CITYTWIN – AI-based Decision Support System for Semantic Search and Analysis of Location-based Information for Urban and Site Planning

Oliver Buchmann[1], Maik Siegmund[2], Robert Kaden[1],Frank Iden[1]

[1] Erfurt University of Applied Sciences, Altonaer Str. 25, 99085 Erfurt, Germany,
oliver.buchmann;
robert.kaden@fh-erfurt.de;
srp@srp-gmbh.de
[2] TEAG Thüringer Energie AG Schwerborner Str. 30, 99087 Erfurt, Germany, maiksiegmund@gmx.de

Keywords: City-Twin, TF-IDF, TextRank, Word2Vec, named entity recognition, part to speach tagging

Abstract

The development of a knowledge-based decision support system for the evaluation and planning of location and urban development concepts was implemented. In order to achieve this goal, cross-domain ontologies were developed for interdisciplinary databases, which are then mapped in semantic networks. The exponential growth in computing power in the hardware sector alone can no longer solve this problem, but at the same time enables the application of new methods for storing and evaluating data. Essentially, it is no longer just about the digital recording of object properties in conventional databases, but also about the digital representation of their significance for specific questions and the linking of meanings across the boundaries of specialist domains. This information is stored in a multimedia knowledge base, together with the methods and rules for its use and the evaluations and decisions based on it. The motivation for this project is the rapidly growing amount of data, which extends across ever new specialist domains and can no longer be sufficiently integrated into the decision-making of experts using conventional methods of knowledge acquisition. After determining this data, it was linked to a georeferencing. Within the framework of the project, documents were analyzed with the help of AI and examined for semantic text corpora. This data was georeferenced. Various algorithms were used to accomplish this task, including TF-IDF, TextRank and Word2Vec.

1. Introduction

Digitalization and the sustainable zeitgeist have significantly altered the requirements for location planning. Urban and regional planners now gather location-related and planning-relevant documents and data from various online sources with confidence. Urban land-use planning, building law, urban development, development concepts for lighting, traffic, urban climate, accompanying greenery, as well as basic geodata and specialist geodata from the surveying authorities are areas of our expertise.

The research project CityTwin developed a decision support system for urban and location planning. The system utilizes artificial intelligence to create an interactive knowledge base and evaluate urban development factors. It includes decisions, strategies, and plans, all of which are crucial for the decision-making process. SRP GmbH implemented the spatial referencing, while Erfurt University of Applied Sciences implemented the semantic referencing. Our team's expertise in machine methods and user interface development ensured successful implementation of the decision support system. Semantic referencing establishes the structural relationship between entities of geo-objects and documents, such as land use plans, development plans, and development concepts. The documents are processed as unstructured text to automatically extract relevant information, such as geocoded addresses, points of interest, and keywords. The extracted information is then confidently assigned to the corresponding geo-object.

2. Related work

Several algorithms were developed to recognize semantic relationships and cluster text, which are crucial for the implementation of the work. This section describes the algorithms used.

2.1 Natural Language Processing

NLP, or Natural Language Processing, is a research field that involves the computer-aided analysis and processing of natural language. This field has its roots in the 1960s and has developed alongside the emergence of artificial intelligence. From the early days of AI, understanding and processing human language has been a central focus. Noam Chomsky's 1957 work(Chomsky, 1957), 'Syntactic Structures,' established the groundwork for the development of NLP. Chomsky's theory presented a formal framework for describing language, which contributed to the emergence of the Chomsky hierarchy of formal languages. The work of Saul Kripke and Richard Montague in the field of language logic also provided significant impetus for the development of NLP. This is an interdisciplinary field that combines artificial intelligence (AI) and computer science. Its main goal is to enable computers to understand, process, and generate human language in text or speech. This research area facilitates text processing, translation, and automated communication by allowing computers to process natural language like humans.

