



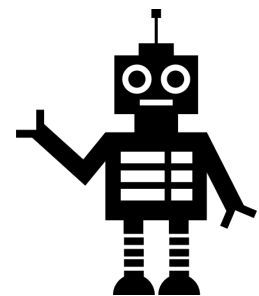
Data Preparation and Analysis I

Dr. Anne Kayem

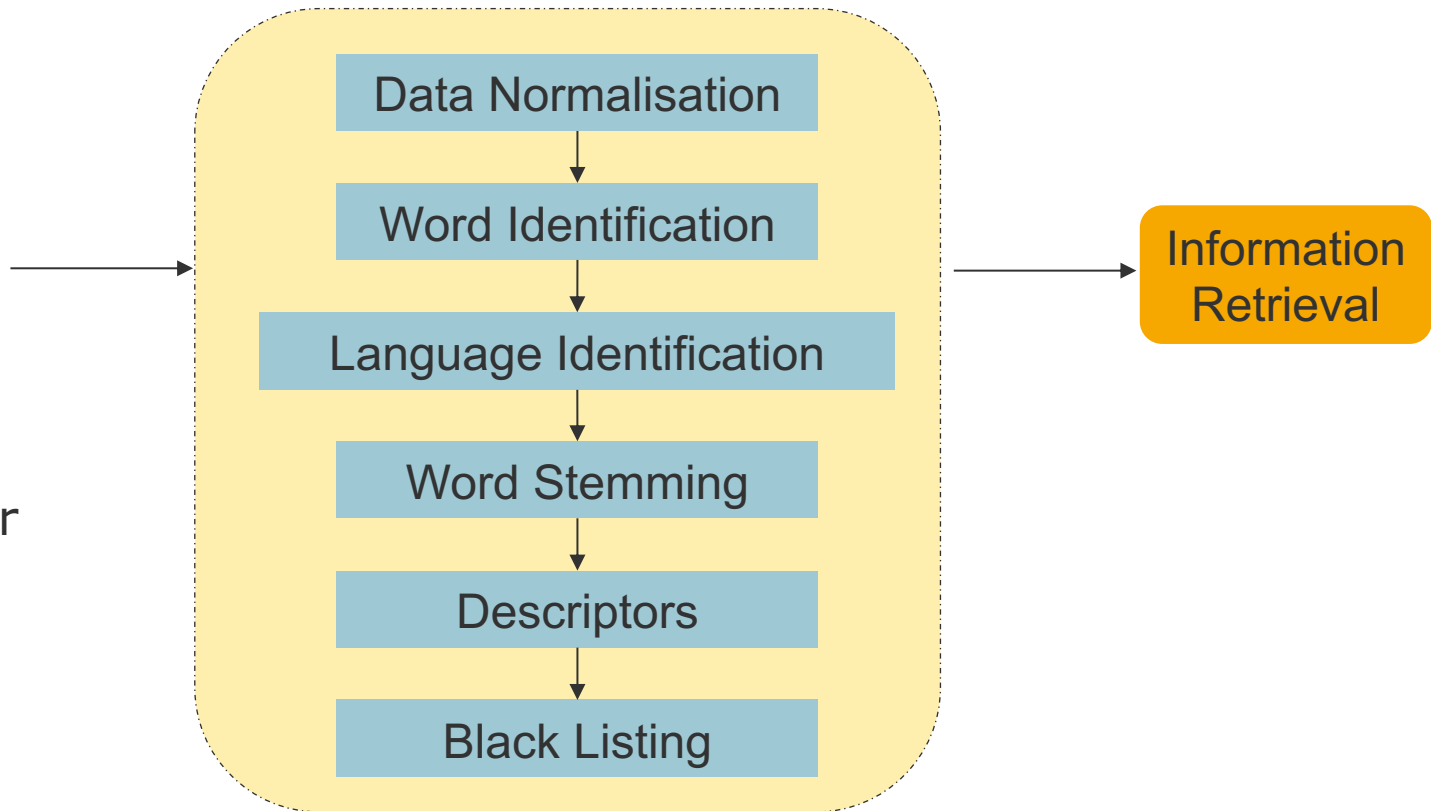
Hasso-Plattner-Institute
University of Potsdam

- 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



WebCrawler



■ Keywords

- Documents should be findable with relatively few keywords and be representative of the content sought by 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



How to make pancakes



[All](#) [Videos](#) [Images](#) [Shopping](#) [News](#) [More](#) [Tools](#)

About 430,000,000 results (0.54 seconds)

Ad · <https://www.rspcaassured.org.uk/recipes/pancakes>

How to Make Pancake Batter - Easy Pancake Recipe

Pancakes Are Simple to Make, Delicious to Eat and a Little Different All Over the World. Simple Recipes for British Pancakes, American Pancakes and Japanese Pancakes. Farm Animal Welfare. Lobby Your Supermarket. Eat Less, Eat Better...

[Welfare Standards](#) · [Farm Animal Welfare](#) · [Farm Assessments](#) · [Modern Slavery](#) · [Farm Inves...](#)



<https://www.bbcgoodfood.com> · Recipes

Easy pancakes recipe | BBC Good Food

Put 100g plain flour, 2 large eggs, 300ml milk, 1 tbsp sunflower or vegetable oil and a pinch of salt into a bowl or large jug, then whisk to a smooth batter.

★★★★★ Rating: 4 · 547 reviews · 30 mins

[How to make pancakes](#) · [Whisk](#) · [Lemon](#) · [How to flip a pancake](#)



pancakes how to



[All](#) [Videos](#) [Images](#) [Shopping](#) [News](#) [More](#) [Tools](#)

About 178,000,000 results (0.50 seconds)

<https://www.bbcgoodfood.com> · Recipes

Easy pancakes recipe | BBC Good Food

Put 100g plain flour, 2 large eggs, 300ml milk, 1 tbsp sunflower or vegetable oil and a pinch of salt into a bowl or large jug, then whisk to a smooth batter.

★★★★★ Rating: 4 · 547 reviews · 30 mins

[How to make pancakes](#) · [Banana oat pancakes](#) · [Whisk](#) · [Lemon](#)



pancakes recipe



[All](#) [Videos](#) [Images](#) [Shopping](#) [News](#) [More](#) [Tools](#)

About 315,000,000 results (0.49 seconds)

Ad · <https://www.rspcaassured.org.uk/recipes/pancakes>

Easy Pancake Recipe - How to Make Pancake Batter

Pancakes Are Simple to Make, Delicious to Eat and a Little Different All Over the World. Simple Recipes for British Pancakes, American Pancakes and Japanese...

[Support Us](#) · [Farm Investigations](#) · [Farm Animal Welfare](#) · [Welfare Standards](#) · [Annual Review...](#)



<https://www.bbcgoodfood.com> · Recipes

Easy pancakes recipe | BBC Good Food

Put 100g plain flour, 2 large eggs, 300ml milk, 1 tbsp sunflower or vegetable oil and a pinch of salt into a bowl or large jug, then whisk to a smooth batter.

★★★★★ Rating: 4 · 547 reviews · 30 mins

[How to make pancakes](#) · [Banana oat pancakes](#) · [Whisk](#) · [Lemon](#)



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.

- 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

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

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

...