: The Hidden Tribe, the Ultra-Runners, and the Greatest...
Example Application: Personalized Search

Google search for "kafka"

Ad: www.confluent.io/ - Confluent | Download Apache Kafka® Today | confluent.io

What is Kafka?

Kafka Definitive Guide
Learn All About Kafka From its Original Developers in this eBook.

Apache Kafka
Kafka® is used for building real-time data pipelines and streaming apps. It is horizontally scalable, fault-tolerant, wicked fast, and runs in production in thousands...

Introduction - Apache Kafka - Apache Software
In Kafka the communication between the clients and the servers is done with a simple, high-performance, language agnostic TCP protocol. This protocol is...

People also ask
What is Kafka used for?
What is meant by Kafka?
What is Kafka and how it works?
Is Kafka free?

Apache Kafka
Apache Kafka is an open-source stream-processing software platform developed by LinkedIn and donated to the Apache Software Foundation, written in Scala and Java. The project aims to provide a unified, high-throughput, low-latency platform for handling real-time data feeds. Wikipedia

License: Apache License 2.0
Developer(s): Apache Software Foundation
Initial release date: January 2011
Stable release: 2.4.0 / December 16, 2019; 48 days ago
Written in: Scala, Java

People also search for
View 15+ more
Apache
Spark
RabbitMQ
Apache ZooKeeper
Apache Cassandra
Apache Hadoop
Example Application: Personalized Search

Google search for "kafka"
3.2 Web Structure Mining

- **Definition**
  
  Discovery and interpretation of patterns in
  1. the hyperlink structure of the Web
  2. the social ties among actors that interact on the Web

- **Typical sources of Web graphs**
  1. Web crawls including HTML pages and hyperlinks
  2. social networks including explicit relations between actors (your Facebook friend network)
  3. other types of community data (discussion forums, email conversations, …)

- **Focuses on the structure, but can of course also be combined with content or usage mining techniques**
Identification of Prominent Nodes

Question: Who are the “most important” actors in a social network?

Centrality
- A central actor is one involved in many edges.
- The direction of lines is not considered.

Prestige
- A prestigious actor is one who is the target of many arcs.
- The direction of arcs is considered.
A **community** is a set of actors between which interactions are (relatively) frequent.

- Finding a community in a social network is to identify a set of nodes such that they interact with each other more frequently than with those nodes outside the group.

- **Methods**: Components, K-Cores, Islands, …

- **Applications**: Recommendation based on communities, visualization of huge networks
Machine Learning with Graphs

**Link Prediction:** Given a snapshot of a social network, can we infer which new interactions among its members are likely to occur in the near future? (Liben-Nowell & Kleinberg, 2007)

- Facebook: recommending possible friends
- Tinder: recommending potential matches

**Node Classification:** Predict node properties based on other node properties as well as neighboring nodes.

- Knowledge Graph: Type of a node: e.g., person, author, athlete
3.3 Web Content Mining

- **Definition**

  Automatic extraction of useful information (facts, patterns) from Web content (text, images, multimedia).

- **Content Mining Tasks**
  - Content Clustering
  - Content Classification
  - Sentiment Analysis
  - Political Scaling
  - Information Extraction
Goal: Previously unseen documents/images should be assigned a class as accurately as possible.

Applications
- News categorization
- Product categorization
- Spam detection

Classification methods
- Naive Bayes, Support Vector Machines, Deep Neural Nets, Transformers
Unsupervised Learning: Given a set of documents and a similarity measure among documents find clusters such that:
- documents in one cluster are more similar to one another
- documents in separate clusters are less similar to one another

Applications
- Search result clustering
- Topic discovery

Techniques
- Algorithms: K-Means, Hierarchical Clustering, S-BERT
- Similarity measures: Cosine, Jaccard, Text Embedding
Mixture of Document Clustering and Classification
Sentiment Analysis

The basic task in sentiment analysis is classifying the polarity of a given text at the document, sentence, or feature/aspect level.

- **Polarity Values**
  - positive, neutral, negative
  - stars

- **Applications**
  - Document-level: vote prediction from tweets
  - Feature/Aspect-level: analysis of product reviews
Hate Speech Detection / Political Scaling