2.2 Knowledge Domain Analysis using Term Frequency-Inverse Document Frequency (TF-IDF) for Information Extraction

TF-IDF is an acronym for 'Term Frequency-Inverse Document Frequency' and is a statistical measure utilized in information

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-4-W5-2024-63-2024 | © Author(s) 2024. CC BY 4.0 License.

retrieval and natural language processing to assess the significance of a word in a document or collection of documents. The TF-IDF measure is employed to recognize keywords in a text document or to rank documents based on their relevance to a specific query. This method has evolved over time from various works by researchers in the fields of information retrieval and text processing. The concept of TF-IDF was first introduced by Karen Spärck Jones, a British computer scientist, in the 1970s(Jones, 2005). By combining the frequency of a word in a document with its rarity in the entire document collection, she proposed a method for determining its importance. The fundamental idea behind TF-IDF is that words with high Term Frequency (TF) in a document but low Inverse Document Frequency (IDF) in the entire document collection are more significant. This score determines the importance of a word in the document. TF-IDF is a fundamental technique in text processing applications like search engines, text classification, text clustering, and information extraction. It analyzes texts and highlights relevant information by assigning a score based on word frequency and uniqueness. Its significance in text analysis and information retrieval cannot be overstated.

2.3 TextRank Algorithm for Automatic Text Summarization and Keyword Extraction

TextRank is an automatic text summarization and keyword extraction algorithm based on graph theory. It was developed by Rada Mihalcea and Paul Tarau in 2004(Mihalcea and Tarau, 2004a). The idea behind TextRank is to represent text documents as graphs, where the words or sentences are represented as nodes and the relationships between them are represented as edges in the graph. The TextRank algorithm uses a variant of the PageRank algorithm originally developed by Google to rank web pages. TextRank evaluates the importance of words or phrases in the text based on their connection to other words or phrases in the document. Words or phrases that have many connections to others are considered more important and are therefore selected for summarization or keyword extraction. TextRank has applications in several areas, including automatic text summarization, keyword extraction from documents, and information extraction from unstructured text. It is an example of an unsupervised text processing algorithm that can work without human guidance to analyze text and extract important information.

In order to stop climate change caused by the human greenhouse effect, funding has been increased for research projects to develop and demonstrate solutions to reduce CO2 emissions. The focus and procedures of these research projects are always presented in text-based descriptions. Natural language processing makes it possible to process and analyze such data. The algorithms TF-IDF and Textrank described above were used for this purpose.(Graph-based research field analysis by the use of natural language processing: An overview of German energy research - ScienceDirect, n.d.)

2.4 Word2Vec Technology for Semantic Vectorization of Words

Word2Vec is a popular word vector calculation algorithm used in Natural Language Processing (NLP) to convert words into numerical vectors that are machine readable. These vectors can be used to capture semantic similarities between words and can be used in various NLP applications such as machine translation, text classification, and named entity recognition.Word2Vec was developed in 2013 by Tomas Mikolov and his team at Google Research (Mikolov et al., 2013). It is an extension of earlier work by Mikolov and others in the field of word vector models. The basic idea behind Word2Vec is to represent words in a high-dimensional vector space so that similar words are close together. This is achieved by the so-called ”Continuous Bag of Words” (CBOW) and the ”Skip-Gram” model, two different approaches to computing word vectors. The CBOW model tries to predict a target word based on its context words, while the skip-gram model tries to predict context words based on a target word. By training a neural network on large text corpora, the vectors are learned and can then be used in NLP applications to capture semantic relationships between words. Word2Vec has revolutionized word representation in NLP research and application by providing an effective and efficient way to extract semantic information from large text data sets. It has also paved the way for many other advanced word vector models, such as GloVe (Global Vectors for Word Representation) and FastText.

Overall, these algorithms demonstrate the diversity and importance of techniques in the field of text processing and semantic analysis, which are of great importance for various applications such as text classification, information extraction, and machine translation, and all of these algorithms were used in the project.

3. Methodology

The content of documents is analyzed for semantic meaning and clustering. The data formats supported are PDF, DOC(x) and HTML. The documents to be examined contain graphics and text. The text passages are structured by headings, footnotes and descriptions of figures and tables. This semantic information is considered as metadata because it relates to a specific topic and the graphics are not considered in the analysis. In addition, headers and footers contain redundant and irrelevant information. The following textual terms are relevant for georeferencing and are extracted: Points of Interest and street names with house numbers or house number suffixes (optional). These textual terms are related to the Official Property Register Information System (ALKIS) and a self-created ontology catalog.This information contains the relevant data and serves as keywords with semantic meaning. The information is stored in a relational database (PostgreSQL) and a graph database (Neo4J) after it has been determined.

