

The Author grants to Chalmers University of Technology and University of Gothenburg the non-exclusive right to publish the Work electronically and in a non-commercial purpose make it accessible on the Internet.

The Author warrants that he/she is the author to the Work, and warrants that the Work does not contain text, pictures or other material that violates copyright law.

The Author shall, when transferring the rights of the Work to a third party (for example a publisher or a company), acknowledge the third party about this agreement. If the Author has signed a copyright agreement with a third party regarding the Work, the Author warrants hereby that he/she has obtained any necessary permission from this third party to let Chalmers University of Technology and University of Gothenburg store the Work electronically and make it accessible on the Internet.

"AuTopEx": Automated Topic Extraction Techniques Applied in the Software Engineering Domain

The design and evaluation of an approach for Automated Topic Extraction

Jonathan Klemetz
Magnus Johansson

© Jonathan Klemetz, June 2016.
© Magnus Johansson, June 2016.

Examiner: Christian Berger
Supervisor: Alessia Knauss
Supervisor: Hang Yin

University of Gothenburg
Chalmers University of Technology
Department of Computer Science and Engineering
SE-412 96 Göteborg
Sweden
Telephone + 46 (0)31-772 1000

Department of Computer Science and Engineering
Göteborg, Sweden June 2016

”AuTopEx”: Automated Topic Extraction Techniques Applied in the Software Engineering Domain

Magnus Johansson
Gothenburg University
Software Engineering & Management
Lindholmsplatsen 1, 412 96 Gothenburg, Sweden
gusjmagn03@student.gu.se

Jonathan Klemetz
Gothenburg University
Software Engineering & Management
gusklejoa@student.gu.se

Abstract

Automatically extracting topics from scientific papers can be very beneficial when a researcher needs to classify a large number of such papers.

In this thesis we develop and evaluate an approach for Automatic Topic Extraction, AuTopEx. The approach is comprised of four parts:

- 1) Text pre-processing.*
- 2) Training a Latent Dirichlet Allocation model on part of a corpus.*
- 3) Manually identifying relevant topics from the model.*
- 4) Querying the model using the rest of the corpus.*

We show that it is possible to automatically extract topics by applying AuTopEx on a corpus of scientific papers on autonomous vehicles.

According to our evaluation AuTopEx works better on full-text articles than texts consisting of just title, abstract and key-words.

Finally we show that this approach is vastly faster than human annotators, although not as accurate.

The source code used to build AuTopEx can be found at:
(<https://github.com/Klemetz/TopicExtraction>).

1 Introduction

In this thesis we design and evaluate an approach for Automated Topic Extraction. Which is evaluated on papers in the Software Engineering domain, more specifically on autonomous vehicles.

1.1 Background

Automated Topic Analysis and Automated Topic Extraction allow researchers to extract the potential topics that are contained in a large text corpus. This has been tried in other scientific domains but (to the best of our knowledge) not in the field of Software Engineering.

1.2 Problem Domain & Motivation

In order to find relevant information, researchers often need to read a large number of published articles. This is especially true when conducting work like mapping studies or Systematic Literature Reviews and can be a very time-consuming process.

There is a lack of automated approaches to topic extraction that could support activities such as Systematic Mapping Studies, especially in the Software Engineering domain.

1.3 Research Goal & Research Questions

Our research goal is to investigate the automation of Topic Extraction from scientific papers in order to support time-consuming activities such as Systematic Mapping Studies [18]. We investigate extraction from both full-text articles and texts containing only title, abstract and keywords using a topic model called Latent Dirichlet Allocation (LDA).

The research goal has been divided into three research questions.

RQ 1: How can we support Automatic Topic Extraction for scientific papers in the Software Engineering domain?

RQ 2: Which approach is better for Automatic Topic Extraction: a) Extraction from title, abstract and keywords or b) Extraction from full text paper?

RQ 3: How well does the approach of using Latent Dirichlet Allocation (with suitable pre-processing) perform compared to a manual method?

1.4 Contributions

In this paper we present an approach (which we call "AuTopEx") for applying Automated Topic Extraction on a large number of scientific papers.

From the extracted data, relevant topics are identified and labeled. Researchers can then also automate the process of finding which papers in the corpus that are most likely to deal with the relevant topics.

Researchers will benefit from AuTopEx as it shows the applicability of Natural Language Processing (NLP) techniques in the Software Engineering Domain.

1.5 Scope

We construct and evaluate an approach for Automatic Topic Extraction using Latent Dirichlet Allocation (LDA). This could contribute in making Automatic Topic Extraction a viable approach in Software Engineering research. We do not evaluate other statistical models such as the n-gram model or term frequency-inverse document frequency. However, their potential use in our approach is discussed in the Conclusions and Future Work section.

1.6 Structure of the Article

Section 2 presents related work on Automated Topic Extraction. Section 3 covers our Research Strategy. In Section 4 we answer our research questions. First we describe AuTopEx and go into detail about its implementation. Then we evaluate the results from implementing AuTopEx compared to human performance. Evaluations are made on two corpuses, one corpus containing articles in full-text and the other only title, abstracts and keywords. Section 5 contains analysis and discussion of the results. In this section we also discuss the validity threats to our findings. Section 6 concludes our findings and discusses what implications they may have for future research.

