

# ML2 Projekt: Image-Text Matching

Menschen Lernen Maschinelles Lernen @ HS Offenburg

# Image Text Matching

**Implementierung einer prototypischen Applikation, welche eine automatische Verknüpfung von Bild- und Textmaterial durchführt**

- Dient als PoC für spätere Analysen von digitalen Asservaten
- Auffinden von Beziehungen zwischen Daten  
(z.B. Verknüpfung von E-Mails mit Bildern bei forensischen Ermittlungen)

## **Meilensteine**

- A: Finden geeigneter Trainings-Datensätze zum Lernen der Modelle für Textsegmentierung und Objekterkennung
  - Alternativ: Evaluation bestehender Modelle
- B: Modelle parallel auf den Daten anwenden
  - Textsegmentierung und Objekterkennung
- C: Zusammenführen der Ergebnisse zum Auffinden von Text-Bild Verknüpfungen
- [D: Aufbau und Verwendung einer Taxonomie zur erweiterten Begriffsfindung (z.B. aus Wiktionary)]

# Analyse-Schritte

1. Modelle finden bzw. lernen
2. Durchsuche Startordner und dessen Unterverzeichnisse nach **Bilddateien**
3. Klassifiziere jede Datei mithilfe des (festgelegten bzw. gelernten) Modells
4. Schreibe Dateipfad und Ergebnis in eine Tabelle
5. Durchsuche Startordner und Unterordner nach **Textdateien** (später weitere Datenformate)
6. Extrahiere Informationen (Stemming) aus diesen Dateien
7. Klassifiziere diese Textteile mithilfe des (festgelegten bzw. gelernten) Modells
8. Schreibe Dateipfad und Ergebnis in eine Tabelle
9. Gleiche Tabellen auf Treffer ab und zeige Beziehungen auf (Text → gefundene Bilder)

parallel

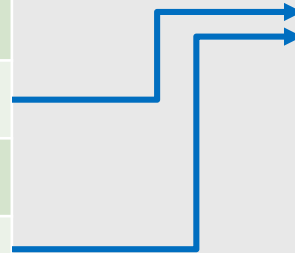
# Aufgaben

## Text Segmentation

- Finde und zerlege Textdateien in wichtige Wörter (Filterung unwichtiger Inhalte)
- Stemming / Lemmatisierung
- Verwende bestehendes Modell zur Verschlagwortung
- Lernen eines neuen Modells?
- Trainingsdatensatz: ?



ID	path	text_word	class	conf
1	/email1.txt	cabrio	car	0.9
2	/email1.txt	bulldog	dog	0.7
3	/email1.txt	bulldog	animal	0.9
4	/email2.txt	pitbull	dog	0.8



## Image Search

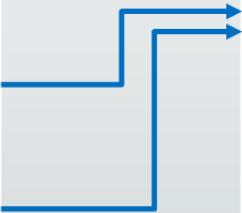
- Klassifizierung von Bildern
- Verwende bestehendes Modell zur Klassifikation
- Lernen eines neuen Modells?
- Trainingsdatensätze: ImageNet, CIFAR-100, Coco, ...



ID	path	class	conf
1	/bulldog.jpg	dog	0.9
2	/cabrio.jpg	car	0.8
3	/IMGo1.jpg	food	0.7
4	/blume.jpg	flower	0.9

# Datenabgleich

- Speichere die Tabellen in einer **Hadoop** Datenbank
  - Hive, HBase (Phoenix)
  - Skalierungsmöglichkeiten im Hadoop ausnutzen können
- Abgleich auf Treffer durch Tabellenabfragen
  - Wenn ein Begriff selbst (oder dessen Klasse) auf die Klasse eines Objekts (Foto) passt → Match



ID	path	text_word	class	conf
1	/email1.txt	cabrio	car	0.9
2	/email1.txt	bulldog	dog	0.7
3	/email1.txt	bulldog	animal	0.9
4	/email2.txt	pitbull	dog	0.8

ID	path	class	conf
1	/bulldog.jpg	dog	0.9
2	/cabrio.jpg	car	0.8
3	/IMGo1.jpg	food	0.7
4	/blume.jpg	flower	0.9

# Projektorganisation

- Docker Image mit TensorFlow Umgebung
- Jupyter Notebook für Coding
- Git Repository, um Ergebnisse zu sichern
  
- Hilfreiche Quellen:
  - <https://github.com/tensorflow/models>
  - Python Natural Language Toolkit (NLTK)
    - <https://pythonprogramming.net/tokenizing-words-sentences-nltk-tutorial/>
    - <https://www.datacamp.com/community/tutorials/text-analytics-beginners-nltk>

# Begrifflichkeiten

- Text Segmentation: Zerteilen von Text in bedeutungsvolle Teile wie Wörter, Sätze oder Themen
- Stemming / Lemmatisierung: Zurückführen eines Wortes auf dessen Wortstamm
  - Stemming: Anwenden von Regeln zur Wortveränderung (Präfix, Suffix, etc.)
  - Lemmatisierung: Verwendung eines Wörterbuchs zur Zurückführung des Wortes (Python NLTK)

# Notizen


- Textzerlegung
  1. Sätze splitten
  2. Wörter innerhalb der Sätze splitten
  3. POS (Part-of-Speech) Tagging: Wortarten (Verb, Adjektiv, Nomen, etc.) klassifizieren
  4. Nomen filtern & auf Grundform zurückbringen
  5. **Wörter klassifizieren**
    - WordNet (<https://www.web3.lu/wordnet-imagenet/>), Word2Vec
- Objekterkennung
  - ImageNet, COCO, Cifar-100, ...



## Text Segmentation

- Finde und zerlege Textdateien in wichtige Wörter (Filterung unwichtiger Inhalte)
- Stemming / Lemmatisierung
- Verwende bestehendes Modell zur Verschlagwortung
- Lernen eines neuen Modells?


- Trainingsdatensatz:



ID	path	text_word	class	conf
1	/email1.txt	cabrio	car	0.9
2	/email1.txt	bulldog	dog	0.7
3	/email1.txt	bulldog	animal	0.9
4	/email2.txt	pitbull	dog	0.8

## Image Search

- Klassifizierung von Bildern
- Verwende bestehendes Modell zur Klassifikation
- Lernen eines neuen Modells?
- Trainingsdatensätze:  
ImageNet, CIFAR-100, COCO, ...



ID	path	class	conf
1	/bulldog.jpg	dog	0.9
2	/cabrio.jpg	car	0.8
3	/IMGo1.jpg	food	0.7
4	/blume.jpg	flower	0.9

