SwissText 2017

2nd Swiss Text Analytics Conference

Poster Presentations

textshuttle.ai

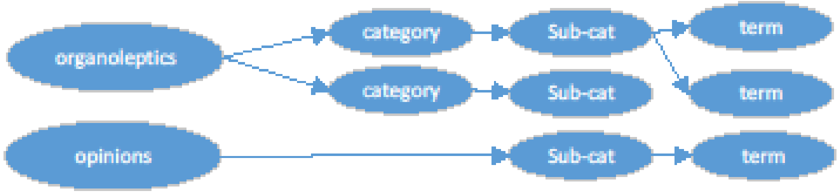
Machine Translation for Film and TV Subtitles

- Subtitles as an excellent basis for machine translation
- Applied language technology:
  Neural **M**achine **T**ranslation
- Translation industry: post-editing savings up to 30%

www.textshuttle.ai
welcome@textshuttle.ai
+41 44 500 73 45

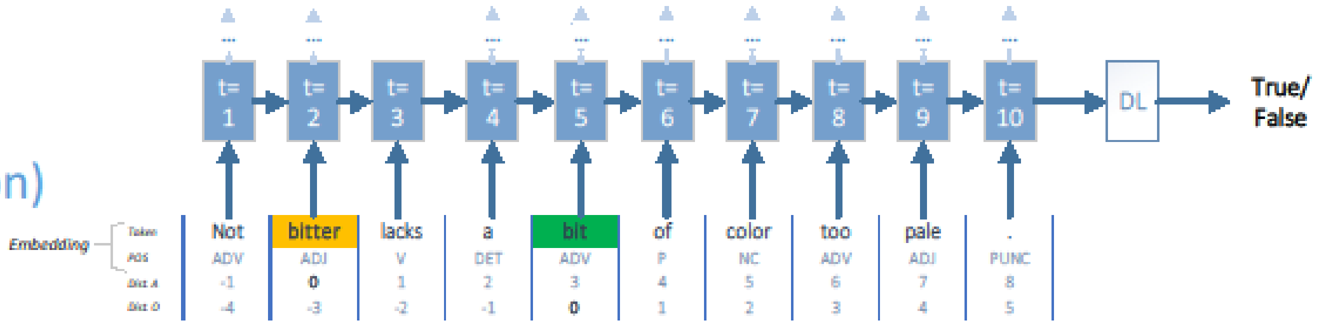
Aspect-based sentiment analysis to extract organoleptic wine profile

Ontology
(aspects and opinions extraction)



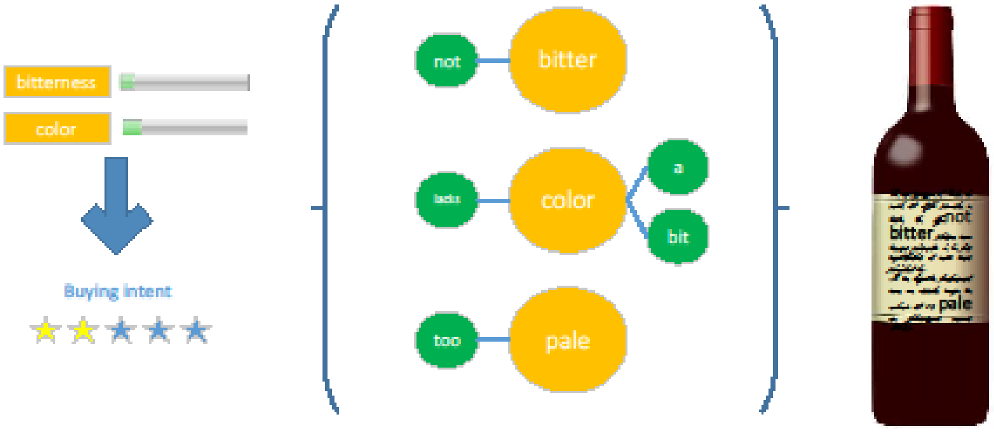
+

Stacked RNN
(relation extraction)