2 Related Work

We have not found any articles dealing with tools specifically tailored towards automation of Systematic Mapping Studies. These do however share some similarities with Systematic Literature Reviews (SLR). Hence we can discuss tools that support the latter.

According to Marshall and Brereton [16], the two most popular frameworks for tools that support SLR:s are the Projection Explorer Pex and ReVis which both make use of Visual Text Mining techniques. Projection Explorer Pex [6] can create a visualization from a set of textual documents either by building a vector representation of the text corpus (which is handled as table data to derive similarity information). It can also compute similarities by directly comparing text against text. It could possibly be used in helping with document classification during a mapping study.

ReVis [5] supports primary studies selection during SLRs. Among its tools is the possibility of visualizing the relationships of potential primary studies. A 2D document map shows content and similarities of different documents. This is based on converting the documents into multi-dimensional vectors which can be reduced using stemming, by eliminating stop words and using projection techniques. ReVis only uses title, abstract and keywords for this document map however, and we ideally want to use full-text articles to discover topics.

CitNetExplorer analyzes citation patterns in scientific literature. The tool collects bibliographical data and constructs a citation network which can then be analyzed and visualized[11]. This could be useful for mapping a specific research topic, since key word searches could miss out on papers that do not contain these key words.

VOSviewer [10] is a tool for creating and visualizing bibliographical networks. It can also use text mining to create term maps from a text corpus. Part of speech-tagging is used to identify noun phrases and a technique for choosing the most relevant noun phrases is applied. Maps and clusters can then be created and visualized.

In [12] the creators of CitNetExplorer and VOSviewer discuss the limitations of both tools. They argue that the loss of information occurring when applying these techniques is very hard to measure and that they should be used as a complement rather than substitute to expert judgment.

One approach for speeding up topic extraction could be automatic summarization of articles. The abstract of a scientific paper is meant to provide a quick overview, but does not necessarily provide enough key information for the researcher. Automatic summarization techniques can capture scientific concepts such as Hypotheses, Method and Background on a sentence level [14] and thus provide more information than just an abstract. However, this work builds on having many domain experts manually annotate a large number of scientific papers used for training the machine learning

classifiers [15]. Such an undertaking is out of scope for this thesis.

The "Latent Dirichlet Allocation" method is widely used and applicable in the discipline of Natural Language Processing (NLP) [2]. As Blei puts it "The simple LDA model provides a powerful tool for discovering and exploiting the hidden thematic structure in large archives of text" [2]. When attempting to extract topics from a large corpus as the purpose is for the AuTopEx approach, a tool like the LDA method is very compelling. Blei has also provided some comparisons with other models which makes the choice of applying LDA an attractive option. He shows that even though LDA is meant to perform "in the spirit of LSI" (Latent semantic indexing) [3], the LDA method outperforms the LSI method regarding perplexity measures [3].

But can we be sure that NLP tools such as the LDA method are fitting for the Software Engineering domain? Studies such as the one performed by Hindle et al [9], shows us that they can be applied but might not always be suitable. Hindle presents in his paper that in the domain of software engineering, neither LDA nor the n-gram analysis approach may be suitable if the intended goal is to extract topics from files containing computer code. However, we do not expect that code snippets will make up anything but a very small portion of the scientific articles that we want to apply Automatic Topic Extraction on.

3 Research Strategy

We have chosen Design Science as our research strategy. Hevner et. al [8] present seven guidelines for conducting, evaluating and presenting Design Science research. These address design as an artifact, problem relevance, design evaluation, research contributions, research rigor, design as a search process, and research communication.

The artifact is in our case an approach based on Natural Language Processing techniques. In this approach we apply a number of pre-processing steps on a large corpus of texts. Then we automatically extract topics from the corpus. Finally we automatically classify the papers based on the extracted topics.

The problem relevance is the fact that doing this manually is a very time-consuming process.

Evaluation will be done by comparing the topics that the machine learning algorithm produces with annotation made manually by humans. Both of the authors will first do manual annotation of the same papers separately and then confirm that there exists an inter-annotator agreement. Basically that both authors have identified the same topics in each paper.

As for research contributions we are transferring knowledge from the domain of language technology to the area of Software Engineering. We also believe the resulting approach can be helpful in future research where speeding up topic extraction can be beneficial.

When it comes to research rigor we are actively selecting and applying appropriate theories and methods both when constructing and evaluating the resulting artifact.

Design as a research process has been a must from the start, since this approach has not been applied on scientific articles about Software Engineering before. Investigating existing tools and techniques that are already being used for similar purposes and how to implement them properly, is the whole foundation of constructing our approach. We have worked in iterations during the whole process, constantly improving our approach by trying out different ways of working with existing tools and developing our own tools where needed.

Finally research communication must be taken into account. Since this thesis is concerned with research, we ensure that we explain our methods as thoroughly as possible so that other researchers can evaluate the approach. All of our code is open source and available under the Gnu General Public License Version 2 at (<https://github.com/Klemetz/TopicExtraction>) along with proper documentation so that others can apply the approach themselves.

4 Results

4.1 AuTopEx

In order to answer **RQ1: "How can we support Automatic Topic Extraction for scientific papers in the Software Engineering domain?"**, we have developed the approach AuTopEx, which supports Automatic Topic Extraction.