The following computer-assisted activities are performed for document keyword analysis and georeferencing:

3.1 Transformation of Special Formats into Standard Text Format

Documents are converted from a variety of file formats into a standardized text format. Formatting elements such as font, font size, and graphics are removed. The basic structure of the document, including **cover page, directories, headings, headers and footers**, remains unchanged. However, the positioning of individual elements may differ from the original document.

3.2 Elimination of Structural Elements

All algorithms used are applied to unstructured text. Before applying these algorithms, the text is converted from special data formats (PDF, DOC, HTML) to a generic text format. The text corpus is then converted from structured text to unstructured

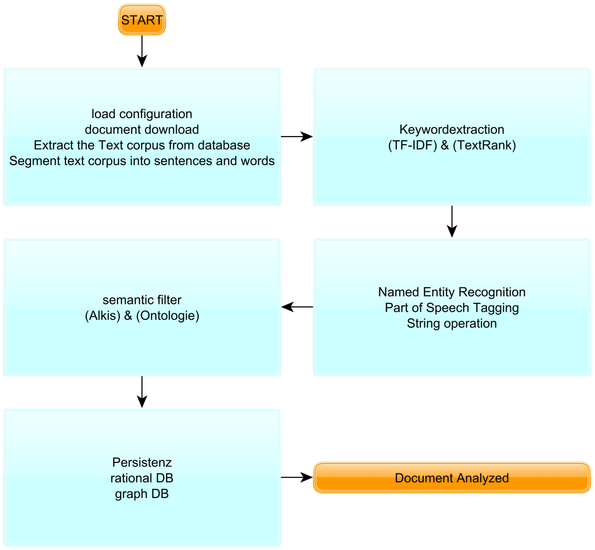


Figure 1. General process of the developed system

text. In text processing and natural language processing, various methods and filters play a crucial role in the efficient processing and analysis of text data. One important technique is stemming, which is used to reduce the root form of words. For example, the word ”houses” is reduced to the base form ”house” for consistent processing. Another step in text processing is segmentation, which is used to separate sentences and words in the text. This is crucial for understanding the structure of a text and isolating individual elements. Filters also play an important role. The ”Keep Word Filter” decides whether a particular word should be kept in the text corpus, while the ”Stop Word Filter” decides whether a word should be removed from the text corpus. The use of ”regular expressions” makes it possible to recognize patterns in the text and keep or remove them as needed. This is particularly useful for extracting specific information or filtering out unwanted content. The ”POS Tag Filter” is another tool that determines which word types should be retained in the text corpus. This is useful for retaining only relevant information. In addition, there are criteria such as Minimum Term Length (minimum length of a word) and Minimum Term Count (minimum number of words in a text segment), which determine the conditions under which a word is considered. Finally, text processing also includes steps such as ”removing line breaks and manual word separations” and ”removing tables of contents and page numbers” to make the actual text content accessible for further analysis or processing.

3.3 Taxonomical Categorization of all Terms

Unstructured text processing uses a variety of powerful algorithms and methods to extract meaning and structure from the available data.These proven approaches include

TF-IDF(Ping and Degen, 2016): This statistical approach is used for keyword identification and allows to highlight keywords in a text by evaluating their frequency in relation to their occurrence in the whole text corpus. **TextRank**(Mihalcea and Tarau, 2004b): A graph-based approach to keyword identification, TextRank analyzes the relationships between words in text to identify keywords. Important words are recognized based on their association with other words. **Word2Vec**(Goldberg and Levy, 2014): This neural network plays a central role in semantic keyword filtering. It allows for the representation of words in a vector space, which allows for the detection of semantic similarities between terms, which in turn improves the accuracy of keyword detection. **Named Entity Recognition**(Ritter et al., 2011): This uses a combined statistical and neural network approach to detect georeferenced and named entities in text. This is particularly useful for identifying places, names, and specific entities in text. **Part of Speech Tagging**(Gimpel et al., 2010): A statistical approach and neural networks are used to determine the part of speech in the text. This is crucial for understanding the meaning and function of words in a sentence. **pattern search (string operation)**: Pattern Search uses string operations to find and edit specific patterns or expressions in the text. This is especially useful when you need to find and modify specific text passages or formats.

Naive Bayes(John and Langley, 1995) algorithm is a simple but powerful algorithm for classification tasks based on Bayes' theorem. It is called ”naïve” because it assumes that the features used for classification are independent of each other. Despite this simplification, the Naive Bayes algorithm has proven effective in many real-world applications, particularly in text classification (such as spam detection) and medical diagnosis.