=

Wine profile
&
Buying intent



SWA SCHWEIZERISCHES WIRTSCHAFTSARCHIV

SPINNINGBYTES

Zurich University of Applied Sciences
zh aw School of Engineering

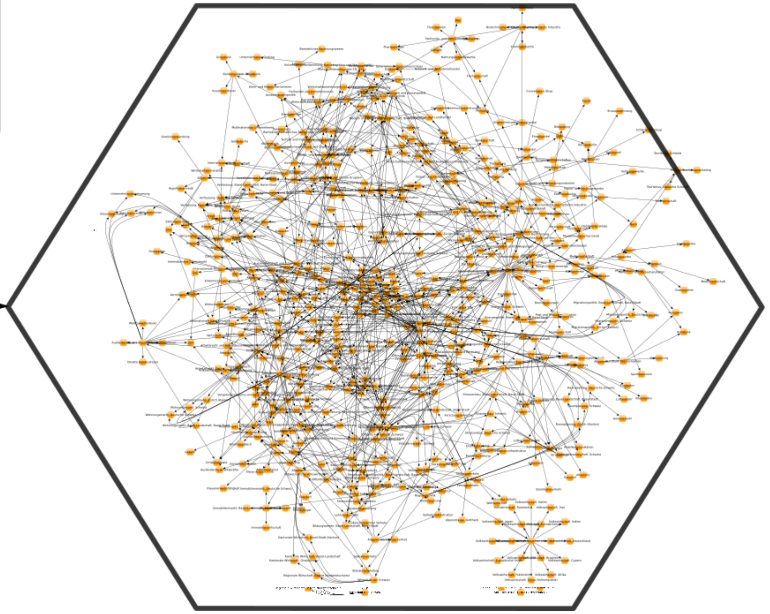
University of Basel

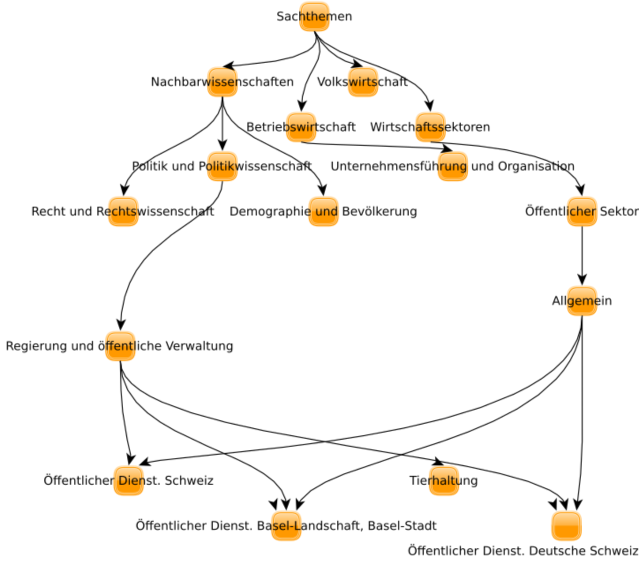
Hierarchical Classification for Economic News Articles

Fernando Benites Mark Cieliebak

News Article

Twint erobert die Region
Basler Kantonalbank und
Bank Coop bieten
Bezahl-App an





Weighted word overlap and word embeddings: A practical ensemble approach to Question Matching in a Dialogue Simulator

Don Tuggener, Institute of Computational Linguistics UZH / SUPSI / LifeLike

- Dialogue simulator for medical interviews (*LifeLike*)
- Replace mouse interaction with voice input
- Match user input to available questions using machine learning
- Handle non-matching inputs, speech-to-text errors

- Medical classifications include between 10 an 70 thousands single codes (or even more in case of SNOMED)
- Stochastic systems have limitations to differentiate a big number of possible outcomes due to exponential growth of the training corpus needed
- Semantic systems are advantageous since analyzing the inputs from inside and applying domain knowledge
- The Concept Molecules facilitate development and maintenance of our interpreting system
- The interpretation result, which contains implicit and explicit information from the input, allows for versatile data handling and analysis

We have answered this before: A study of the characteristics and solutions of the question retrieval and equivalence detection problem