AuTopEx can be broken down into four steps:

1. Pre-processing the articles of a large corpus of scientific paper.
2. Training a Latent Dirichlet Allocation (LDA) model using 10 percent of the pre-processed papers.
3. Manual identification and labeling of relevant topics returned by the model.
4. Automatic classification (extracting the topics) of the rest of the corpus by querying the LDA model.

Some of the papers will be annotated manually before automatic classification. We can evaluate the accu-

racy of the model by comparing this manual annotation with the automatic classification of the same papers.

We have chosen existing tools to help answering our research questions and to construct AuTopEx. Calibre (<https://calibre-ebook.com/>) is used for pdf-to-text conversion. The Natural Language ToolKit(NLTK) (<http://www.nltk.org/>) is used for pre-processing individual texts and Gensim (<https://radimrehurek.com/gensim/>) is used for applying the machine learning algorithms. Complementary tools in the form of Python scripts have been developed by the authors when needed.

4.1.1 Text Pre-Processing

Pre-processing the scientific papers follows a pipe-line of six consecutive steps:

1. Pdf-to-text conversion
2. Converting all text to lower-case
3. Tokenization
4. Removal of stop words, numbers and punctuation
5. Lemmatization
6. Removal of references section

Pdf-to-text conversion

The first step of preparing the individual scientific papers is to convert them from pdf to text format. We use the free and open-source Calibre software. The reasons are two-fold: a) Unlike other tools we tried, Calibre handles ligatures well and b) dehyphenation. If a word is cut off with a hyphen at the end of a column, the program checks if the hyphenated word exists elsewhere in the document without the hyphen and dehyphenates it if that is the case. This ensures that more accurate words remain in the document.

Converting all text into lower-case

All text is transformed into lower-case format, to ensure correct multiplicity of words even if they appear at the start of a sentence. It is also important that the list of stop words (introduced below) are all in lowercase.

Tokenization

Tokenization means we break down the stream of text into meaningful elements. In our case this means individual words (all contiguous alphabetic characters become part of a token) which are then separated from symbols such as punctuation. The Natural Language Toolkit(NLTK) has a number of tokenizers, we recommend their "regextokenizer" for this.

Removal of stop words, numbers and punctuation

Latent Dirichlet Allocation uses a Bag-of-words model [3]. This means that neither grammar or word order is important, only the multiplicity of the words. Thus we can now safely remove all punctuation and also stop words (such as "a", "and", "if", "or" etcetera).

Greek letters are often used as mathematical notation in scientific articles. Such a symbol on it's own has little to no semantic value for an annotator examining our results. Neither do we expect numbers from the articles to hold any semantic importance in the topic extraction so these are removed as well. This is easily solved by only allowing alphabetical words in the tokenizer.

Punctuation is also removed using regular expressions.

Stop words are removed by using a stop word list. We use the default stop words list from NLTK, and supplement it with more words that we deem have no semantic value. For example the word "fig." very commonly appears next to images and graphs in research papers. This word pollutes the results rather than give the topics any semantic meaning.

Lemmatization

Within a document a word can use several forms (such as "organize", "organizing" or "organizes") while all referring to the same concept, and we are interested in the multiplicity of this concept. Stemming is the process of reducing inflected words to their word stem. A stemmer only operates on the word at hand by cutting of the word stem. "Organize", "organizing" and "organizes" would all be reduced to "organ" using a stemmer, which is not what we want since the word now has an entirely new meaning.

Lemmatizing is closely related to stemming in that it reduces inflected words, however it reduces the word form to linguistically valid lemmas, using algorithms that deal with grammar and a built-in dictionary. For this purpose we use the WordNet Lemmatizer included with NLTK.

Here are a few example sentences. "They walk down the road. She walked by him. The elephant walks on four legs while we are used to walking on two."

Using the most common stemmer (Porter) we get: "They walk down the road . She walk by him . The eleph walk on four leg while we are use to walk on two ."

Using the WordNet lemmatizer we instead get: "They walk down the road . She walked by him . The

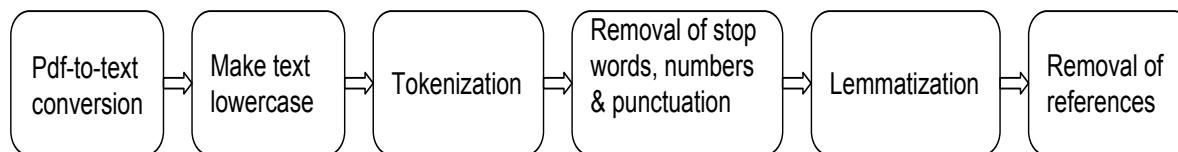


Figure 1: The steps involved in text pre-processing.

elephant walk on four leg while we are used to walking on two .”

The differences between stemmers and the basic differences between stemmers and lemmatizers are discussed in [13].

Removal of References section

Finally, all of the papers have a ”references” section at the end. We do not want the words in this section to pollute the article at hand. This section is removed by finding the last occurrence of the word ”reference” (remember we have lemmatized all words) and removing all remaining words in the document including ”reference”. In the rare event that a reference has the word ”reference” in it, some references might remain in the document. On the whole however we do not expect this to have a major impact on the results.

Pre-processing Title, Abstract, Keyword texts

Pre-processing these texts uses the same pipe-line as the full text method outlined at the beginning of this section but with the first and last step removed. This is because we already have the abstracts available in text format and they don’t contain any references.