The basic idea is to calculate the probability of a particular event occurring, given the probabilities of associated conditions. In practice, this means that for a given input, it is calculated which class it most likely belongs to based on previous observations. The algorithm uses existing data to estimate class membership probabilities and then applies Bayes' Theorem to calculate the probability that a new input belongs to a particular class.

These algorithms and methods play a crucial role in transforming unstructured text into structured, interpretable information and are essential for analyzing and extracting knowledge from text data.

3.4 Georeferencing: Determination of Addresses and Points of Interest

The documents are georeferenced to the converted standard text. Addresses are searched using regular expressions and string patterns. The search is performed on the standard text, preserving the original structure and excluding graphics. The search for places of interest, on the other hand, is performed on the unstructured text corpus. The documents in the corpus contain a wealth of georeferences that need to be analyzed. The available documents are rich in georeferencing, covering a wide range of location information. This includes place names, federal states, district names, neighborhoods, street names, squares such as Alexanderplatz, landmarks, and even addresses with house numbers and optional house number suffixes. This variety of information provides a comprehensive basis for searching for geographic landmarks. However, our focus is on quality rather than quantity and precise positioning. Therefore, much of the information listed above is excluded from the search. Our targeted search is limited to the following key information:

Sights: We are particularly interested in identifying points of interest mentioned in the documents. This information is of high value and interest. **Addresses:** The addresses, including house numbers and optional house number suffixes, have been successfully captured.This allows for accurate location. Written descriptions like ”... Anton-Saefkow-Park, Höhe der Zufahrt zur Bötzowstraße...” are not considered in the search,

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-4-W5-2024-63-2024 | © Author(s) 2024. CC BY 4.0 License.

as this information cannot be processed by the algorithms and methods we use. The search for landmarks and addresses in the converted unstructured text corpus is performed using named entity recognition. Several models have been trained using supervised learning, and the model with the highest accuracy is used for our purposes. This precise approach ensures that the resulting geographic information is of the highest quality and can be used for a wide range of applications.

3.5 Semantic Referencing in Relation to Knowledge Representation

Some of the keywords found are generic results of the algorithm that do not add value in the context of urban and spatial planning. The relevant keywords for this context are rather abstract. In order to establish a concrete semantic reference, all identified keywords are compared with catalogs.

The first catalog is based on ALKIS(DOI, 2023) and contains names and categories. ALKIS is the nationwide data model that provides the technical basis for the content and structure of the cadastre in Germany. If a keyword and a name match, the corresponding category is derived.

The second catalog used is the system ontology catalog. In this context, an ontology, similar to that used in computer science, is used to describe information that has a logical connection. This catalog was created independently and is based on a variety of data sources, including the technical expertise and experience of an urban and spatial planner. It also includes the manual analysis of documents by urban and spatial planners and software developers, including justifications for development plans, land use plans, policy development, and other specialized planning documents. The results of the algorithms used are also manually validated. The keywords are linked to the ALKIS catalog and the system ontology catalog by means of technical terms. However, it is not guaranteed that every keyword is present in the catalog. A search for the technical terms in the document often results in only a slight match, which is determined by checking the character strings using various methods.

The problem described above is solved using Word2Vec. Word2Vec is a simple neural network that converts words into vectors. These vectors can be used to perform simple mathematical operations such as comparison and addition. A model was trained using unsupervised learning from 5.2 million German Wikipedia articles. By applying a defined similarity threshold, the keywords extracted from the document are compared with the entries in the ALKIS catalog and in the catalog of the system ontology. If the similarity value is reached or exceeded, an assignment is made.

3.6 Methods and Strategies for Data Management and Storage

After the analysis the results can be saved in PostGreSQL(Group, 2024) and Neo4J(Neo4j, 2023). The following figure 2 shows a subgraph representing the links from the document to the ALKIS catalog and to the catalog of the system ontology. This is the result of the semantic document analysis. Different questions can be answered by specific queries to the DBMS.