Alireza Ghasemi, Silvia Quarteroni
ELCA Informatique SA

Vidatics GmbH - Dr. Christian Rohrdantz

Improving Product Development and Customer Journey through Text Analytics

Thousands of Reviews

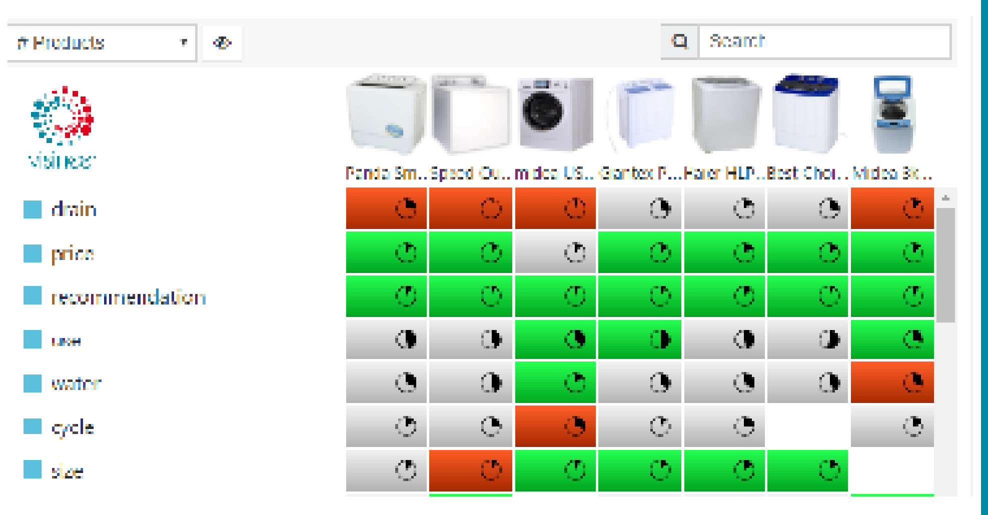
Our Text Analytics Engine

VIDATICS

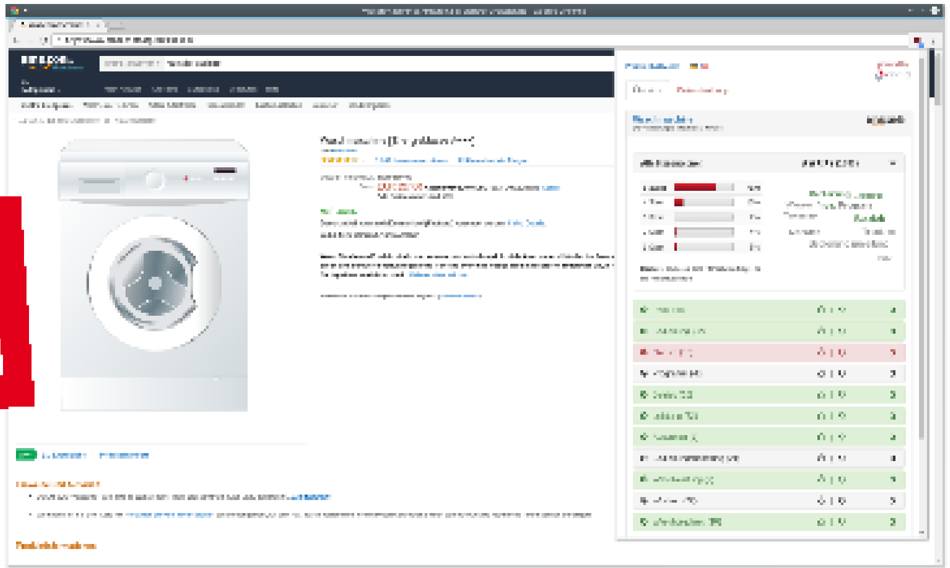
Domain Models | Linguistic Rules

Artificial Intelligence / Machine Learning

B2B Solution: Visineer Dashboard



B2C Solution: Product Buddy Browser App



Free download for YOU!
www.product-buddy.com

Firefox

available in the chrome web store

Making chatbots for the Swiss industry: Lessons learned

Lancelot Caron, Silvia Quarteroni
ELCA Informatique SA

University of Zurich[UZH]

Institute of Computational Linguistics

Website
http://pub.cl.uzh.ch/demo/diachronvis/bbdia/

BBDia: Diachronic Visualization of Semantically Related N-grams Using Word Embeddings