Extracting title, abstracts and keywords from a full-text paper can be difficult, since not all articles are formatted in the same way. Some papers do not even include the keywords in the article document. Therefore we chose to extract this information using meta data stored in the software Endnote used by the researchers who provided us with the data used for the evaluation. Endnote can produce a single text file that contains author names, publishing year, publication, title, abstract and keyword for all articles that you want to perform topic extraction on.

A Python script extracts the relevant meta data (title, abstract and keywords) and saves a separate text file for each article (the model needs an entire corpus of papers to work with) naming them in the format author-publication-year-title for identification purposes when doing the evaluation.

Then we clean each document the same way as we

did for the full-text articles (tokenization, removal of stop-words, lemmatization).

4.1.2 Training the model

Latent Dirichlet Allocation

The intent of using LDA in this study is to get topics from the documents in the supplied data sets. LDA can do this through its probabilistic, generative functionalities. So a trained LDA model will be able to point out topics for documents[19]. There are however quite a few ways of training an LDA model to achieve the queryable functionalities [19].

Most of the ways of training a model boils down to a guessing game. This guessing game begins when for every document every word has been assigned to a random topic. A topic is a list of words and how many instances of them there are. A word can reoccur several times in a topic, this gives it an increased chance to be dominant within this topic, and get more instances of itself within this topic. Then the algorithm, for every word in every document, looks for what topic the current word could fit in as well, then moves it there. Depending on how many instances of the current word there are in that topic already, the chance that the word will be moved there varies. Now when the current word is moved, the current words new topic which has received the current word will have an increased chance of receiving another instance of this word.

In short the guessing game can be described as that the LDA model gets better at guessing as it keeps at it, and a measurement of measuring how well a model guesses is it’s perplexity value [3].

Using the Gensim Framework

Gensim is a framework that is accessible through the programming language Python. The Gensim framework allows the user to build and train their own unique LDA model based on the users own corpora. Gensim also offers other kinds of machine learning algorithms outside the scope of LDA [19].

When creating an LDA model with Gensim, it requires a corpus that has been tokenized. In the case

of AuTopEx, the models trained are handed a number of the pre-processed text files. AuTopEx can then through the Gensim framework train an LDA model given a sample from the whole corpus.

Throughout the training process, the Gensim framework tells the user whether or not the model is improving by printing out what is called a perplexity measure [2]. An indication of whether the model is improving is, if the perplexity measure is decreasing for each iteration [3].

Then, when the perplexity measure is down to a pre-determined value, the LDA model can be saved down on the hard drive of the users system and reused on the entire corpus. This is where AuTopEx can return which topics are deemed most relevant for each document.

To find measures that act as good examples when training a model to perform as well as possible one can observe Bleis experiments[3]. When asking for more than a hundred topics in these experiments, the perplexity measure is not improving as much anymore unlike when the number of topics approach a hundred. Blei also presents in his paper that when training models, if training a model with a larger sample of ten percent, of the entire corpus, the gain in accuracy is not significant. However, approaching ten percent of the entire corpus for a training sample the gain in accuracy is certainly appealing [3].

4.1.3 Identifying and labeling relevant topics

The trained model provides us with up to 100 topics, each topic consisting of a set number of words. A script exports this data to a spreadsheet for easy access by the annotators.

An example of a complex topic (10 words) outputted after training could look like this:

```
(0.005): 0.012*communication + 0.009*channel
+ 0.008*packet + 0.008*velocity + 0.007*protocol
+ 0.006*follower + 0.006*platoon + 0.006*leader +
0.006*transmission + 0.006*controller.
```

What can be observed from this topic is that the words that follow a number and a star is related to the topic. Inside this topic there are several expressions, for example "0.012*communication". This expression and all other expressions that follows inside this topic will combined provide the interpreter guidance towards labeling the topic.

At a first glance it seems like the topic could be labeled as one of the words that it already contains, "Communication". The way this could be argued is be-

cause the topic also contains "packet" and "protocol". These words are tightly related with communication solutions/properties in software and computers in general. At a closer look there are other options for the label. Since the topic contains "platoon", "follower" and "leader" which all probably refer to a platoon of vehicles (the corpus being related to autonomous vehicles). The label could then arguably be something like "Networked vehicles" or "Cooperating Vehicles". Then again "Transmission" might be related to communication but could also refer to gearbox and we also have "velocity" and "controller" in the topic.

As you can see the labeling phase can prove quite difficult based on the number of words and their semantic relations.

We chose 7 words per topic but we encourage those who want to try this approach to experiment with the number of words per topic. In our experience, with fewer words the topics became more general (e.g. "Network") and with more words the topics became more specific ("Networked vehicles grouped in platoon"). Seven to us seemed like a good compromise because for this specific corpus of papers it gave us a large amount of varied topics.

It's important to note that not all topics from the trained model will be interpret-able by humans. This is due to the generative and probabilistic nature of the LDA model. A model will produce a number of bad topics with low scores. From our experience these topics will never be assigned to papers during the classification, so this is not a problem.

A very large majority of the topics from our corpus were however indeed interpret-able and covered a wide variety of areas (please refer to the Appendix for examples of the topics we got from the model and how we labeled them).

4.1.4 Querying the model

When an LDA model is finished and saved onto the hard drive, one can query this model with pre-processed documents in order to get the models opinion of what topics might exist in each specific document.