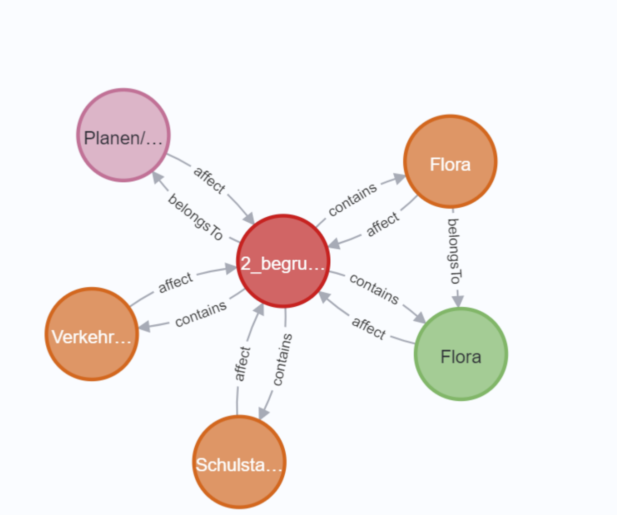
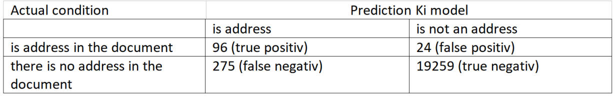


Figure 2. Subgraph of semantic referencing

4. Results

4.1 Quantitative and Qualitative Evaluation of the Used AI Model

The models were evaluated using the confusion matrix. The following table 3 shows this matrix. It is a binary classification with the categories ’true’ or ’false’ for the property: **Is a term an address**.



Actual condition	Prediction Ki model	
	is address	is not an address
is address in the document	96 (true positiv)	24 (false positiv)
there is no address in the document	275 (false negativ)	19259 (true negativ)

Figure 3. Confusion matrix

There are a total of 120 addresses in the document that were manually classified. The model correctly recognized and classified 96 of these addresses (true positives). However, it incorrectly classified 275 addresses as not addresses (false negatives) and identified 24 addresses as correct addresses when they were not (false positives). In addition, the model correctly recognized and classified 19259 non-addresses (true negatives).

This matrix can be used to calculate the following metrics that provide information about the quality of the model, as shown in the figure below:

$$Precision = \frac{TP}{TP - FP} = \frac{96}{96 - 24} = 0,8 \tag{1}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \tag{2}$$

$$Accuracy = \frac{96 + 19259}{96 + 19259 + 24 + 275} = 0,98 \tag{3}$$

$$Recall = \frac{TP}{TP - FN} = \frac{96}{96 - 275} = 0,25 \tag{4}$$

$$F1Score = 2 * \frac{Precision * Recall}{Precision + Recall} = \tag{5}$$

$$2 * \frac{0,8 * 0,25}{0,8 + 0,25} = 0,38 \tag{6}$$

where TP = true positiv

FP = false positiv
TN = true negativ
FN = false negativ

The model has an accuracy rate of 80%(1). The overall accuracy is 98%(4), while the rate at which the model recognizes correct classifications is 25%(4). The overall evaluation score is 38%(6). The main priority was to maximize precision, which is why the metrics for recall and F1 score were neglected for the time being, as the results will be validated in a later step. The priority was to maximize accuracy. This made it possible to neglect the recall and F1 score metrics for the time being, as the results will be validated in the next step. Several models were trained, with the training data being continuously adjusted. The following figure 4 shows a part of the results of the first model training. Each training was done automatically with different parameters. The algorithm was selected, either Maxent or Naive Bayes as statistical method or Perceptron as neural network. The CutOff value indicates how many times the feature must be present in the training data during the current iteration before it is discarded. The number of iterations represents the number of training runs. During training, the term found was determined and its probability was also recorded. In summary, automated training runs were performed with different parameters using different algorithms such as Maxent, Naivebayes and Perceptron. The cutoff value determined how many times a feature had to be present in the training data to be considered. The iterations and the probability of the term found were also documented.

The test data contained five addresses, of which Nußbaumweg 4 was not found in any of the evaluation runs. The model used was trained with the following parameters: Naïve Bayes classifier, CutOff: 2, Iterations: 30.Improvement of the models is possible, but requires additional training data in the form of unstructured text in standard format. The prototype uses this model.