Similar Words

Similar Spellings

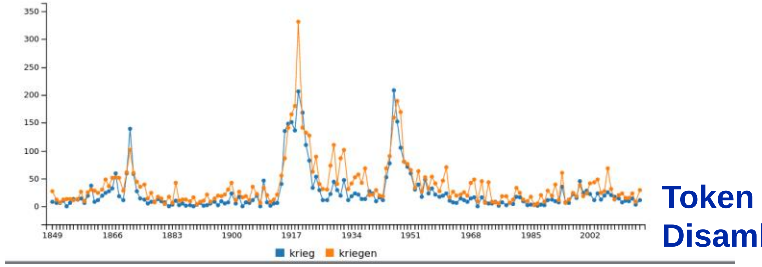
Temporal Trends

Suche

Krieg

ähnliche Bedeutung:
Weltkrieg
Feindseligkeit
Bürgerkrieg
Belagerung
Kriegsausbruch
Kriegsereignis
Wirtschaftskrise
Finanzkrise
Revolution
Golfkrieg

ähnliche Schreibweise:
Krise



Token Disambiguation

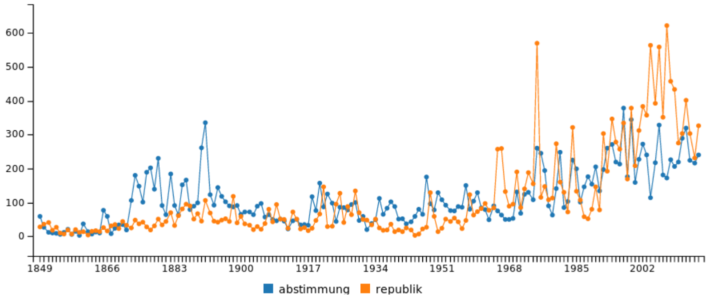
Suche

Abstimmung, Republik

ähnliche Bedeutung:
Volksrepublik
Dominikanische
Demokratischen
Kongo-Brazzaville
Ministerpräsident
Ministerkabinett
Freistaat
Mauretanien

ähnliche Schreibweise:
Eepublik
Bepublik
Kepublik

Finding OCR-Errors

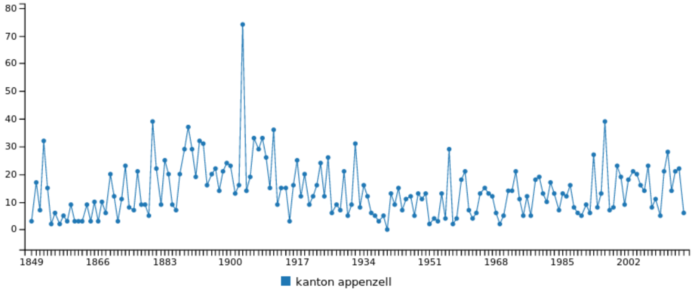


Bigram Search

Suche

Kanton Appenzell

ähnliche Bedeutung:
kanton uri
kanton basel-landschaft
kanton schwyz
kanton thurgau
kanton basel-stadt
kanton obwalden
kanton appenzell
kanton baselland
kanton glarus
kanton baselstadt



A Twitter Corpus and Benchmark Resources for
German Sentiment Analysis

Zurich University of Applied Sciences (ZHAW) - Winterthur, Switzerland
SpinningBytes AG, Küsnacht, Switzerland

Mark Cieliebak **Jan Deriu** **Dominik Egger** **Fatih Uzdilli**

New Corpus: SB-10k

- 9738 German tweets
- Labels: "positive", "neutral", "negative" and "mixed"
- Each tweet annotated by 3 annotators
- Designed to cover a wide variety of unigrams and topics

Positive	Negative	Neutral	Mixed	Unknown	Total
1682	1077	5266	330	1428	9738

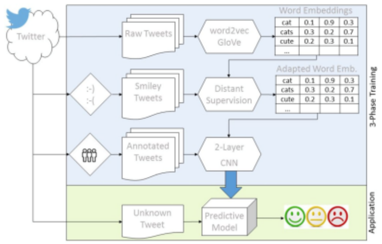
Benchmark for Sentiment Analysis in German

SVM System

Features

- n-grams, n = 1...4, POS-n-grams, n = 3...5, non-contiguous n-grams, n = 3...5
- Character n-grams, n = 3...6
- # upper-cased tokens, # of hashtags, # of POS tags
- # continuous punctuations (max), last token punctuation (?, !)
- # elongated words, # negated tokens
- Lexicons: NRC-emotion, BingLiu, MQA, NRC-HashtagSentiment, Sentiment140, Sentiment140-3-class, RottenTomatoes-3-class)