As an example, when we ask the model to return three possible topics for the paper "A Real-Time Multi-Sensor Fusion Platform for Automated Driving Application Development", the model outputs: "(37, 0.81872937773255205), (55, 0.078783842923631039), (78, 0.034186934006349756)"

This means that according to the model, topic 37 is the most probable topic for this document, followed by topics 55 and 78. These results are exported to a spreadsheet, where the researcher can look up the

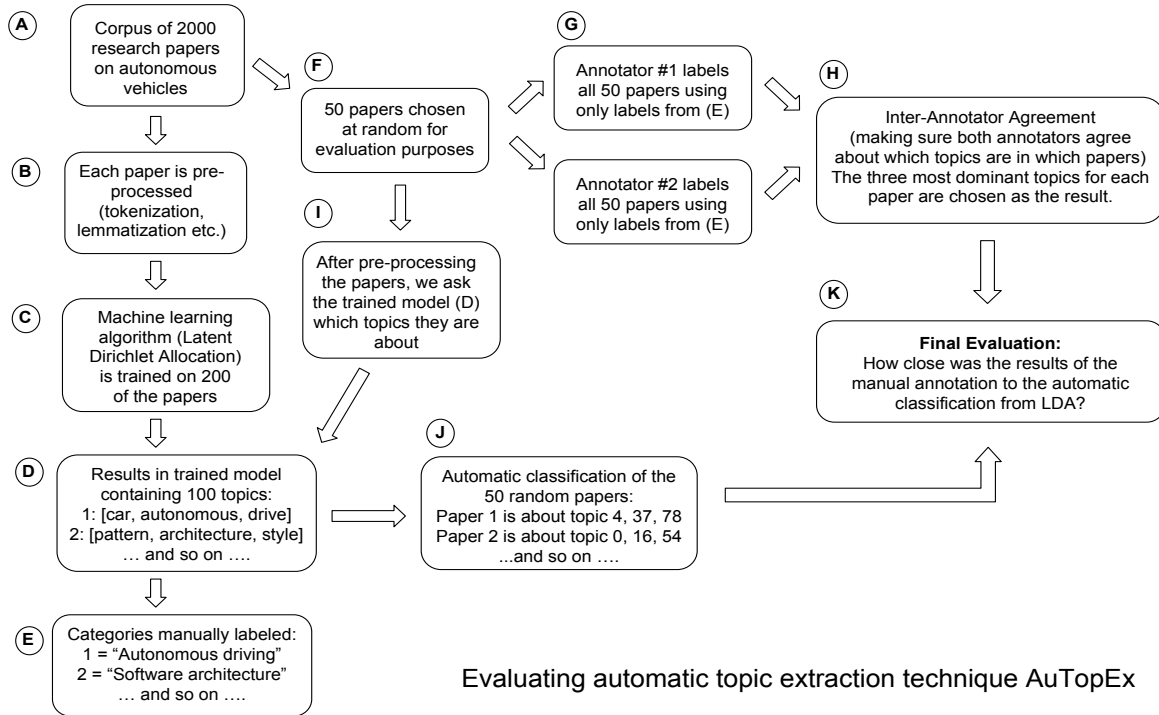


Figure 2: A simplified overview of how we evaluate the AuTopEx approach.

labels corresponding to these numbers.

4.2 Evaluating AuTopEx

4.2.1 Setting up the evaluation

The data set consists of 425 scientific articles related to autonomous vehicles. These papers had been screened based on certain inclusion and exclusion criteria for an actual Systematic Mapping Study being performed by researchers at Chalmers University of Technology, thus we deemed it an excellent data set for performing our evaluation.

For each of our two evaluations 200 of the 425 scientific papers were selected at random for training the LDA model. This number was chosen because we expect the final mapping study to include at least 2000 articles, and it is considered good practice to use ten percent of the data set for training purposes when implementing LDA.

The 100 topics (containing 7 words each) from the model are now manually labeled by the authors. First each author labels all of the topics on their own and then check whether they disagree on any topic label. Any disagreements are solved by discussing the topic at hand. The labeling phase is arguably the most difficult part of the entire process because it requires the

annotators to have very good language skills as well as domain expertise. More on that in the "Threats to Validity" section of this thesis.

4.2.2 Evaluation method

From the remaining 225 papers 50 are chosen at random for evaluation purposes. We use a Python script for random selection as well, in order to eliminate any potential bias where an annotator could choose documents with very clear titles that were similar to the topics we already knew existed in the corpus.

All of the 50 documents are now read and annotated by each of the authors, if a document talks about a topic labeled in the previous step, it gets the same label.

After the human labeling is completed we process the same 50 documents using our trained LDA model. A Python script exports the most probable topics for each paper to a spreadsheet. We chose a two-fold approach for both full-text and title-abstract-keyword evaluations here: First we export the three topics with the highest probability weight according to the algorithm. Then we separately export all probable topics, no matter how low the probability is. This might give us insights into both how the Gensim implementation

of LDA works as well as tell us something about the documents being analyzed (mainly the number of probable topics per paper and their respective probability weight according to the algorithm).

For the purpose of supporting tasks such as document classification in Systematic Mapping Studies we are interested in knowing whether AuTopEx performs better with a data set consisting of full-text articles or a set where the articles only contain titles, abstract and keywords. In order to evaluate this, as well as getting a measure on how well the human annotators and the system agree with each other, we use an evaluation technique called precision and recall [1].

