
# Data Preparation and Analysis I

**Dr. Anne Kayem**

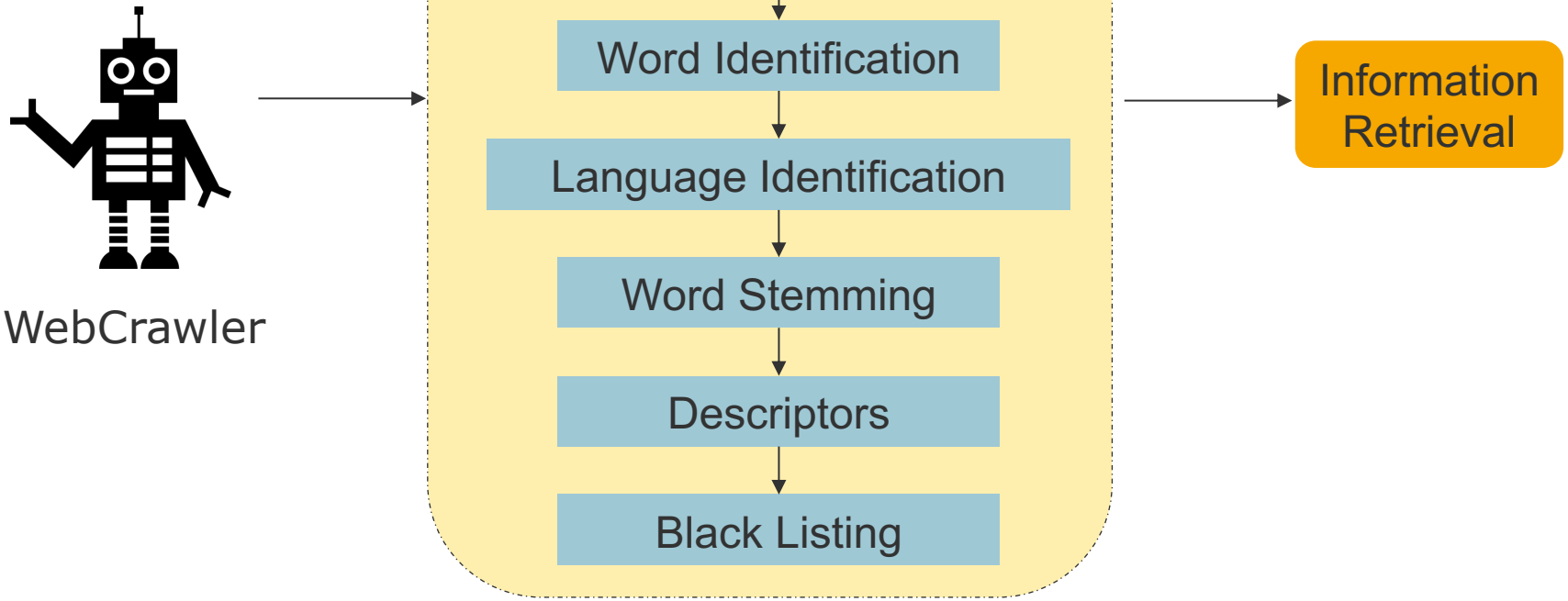
Hasso-Plattner-Institute
University of Potsdam

# Data Preparation and Analysis

- Content indexing of text documents:
  - conversion of documents to uniform document type (HTML, PostScript, PDF, DOCX, PPTX to text)

- Efficiently searchable database – to enable finding relevant strings by semantic analysis of the text file: Search and analysis of:
  - keywords
  - headings
  - Bullet points
  - ...
- Assignment of keywords (descriptors) to documents
- Form ranking order taking into account evaluation criteria