CNN System



Classifier	Training Corpus	Test Corpus	F1
SVM	SB10k	SB10k	56.98
CNN	SB10k	SB10k	**65.09**
SVM	SB10k	MGS	44.06
CNN	SB10k	MGS	**47.30**
SVM	SB10k	DAI (full)	61.85
CNN	SB10k	DAI (full)	60.61
SVM	MGS	SB10k	60.50
CNN	MGS	SB10k	**61.07**
SVM	MGS	MGS	58.41
CNN	MGS	MGS	**59.80**
SVM	MGS	DAI (full)	57.68
CNN	MGS	DAI (full)	**58.38**

Results

- CNN outperforms SVM in all but one case (red)
- SB-10k generalizes better than MGS to unseen data
- Resulting F1-Scores match state-of-the-art

Corpus and source code are publicly available at www.spinningbytes.com/resources

Johannes Forster (M.Sc.)
jforster@hm.edu

Nora Hollenstein (M.Sc.)
n.hollenstein@de.ibm.com

Hochschule für angewandte Wissenschaften München
Universität Regensburg

Understanding

Natural Language Processing
Image Recognition

Designing Cognitive Computing Architectures for Domain-Specific Decision Support Systems

Reasoning
Provide user-specific recommendations

Learning
Machine Learning to generate or verify hypotheses

Taxonomy Induction using Hypernym Subsequences

Amit Gupta, Remi Lebret, Hamza Harkous, Karl Aberer
EPFL

An end-to-end pipeline for detecting and categorising customer complaints

Thomas Bögel, Henrik Matzen
IBM

Privacy in the Time of Bots: Answering Free-form Questions about Privacy Policies with Deep Learning

Hamza Harkous, Kassem Fawaz, Rémi Lebret,
Florian Schaub, Kang G. Shin, Karl Aberer
EPFL, University of Michigan

KuBu - a chatbot for public transport

- Engage users
- Unique conversations
- Parse input via wit.ai

→ beta available: kubu.chat

Four Different Ways to Build a Chatbot About Movies

SPINNINGBYTES

Erland Xhoxhaj
Yusuf Koc
Sandro Panighetti
Matteo Togni

Dirk Von Grünigen
Martin Weilenmann
Hans Daniel Graf
Daniel Zürrer

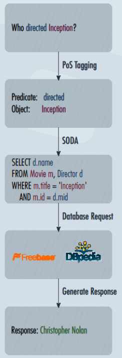
Fernando Benites
Jan Deriu
Nico Neureiter
Pius von Däniken

Mark Cieliebak
Walter Eich
Stephan Neuhaus
Kurt Stockinger

Zürcher Hochschule für Angewandte Wissenschaften
zhaw

Rule Based Question Answering

We developed a bot which is able to answer basic factoid questions on movies, like "Who directed Inception?" or "When was Brad Pitt born?". It parses the question using natural language processing and machine learning techniques and extracts the relevant information from an existing database. This project shows a possible way to replace formal database queries by natural language questions. In that way we could provide an interface for non-technical users to databases from virtually any domain.



System Description

- Part-of-Speech Tagging (Stanford CoreNLP):
  We first extract the subject, object and predicate from the input sentence. We also use synonyms for these words.
- Bag-of-Words similarity (Gensim):
  We compare the input sentence to a set of known questions. If we find a match, we can directly infer the question type.
- Simple Object Data Access (SODA):
  SODA allows to search an entire database for keywords. They are matched to the database's metadata and transformed into an SQL query. We pass the previously found keywords (e.g. subject and predicate) to SODA and use the extracted data to generate our response.
- Database:
  We use the well known knowledge bases Freebase and DBpedia as our information sources.

Results

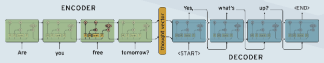
This chatbot demonstrates a way to implement a natural language interface to structured information. To the right you can see one example from the movie domain (flow chart) and three examples on geographical information.

Where is London?	United Kingdom
What ist the country of Dallas?	United States
In which country is Salzburg	Austria

Example questions (left) and the answers of our chatbot (right).