Before one can calculate the values for precision and recall one must first collect the required data. Rather than just presenting this data in tabular form, it helps to produce a confusion matrix, consisting of four fields. See the model below as an example. The four fields are labeled true positive, false positive, true negative and false negative.

In this study, true positives are the topics that are deemed by both the machine and the annotator as relevant for the given articles. False negatives are topics that have not been deemed relevant by the machine but have been deemed relevant by the human annotator. False positives are topics that the machine is returning as relevant topics but have not been deemed relevant by a human annotator. Lastly true negatives, are basically just the rest of the topics that have not been returned by the machine and that should not have been returned according to the human.

Figure 3: Example of the confusion matrix

		Returned by LDA?	
		Yes	No
Relevant?	Yes	True Positive	False Negative
	No	False Positive	True Negative

These boxes

would be filled with the values that has been described previously in the respective box. So to show an example of how this would be performed, please refer to

the data supplied in the first appendix. When looking at the first sheet in this spreadsheet, there are four columns, true positive, false negative, false positive and true negative that are of importance. The papers are listed on the left and for each papers corresponding row the values for each of these elements are represented. Since this study is focusing on how the different data sets (full-text vs title, abstract and keywords) perform against each other, one can observe at the bottom part of the sheets, the sums of all the precision and recall values are stored. Here the values from the entire data set are added together and presented. It is these sums of the true positives, false negatives, false positives and true negatives for each data set that are used and later presented inside these confusion matrices that is exemplified above.

When this data has been collected, the following equations can be applied to get the values of precision and recall.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

To bring a bit more clarity to what these values will indicate in the case of this study, lets quickly summarize. **Precision** serves as an indication of how many of the topics that are returned as relevant, are truly relevant. **Recall** represents how many relevant topics were returned by the system.

This study investigates if there is any preference for what type of documents to use when performing Automatic Topic Extraction. Thus, a value called an F-measure, which is a harmonic mean of precision and recall will be used in comparing the different results [1]. The F-measure can be a number between 0 and 1 and measures the accuracy of the test. The closer the result is to 1 the better.

$$FMeasure = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{3}$$

The harmonic mean from precision and recall gives us a good measure of which method is better: Applying LDA on full-text papers or on title, abstract and keywords.

With every query executed in the two LDA models (one for full-text, another for title, abstract and keywords) and all the human annotated data collected, we will now outline what the confusion matrices looks like with the corresponding values.

4.2.3 Evaluation 1: All LDA topics, full-text articles vs title, abstract & keywords

Figure 4: Full text, all topics

		Returned by LDA?	
		Yes	No
Relevant?	Yes	102	36
	No	813	3198

The values generated by the table above is:

$$Precision = \frac{102}{102 + 813} = 0,111 \quad 4$$

$$Recall = \frac{102}{102 + 36} = 0,739 \quad 5$$

$$FMeasure = 2 * \frac{0,111 * 0,739}{0,111 + 0,739} = \mathbf{0,193} \quad 6$$

Figure 5: Title, abstract and keywords, all topics

		Returned by LDA?	
		Yes	No
Relevant?	Yes	83	54
	No	591	2273

The values generated by the table above is:

$$Precision = \frac{83}{83 + 591} = 0,123 \quad 7$$

$$Recall = \frac{83}{83 + 54} = 0,606 \quad 8$$

$$FMeasure = 2 * \frac{0,123 * 0,606}{0,123 + 0,606} = \mathbf{0,204} \quad 9$$

Regarding "RQ 2: Which approach is better for Automatic Topic Extraction: a) Extraction from title, abstract and keywords or b) Extraction from full text paper?" The F-Measure is slightly higher for title, abstract and keywords. However with such a small difference we can't safely say that one type is better than the other.

To answer "RQ 3: How well does the approach of using Latent Dirichlet Allocation (with suitable pre-processing) perform compared to a manual method?" We assume that the human performance is perfect, since that is what is accepted and applied today in the Software Engineering domain. So when looking at the amount of false negatives stored (the amount of topics that should have been returned by the machine, but were not) in these two confusion matrices. The full-text gives us 36 and the title, abstract and keyword set 54. So that tells us that full-text data set returns the relevant topics more often than the title, abstract and keywords data set. So the full-text missed 36 topics that the humans had deemed relevant and the title, abstract and keyword missed 54. This is the indication of how much the humans and the algorithm disagree

4.2.4 Evaluation 2: Top 3 LDA topics, full-text articles vs title, abstract & keywords

Figure 6: Full text, top three topics

		Returned by LDA?	
		Yes	No
Relevant?	Yes	46	103
	No	103	3897

The values generated by the table above is:

$$Precision = \frac{46}{46 + 103} = 0,309 \quad 10$$

$$Recall = \frac{46}{46 + 103} = 0,309 \quad 11$$

$$FMeasure = 2 * \frac{0,309 * 0,309}{0,309 + 0,309} = \mathbf{0,309} \quad 12$$

Figure 7: Title, abstract and keywords, top three topics

		Returned by LDA?	
		Yes	No
Relevant?	Yes	35	115
	No	115	2735

The values generated by the table above is:

$$Precision = \frac{35}{35 + 115} = 0,233 \quad 13$$

$$Recall = \frac{35}{35 + 115} = 0,233 \quad 14$$

$$FMeasure = 2 * \frac{0,233 * 0,233}{0,233 + 0,233} = \mathbf{0,233} \quad 15$$

In regards of "RQ 2: Which approach is better for Automatic Topic Extraction: a) Extraction from title, abstract and keywords or b) Extraction from full text paper?" When we ask the model to only return the three most probable topics per paper we get a higher F-measure for the full-text articles and the title, abstract and keywords, than when we asked it to return all topics. This is probably because when

the Gensim framework only returns three topics, it returns fewer false positives, thus the value of precision is higher. Though due to probability, there is a smaller chance for the annotators to agree with the machine with only three returned topics. So the recall value is smaller, due to the higher value of false negatives.