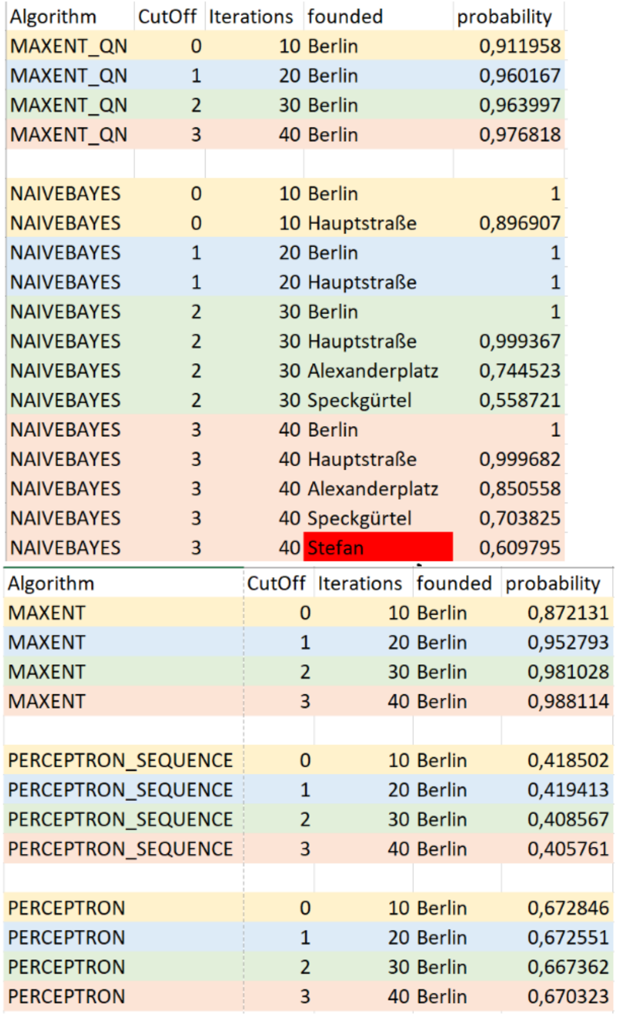
4.2 Identification of Text Patterns in the Documents

Addresses in the format 'Name house number (house number suffix)' are very unlikely to be recognized by the AI. This is compensated for by using a text pattern search. This significantly increases the number of recognized results, but many of these results must be classified as false negatives. The results found by both methods are then combined and validated using several database tables.

The first table contains landmarks based on the data source of https://geonames.org..

4.3 Validating the Results of Georeferencing

Addresses are validated using the following criteria: The address must contain both a text and a numeric part, the street name, the house number and, if applicable, the house number suffix must be listed in the address directory, and the district in which the street is located must be mentioned in the document. Points of Interest are validated based on geographic conditions.If an entry is found in the database table, it must either be within the specified bounding box or the distance to the set center must not exceed the maximum value. In summary, Points of Interest are validated based on geographic criteria by checking if the entries are within the specified bounding box or if the distance to the center does not exceed the maximum value.



Algorithm	CutOff	Iterations	founded	probability
MAXENT_QN	0	10	Berlin	0,911958
MAXENT_QN	1	20	Berlin	0,960167
MAXENT_QN	2	30	Berlin	0,963997
MAXENT_QN	3	40	Berlin	0,976818
NAIVEBAYES	0	10	Berlin	1
NAIVEBAYES	0	10	Hauptstraße	0,896907
NAIVEBAYES	1	20	Berlin	1
NAIVEBAYES	1	20	Hauptstraße	1
NAIVEBAYES	2	30	Berlin	1
NAIVEBAYES	2	30	Hauptstraße	0,999367
NAIVEBAYES	2	30	Alexanderplatz	0,744523
NAIVEBAYES	2	30	Speckgürtel	0,558721
NAIVEBAYES	3	40	Berlin	1
NAIVEBAYES	3	40	Hauptstraße	0,999682
NAIVEBAYES	3	40	Alexanderplatz	0,850558
NAIVEBAYES	3	40	Speckgürtel	0,703825
NAIVEBAYES	3	40	Stefan	0,609795

Algorithm	CutOff	Iterations	founded	probability
MAXENT	0	10	Berlin	0,872131
MAXENT	1	20	Berlin	0,952793
MAXENT	2	30	Berlin	0,981028
MAXENT	3	40	Berlin	0,988114
PERCEPTRON_SEQUENCE	0	10	Berlin	0,418502
PERCEPTRON_SEQUENCE	1	20	Berlin	0,419413
PERCEPTRON_SEQUENCE	2	30	Berlin	0,408567
PERCEPTRON_SEQUENCE	3	40	Berlin	0,405761
PERCEPTRON	0	10	Berlin	0,672846
PERCEPTRON	1	20	Berlin	0,672551
PERCEPTRON	2	30	Berlin	0,667362
PERCEPTRON	3	40	Berlin	0,670323

Figure 4. Results of various trained models

4.4 Visualization of Results in the Neo4J Browser

The following figure 5 shows a subgraph with a specific query to the DBMS. The query is as follows Show me all addresses and points of interest from the document 'Planwerk Innenstadt.pdf' that are related to the keyword 'Wohnen'.