Learning Dialogues End-to-End

We developed a neural network which learns how to respond to a dialogue partner. In contrast to the two chatbots above, the focus was not on correctly answering factoid questions, but rather on learning end-to-end to interpret input and generate appropriate output without any external source of information.



System Description

- Word embeddings:
  The input sentence is transformed into a sequence of word vectors.
- Sequence-to-sequence model:
  The model consists of two long short-term memory networks (LSTMs): an encoder, which maps the input sequence to a thought vector, and a decoder, which maps the thought vector to the output sentence. The encoder and decoder are trained together on a set of training dialogues.
- Training set:
  We trained the model once on a set of reddit.com comments and once on the OpenSubtitles dataset. Empirically the results obtained from the OpenSubtitles set were more convincing.

Results

In contrast to the other presented chatbots, this one does not extract the information for a response from a given knowledge base, but learns how to answer questions in general. This makes it hard to evaluate the performance of the chatbot. We think the fact that the model learns to give sound responses, which fit the type of the question, is exciting by itself. Also, in many cases the system actually gives correct or interesting responses (see examples).

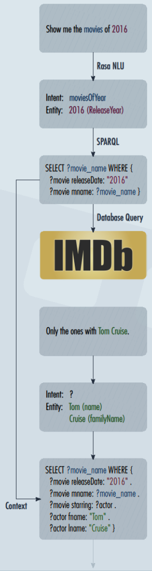
Hi, how are you?	good
What color is the sky?	blue
What color is fire?	red
I wear a blue jeans. What color do my jeans have?	i m not a fan
Tell me a joke!	i m not a joke
Are you a man or a woman?	i m not a woman
Do you think war is good?	i don t think so

Example questions (left) and the answers of our chatbot (right).

Modelling Conversation Context

In a second project we approached the same problem - automatically answering factoid questions about movies - with different tools and a different focus. We developed a system that keeps track of the context of the current conversation and is able to answer follow-up questions.

Imagine a conversation where you are talking about a specific actor and want to know his date of birth. Instead of restating the title of the movie, you will probably just ask "Who is the director of this movie?" and expect your dialogue partner to remember which movie you were talking about. We present a simple approach to resolve this kind of state dependency in certain scenarios, by simply combining database queries.



System Description

- Rasa NLU:
  Entity extraction and intent classification.

```
Input: Show me movies by Clint Eastwood!
Intent: moviesOfDirector
Entities: (Clint, name), (Eastwood, familyName)
```

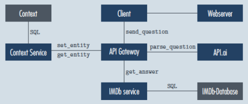
- GraphDB, SPARQL:
  We use data from IMDb, stored as RDF Triples in GraphDB:

```
636307, firstName, Clint
636307, lastName, Eastwood
Million Dollar Baby, director, 636307
```

- Context Tracking:
  We keep track of the context by combining the current SPARQL query with the previous one. E.g. if the previous query asked only for movies from a specific director, we can keep this filter for the new query as well.
  It is not obvious, when we want use this context and when to treat a new query independently. We decided to only use the context when the current input shows no clear intent.

Microservice Architectures

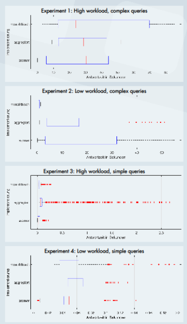
The recent developments and interest in cloud computing lead to increased use of microservice architectures, where classical monolithic applications are replaced by independent services, which rely on very little communication. This kind of architecture can have great advantages in scalability, robustness and innovative potential.



Component diagram of the microservice architecture with aggregate data interface. The atomic variant replaces the IMDb service by separate interfaces for different kinds of requests (get_movies_by_director, get_movies_by_genre,...).

We implemented three different versions of a chatbot and investigated how suitable each approach is in different settings. The three versions were:

1) a classical monolithic implementation.
2) a microservice implementation with an aggregated data interface (few services, diagram).
3) a microservice implementation with atomic data access (more services).