Full-text also performs somewhat better than title, abstract and keywords when looking at the top 3 most probable topics.

Regarding "RQ 3: How well does the approach of using Latent Dirichlet Allocation (with suitable pre-processing) perform compared to a manual method?" The topics that the machine should have returned. When only using the three most likely topics, there are a lot more topics in the false negative boxes than when returning all topics. There is a bigger chance when returning all topics that the topic the annotator deemed relevant will show up. However, between the two data sets when only returning the three most likely topics, yet again, the full-texts model returns more relevant topics than the title, abstract and keywords. This is since the full-texts confusion matrices only contains 103 false negatives and the other 115.

4.2.5 Evaluation 3: Most probable LDA topic, full-text articles vs title, abstract & keywords

However we are also interested in looking at the most probable topic for each paper (the topic with the highest probability weight according to the algorithm) and comparing this to the human evaluation.

Therefore (for each paper) we also do a simple binary comparison to see if the most probable topic according to the machine is among the three topics identified by the human annotators.

Figure 8 provides a simplified overview of how this evaluation was performed. First we compared the labeling made by human annotators with the machines categorization for the full-text articles and secondly we compared the same results for title, keyword and abstracts.

For RQ 2: Which approach is better for Automatic Topic Extraction: a) Extraction from title, abstract and keywords or b) Extraction from full text paper? its a bit difficult to motivate using precision and recall since if the machine would correctly return a relevant topic, there would still be two false negatives left. So a more simple approach is applied for this evaluation. One where if the machine returned a topic that was among the three the humans had deemed relevant it is labeled as a hit. The data

Most probable topic according to the Model	Documents with a hit	Missed documents	Hit-ratio
Full-text articles	17	33	0,34
Title/abstract/key-words	13	37	0,26

Figure 8: Only the top favorable topic returned from the queries

This is the result of a comparison of how often the most favorable topic returned from a query was among the three topics assigned from the annotators

sets model with most hits should therefore have returned the most relevant topic as their most probable topic. In the case of this study, please refer to figure 8 to observe that the full-text has a hit rate of 0.34 and title, abstract and keywords only have 0.26. So in this case it seems that the full-text data set has outperformed the title, abstract and keywords.

This is our final evaluation in regards to **”RQ 3: How well does the approach of using Latent Dirichlet Allocation (with suitable pre-processing) perform compared to a manual method?”**.

We simply check if the most probable topic according to LDA is among the three topics chosen by the human annotators for each article (see figure 8). Here the model also performs slightly better on full-text articles than on title, abstract and keywords. For 17 out of 50 documents, the most probable topic according to the model is also among the topics chosen by the annotators. For title, abstract and keywords. the same number is 14 out of 50. This gives a hit-ratio of 0.34 for full-text and 0.26 for title, abstract and keywords.

5 Analysis & Discussion

5.1 Analysis

With the result from the human annotators compared to the model, it seems fair to argue that the machine and humans agree more when both are supplied the articles in their entirety.

From the evaluation results using Precision And Recall we can see that the algorithm performs better when evaluating full-text articles rather than title, abstract and keywords and only looking at the top 3 topics.

When comparing the full-text, all topics result with the Abstract and keywords, all topics result, the F-measure of the lastly mentioned is however actually 0.011 higher than the F-measure of the full-text evaluation.

The reason why it still seems fair to argue that the full-text evaluation outperforms the Title, abstract and keywords, is because of when the machine presents its most probable choices of topics. Then the F-measure is much higher in the full-text evaluation. Just to add to this reasoning, another comparison was made with the singular most probable topic according to the machines and the annotators topics, as shown in figure 8. Yet again (with other measurements however) it is clear that when supplying full-text data sets to the machine, it performs better.

Worth mentioning is that when the model for title, abstract and keywords had been trained, it generated far fewer interpret-able topics when the time came to label them. In fact, for the full text model, 83 clear and usable topics were generated as for the abstract and keywords model, only 60 clear and usable topics were generated. So that explains the lower values of the true negatives in the abstract and keywords data sets.

Another reason why we wanted to compare the differences between the results of asking the model for all topics with the model’s top 3 topics was to show how the Gensim implementation of LDA produces a lot of topics for some documents with this data set. A lot of these topics get very low probability scores (see appendix) which is why there are a lot less false positives when we just look at the top 3 topics.

5.2 Discussion

Using AuTopEx for Topic Extraction

With all the tools in place a researcher only needs to do the following in order to perform automatic topic extraction:

1. Batch-convert all desired pdf:s.
2. Run the pre-processing script.
3. Train the LDA model using part of the corpus.
4. Query the model with the desired number of

remaining documents from the corpus.