5. Conclusion and Further Work

The conclusion of the extensive research project "CityTwin", which deals with the development of a decision support system for urban and location planning, highlights important results and challenges. By using advanced algorithms such as TF-IDF, TextRank and Word2Vec, as well as techniques such as Named Entity Recognition and Part-of-Speech Tagging, an efficient analysis and semantic referencing of large text corpora could be achieved. Particularly noteworthy is the successful integration of geo-referenced data into the analysis process, which enables precise spatial positioning and analysis. However, challenges also arose, particularly in the accurate identification and classification of addresses and in the integration and validation of the results in databases. While the accuracy of the models was satisfactory in some areas, there were limitations in recall and F1 scores, indicating a need for further training and improvement of the algorithms. Overall, the CityTwin project impressively demonstrates how the combination of different methodological

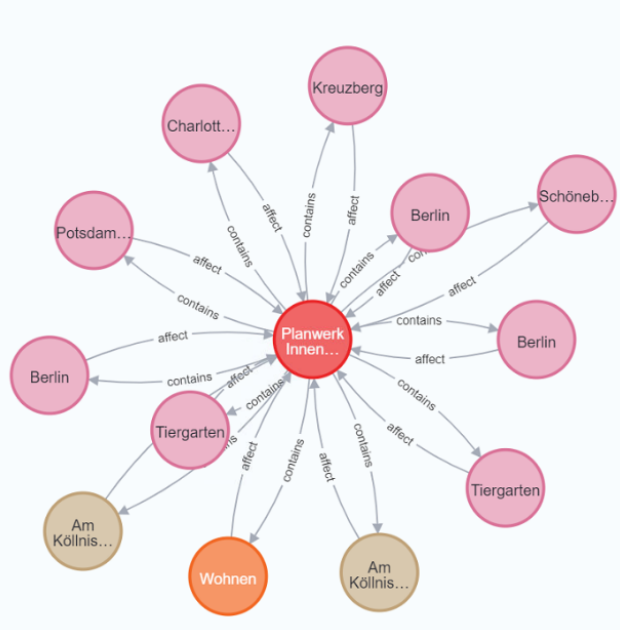


Figure 5. Partial graph with document, points of interest and addresses

approaches enables a comprehensive and in-depth analysis of documents and data relevant to urban planning. The results of this project provide valuable insights and tools for urban and regional planners to better address the challenges of modern urban planning.

The schematic diagram shown in Figure 6 illustrates the final document analysis flowchart in a detailed and clear manner. This flowchart includes both the sequential and conditional processing steps required to analyze the documents, and thus provides an essential basis for understanding the underlying processes. Within this schematic overview, all components used during the development process, as well as the specially developed functions and algorithms required to perform the analysis, are precisely listed. These components and functions have been developed with the aim of enabling efficient, reliable and accurate analysis of document content by integrating modern data processing and text analysis techniques.It should be emphasized that the entire source code, including all components used and the independently developed functions, is made publicly available on the GitHub platform. Making the code available on GitHub allows transparent insight into the development process, promotes the traceability of research results, and supports the collaborative further development of the project components by the scientific community. By making these resources available on GitHub, other researchers and developers can access, examine, modify, and adapt the developed tools for their own projects. This opens up perspectives for future research and the development of innovative approaches in document analysis and related fields. Integration and open access to such resources are crucial for the advancement of knowledge and technologies in the field.

Future research could focus on further increasing the accuracy of the models and integrating additional data sources such as graphs to further improve the analytical capabilities and applicability of the system. Due to the fact that the graphs have not been considered in the research so far, there is an opportunity to perform a complementary analysis of these graphs and integrate them into the existing process. The graphics have various characteristics, including geometries, textures, and text elements. To capture and identify these features, one possible approach could be to use Optical Character Recognition (OCR) or Optical Text Recognition.