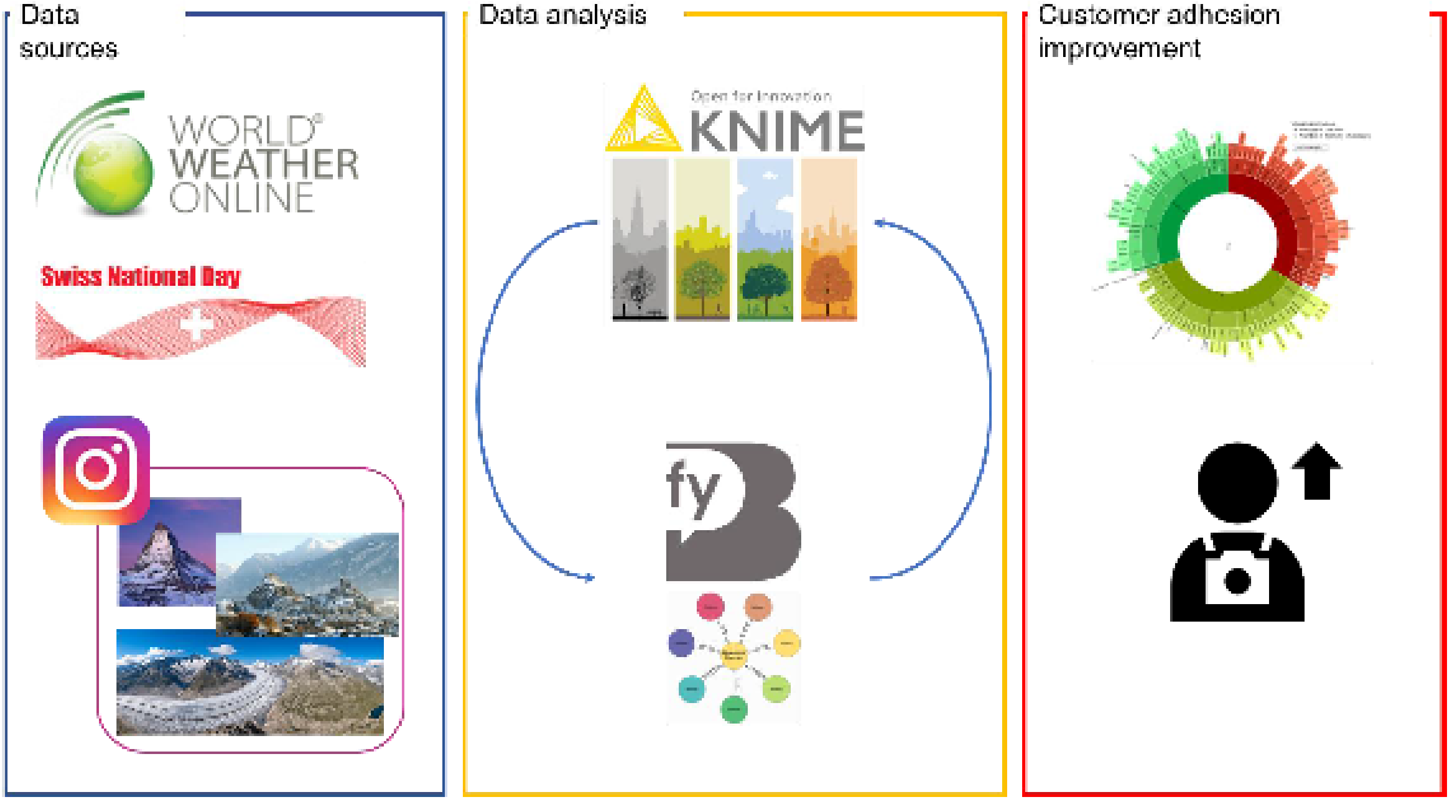
Results

We performed experiments in four different settings:

- Low or high workload: The chatbot is confronted with 2 or 100 queries per second respectively.
- Simple or complex queries: Queries filtering by one or multiple criteria (e.g. "Show me action movies from 2016") respectively.

Our results show that the performance of each architecture strongly depends on the setting. In particular splitting the data-interface into atomic services, requires an application-level-join for complex queries. This results in a significant drop in performance. Overall, we conclude that there is no general answer to the question monolith vs. microservices. The appropriate architecture depends on the requirements of the application.

Improving tourism marketing strategies by predicting the behavior of travelers using social media networks



- +200'000 comments and captions
- 3 years of pictures meta-information

Text Analysis – Semantic Concepts – Marketing Message