- Hate Speech Detection in Social Media Content

<table>
<thead>
<tr>
<th></th>
<th>Twitter</th>
<th>% posts</th>
<th>Whisper</th>
<th>% posts</th>
</tr>
</thead>
<tbody>
<tr>
<td>I hate</td>
<td>70.5</td>
<td>I hate</td>
<td>66.4</td>
<td></td>
</tr>
<tr>
<td>I can’t stand</td>
<td>7.7</td>
<td>I don’t like</td>
<td>9.1</td>
<td></td>
</tr>
<tr>
<td>I don’t like</td>
<td>7.2</td>
<td>I can’t stand</td>
<td>7.4</td>
<td></td>
</tr>
<tr>
<td>I really hate</td>
<td>4.9</td>
<td>I really hate</td>
<td>3.1</td>
<td></td>
</tr>
<tr>
<td>I fucking hate</td>
<td>1.8</td>
<td>I fucking hate</td>
<td>3.0</td>
<td></td>
</tr>
<tr>
<td>I’m sick of</td>
<td>0.8</td>
<td>I’m sick of</td>
<td>1.4</td>
<td></td>
</tr>
<tr>
<td>I cannot stand</td>
<td>0.7</td>
<td>I’m so sick of</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>I fuckin hate</td>
<td>0.6</td>
<td>I just hate</td>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td>I just hate</td>
<td>0.6</td>
<td>I just don’t like</td>
<td>0.8</td>
<td></td>
</tr>
<tr>
<td>I’m so sick of</td>
<td>0.6</td>
<td>I secretly hate</td>
<td>0.7</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Twitter</th>
<th>% posts</th>
<th>Whisper</th>
<th>% posts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nigga</td>
<td>31.11</td>
<td>Black people</td>
<td>10.10</td>
<td></td>
</tr>
<tr>
<td>White people</td>
<td>9.76</td>
<td>Fake people</td>
<td>9.77</td>
<td></td>
</tr>
<tr>
<td>Fake people</td>
<td>5.07</td>
<td>Fat people</td>
<td>8.46</td>
<td></td>
</tr>
<tr>
<td>Black people</td>
<td>4.91</td>
<td>Stupid people</td>
<td>7.84</td>
<td></td>
</tr>
<tr>
<td>Stupid people</td>
<td>2.62</td>
<td>Gay people</td>
<td>7.06</td>
<td></td>
</tr>
<tr>
<td>Rude people</td>
<td>2.60</td>
<td>White people</td>
<td>5.62</td>
<td></td>
</tr>
<tr>
<td>Negative people</td>
<td>2.53</td>
<td>Racist people</td>
<td>3.35</td>
<td></td>
</tr>
<tr>
<td>Ignorant people</td>
<td>2.13</td>
<td>Ignorant people</td>
<td>3.10</td>
<td></td>
</tr>
<tr>
<td>Nigger</td>
<td>1.84</td>
<td>Rude people</td>
<td>2.45</td>
<td></td>
</tr>
<tr>
<td>Ungrateful people</td>
<td>1.80</td>
<td>Old people</td>
<td>2.18</td>
<td></td>
</tr>
</tbody>
</table>

- Political Scaling of Opinionated Texts

![Political Scaling Diagram](image)
Information Extraction

Goal: Automatic extraction of structured information from unstructured or semi-structured Web content.

- Example of below 1NF data:

- The difficulty of the extraction depends on the structuredness

Diagram:
- Structuredness vs. Difficulty of information extraction:
  - Web APIs
  - HTML-embedded Data
  - HTML Tables
  - DOM Trees
  - Free text
  - Parsers
  - LLMs
Examples: Information Extraction from the Web

Equal to the standard data mining process with the difference that data is gathered from the Web.
Gathering and Exploration

1. Gathering of Web Data
   - Crawl documents or data
   - Retrieve data via Web API
   - Download pre-gathered data sets

2. Exploration
   - Get an initial understanding of the data
   - Calculate basic summarization statistics
   - Visualize the data
   - Identify data problems such as outliers, missing values, duplicate records
Preprocessing and Transformation

- Transform data into a representation that is suitable for the chosen data mining methods
  - amount of data (determines hardware requirements)
  - number of dimensions (represent relevant information using less attributes)
  - scales of attributes (nominal, ordinal, numeric)

- Methods
  - discretization and binarization
  - feature subset selection / dimensionality reduction
  - attribute transformation / text to term vector / embeddings
  - aggregation, sampling
  - integrate data from multiple sources

- Good data preparation is key to producing valid and reliable models
- Data integration and preparation is estimated to take 70-80% of the time and effort of a data mining project
Actual Data Mining

- **Input:** Preprocessed Data
- **Output:** Model / Patterns

1. **Apply data mining method**
2. **Evaluate resulting model / patterns**
3. **Iterate**
   - experiment with different hyperparameter settings
   - experiment with multiple alternative methods
   - improve preprocessing and feature generation
   - increase amount or quality of the training data
Deployment

- Use model in the business context
- Keep iterating to maintain and improve model

CRISP-DM Process Model
Questions?

- This week: No lab!
- Next week: Lecture and Lab: Web Usage Mining and Recommender Systems