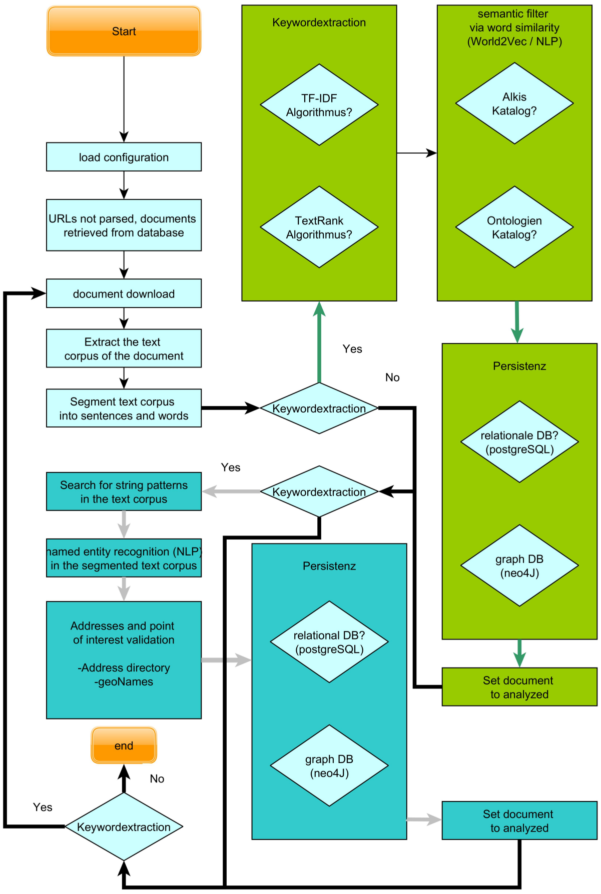

Start
load configuration
URLs not parsed, documents retrieved from database
document download
Extract the text corpus of the document
Segment text corpus into sentences and words
Keywordextraction
TF-IDF Algorithmus?
TextRank Algorithmus?
semantic filter via word similarity (World2Vec / NLP)
Alkis Katalog?
Ontologien Katalog?
Persistenz
relationale DB? (postgreSQL)
graph DB (neo4J)
Set document to analyzed
Yes
No
Keywordextraction
Search for string patterns in the text corpus
Keywordextraction
Yes
amed entity recognition (NLP) in the segmented text corpus
Addresses and point of interest validation
-Address directory
-geoNames
Persistenz
relational DB? (postgreSQL)
graph DB (neo4J)
Set document to analyzed
end
No
Yes
Keywordextraction

Figure 6. Document analysis program flowchart

References

Chomsky, N., 1957. *Syntactic Structures*. De Gruyter Mouton, Berlin, Boston.

DOI, 2023. Amtliches liegenschaftskatasterinformationssystem (alkis).

Gimpel, K., Schneider, N., O'Connor, B. T., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., Smith, N. A., 2010. Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments. https://api.semanticscholar.org/CorpusID:14113765.

Goldberg, Y., Levy, O., 2014. word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. *ArXiv*, abs/1402.3722. https://api.semanticscholar.org/CorpusID:12890187.

Graph-based research field analysis by the use of natural language processing: An overview of German energy research - ScienceDirect, n.d. `https://www.sciencedirect.com/science/article/pii/S0040162522006606`. (Accessed on 05/24/2024).

Group, P. G. D., 2024.

John, G. H., Langley, P., 1995. Naive Bayes classifiers. *Machine Learning*, 5(1), 241–273.

Jones, K., 2005. Some thoughts on classification for retrieval. *Journal of Documentation*, 61, 571-581.

Mihalcea, R., Tarau, P., 2004a. TextRank: Bringing Order into Text. 404–411. https://aclanthology.org/W04-3252.

Mihalcea, R., Tarau, P., 2004b. Textrank: Bringing order into text.

Mikolov, T., Yih, W.-t., Zweig, G., 2013. Linguistic regularities in continuous space word representations. 746–751.

Neo4j, 2023. Neo4j graph database analytics – the leader in graph databases.

Ping, N., Degen, H. E., 2016. Tf-idf and rules based automatic extraction of chinese keywords.

Ritter, A., Clark, S., Mausam, Etzioni, O., 2011. Named entity recognition in tweets: An experimental study.