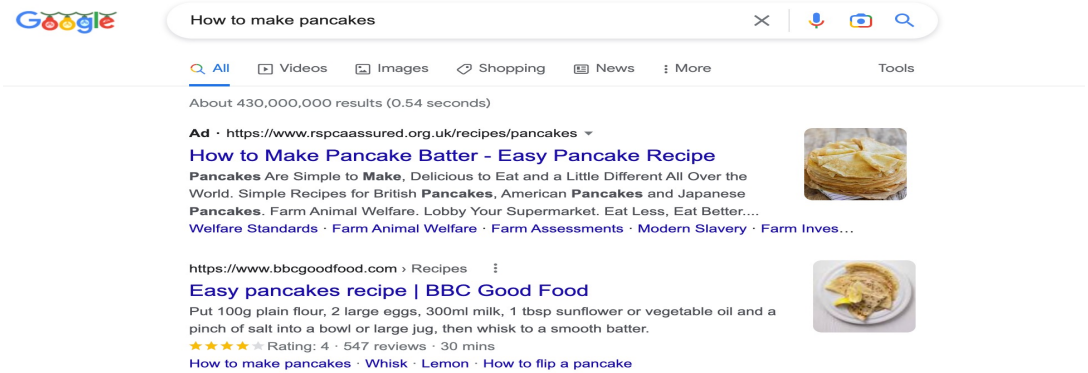
# Text Preparation Phase

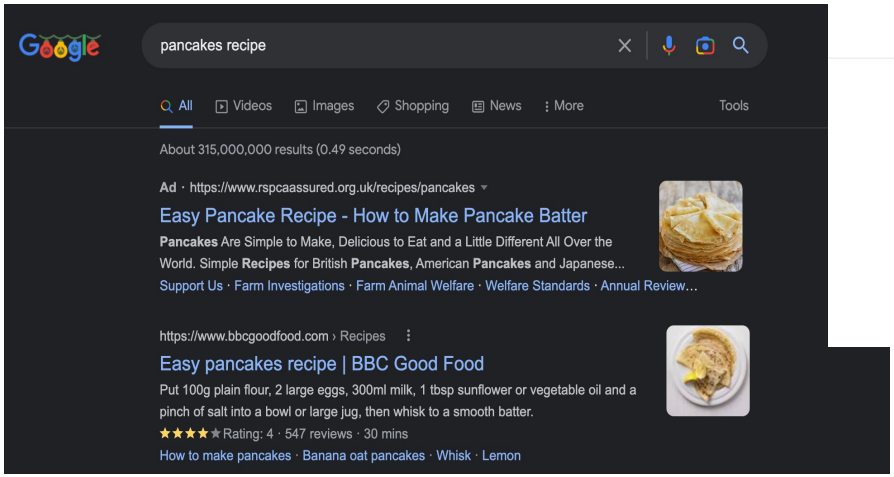


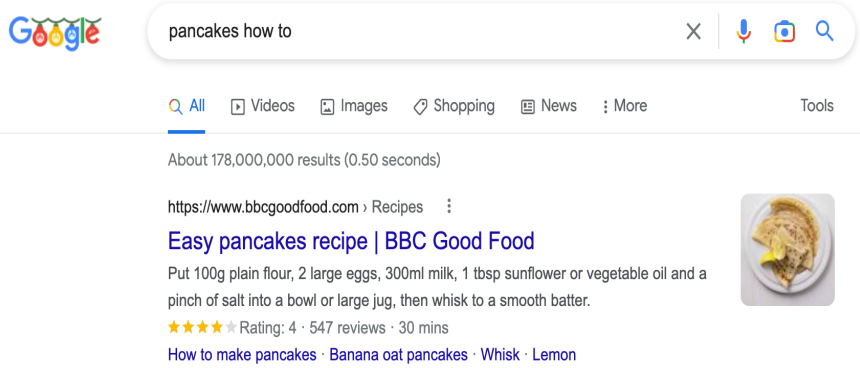
# Descriptor Extraction

- **Keywords**
  - Documents should be findable with relatively few keywords and be representative of the content sought be users.
  - Application of keyword relevance filters for partial context analysis, e.g.
  - HTML **<h$_x$>**-Tags, Text Highlighting, etc…
    - Omitting filler words, linking words, pronouns, etc....
  - Frequency analysis of keywords → **Relevance**

# Descriptor Extraction - Example



# George Kingsley Zipf (1902 – 1950)

**Zipf's Law:** It is always easier for the author of a text to repeat certain words describing a subject than to constantly search for new terms.

# Vector Space Model

- Proven method to support information retrieval

- **Idea:**
  - Document is considered as a vector in an n-dimensional vector space (n - number of descriptors).
  - basic vectors represent one word each
  - document vector is a linear combination of the basis vectors, where each basis vector is multiplied by the number of occurrences of the word
  - Document analysis can then be performed using methods from linear algebra

# Example

Apple is a fruit and not a vegetable.

A Potato is neither a vegetable nor a fruit. Onions are a vegetable.

| Word | Vector | #d1 | #d2 |
|---|---|---|---|
| Apple | e1=(1,0,0,0,0,0) | 1 | |
| Fruit | e2=(0,1,0,0,0,0) | 1 | 1 |
| Potato | e3=(0,0,1,0,0,0) | | 1 |
| Onions | e4=(0,0,0,1,0,0) | | 1 |
| Vegetable | e5=(0,0,0,0,1,0) | 1 | **2** |

$$d1 = e1 + e2 + e5 = (1,1,0,0,1)$$

$$d2 = e2 + e3 + e4 + \mathbf{2} * \boldsymbol{e5} = (0,1,1,1,\mathbf{2})$$

…