From our experience document conversion and text-cleaning takes the longest time. For a large corpus (> 2000 scientific papers for example) each of these steps can take several hours. The researcher however does not need to be present while the programs are running. Training a model on 200 full-text papers took 40 minutes using a cheap laptop with a Celeron processor clocked at 2.0 GHz (utilizing two of the processor cores). Querying the trained model with 50 papers using the same computer is done in a couple of minutes.

Seeing as how it takes a human reader many hours to read and annotate 50 scientific articles, using an approach such as AuTopEx can greatly speed up topic extraction. Especially during tasks that require a researcher to read a large amount of articles, (such as when doing document classification in a Systematic Mapping Study).

Of course this requires that the model classifies the papers accurately enough, and there is room for improving AuTopEx here.

General Discussion

For the full text evaluation, the most probable topic identified by the algorithm was indeed a topic in the paper in 34 % of the cases according to the human annotators. This might not sound as a huge percentage, but seeing as this was the very first evaluation of the AuTopEx approach it seems very promising. Especially when one compares the many hours it takes for a human to read 50 scientific papers compared to the mere minutes it took the algorithm to produce this result.

It can be a good idea to perform word analysis on the corpus using NLTK after text pre-processing, for example checking a lot of the most popular words in the corpus. While time-consuming it can give insights into if some of the pre-processing steps might need adjusted. For example, perhaps there are still words in the corpus that could be considered stop words.

If batch-converting a large number of documents we recommend that the file sizes of the documents are checked afterwards. If any of the text-files have a size of 0 kilobytes the conversion has failed.

Discussion on Topics and their labeling

Labeling topics manually when performing evaluation can be a very difficult task. It requires both language skills as well as domain knowledge. Sometimes the words in a topic are acronyms or words that have no

meaning to those not familiar with the domain. Making sure that you found the correct meaning of the acronym (often an acronym has a number of meanings in a multitude of fields) or finding an explanation of a very niche word can be quite time consuming.

Interestingly enough adjectives were very uncommon in the results from our corpus. Besides "autonomous", which came up in 17 topics, the results were dominated by nouns, followed by verbs. For the full-text experiment only two other adjectives appeared, "intelligent" and "content", and the latter is also a noun. For title/abstracts/keywords the adjectives were more varied: "Intelligent", "dynamic", "generalized", "industrial", "artificial", "automatic" and "natural" appeared. The word "real" appeared three times (and was always accompanied by "time" in the same topic).

The dominance of nouns was quite helpful when labeling the technology-oriented topics often found in the software engineering domain. This is especially true when doing classification that does not take positives and negatives into account (we don't need colorful adjectives criticizing or praising something in a topic). Words like "car", "architecture" or "network" tells us a great deal on their own. Verbs are helpful in a supporting role (such as "driving" appearing in a topic with "autonomous" and "vehicle").

Another interesting note was that even though the entire corpus consisted of scientific papers, none of the topics produced in either of the two evaluation experiments were about scientific methodology. This is useful data to extract when performing tasks like Systematic Mapping Studies.

We found it interesting how the LDA model produced a lot of potential topics with low probabilities on the corpus on autonomous vehicle research. This could be due to how the scientific articles are written, but requires further study before any conclusions can be drawn.

5.3 Threats to Validity

It's important to remember that LDA is a probabilistic topic model, thus we are dealing with probabilities. If a human claims that a paper is about a certain topic and the machine claims that this probability is high, we only argue that the likelihood of this to be true is very high.

Properly labeling topics and scientific papers requires a lot from the human annotators. They must have excellent language skills as well as domain expertise in order to interpret each topic supplied by the model. One misunderstanding of a word could result

in an improper label, and this could impact the results of the evaluation.

We mitigated this by reading about concepts we were not familiar with before finishing the topic labeling, and looking up the meaning of any acronyms that appeared in the results. Both authors are software engineering students. Having previously studied concepts such as image processing or lane following for autonomous vehicles meant that we had a good understanding of a large majority of the topics produced by the model.

Then again, labeling 100 topics and reading 50 scientific articles (for two separate evaluations) can be difficult for humans. Stress, fatigue or just having a bad day can impact the accuracy both when performing topic labeling and when manually assigning topics to documents. We tried to mitigate this by taking breaks regularly during the evaluation. However if other researchers would redo our evaluation, using the same articles, we can not say for certain that they would label every single topic or classify the papers in exactly the same way.

Our main mitigation strategy for human error was that there were two of us doing the same work in parallel. We continuously compared the results between ourselves and where there were any disagreements regarding topic labels or which topics belong to a certain paper, we tried to reason with each other until we came to a result we could both agree on.

Another thing to consider when performing this kind of automatic topic extraction is that there is no way of handling positives and negatives. A paper that deals with a certain topic may actually reject the idea behind that topic. We mitigate this by not making any specific claims regarding the documents. We only state that in the results where human and machine agree that a topic exists in a document, that topic is indeed discussed in that specific document.

AuTopEx has only been evaluated on a corpus in the scientific domain of Autonomous Vehicles. We can't say with certainty how different the evaluation results would be if applying the approach on corpora from other domains. However, steps have been taken to make AuTopEx as generally applicable as possible. Especially by only limiting the tokenization to alphabetical words and using a very general stop word list. We recommend that anyone who uses this approach carefully consider if there are any special measures to be taken in the text pre-processing stage (e.g. adding words to the stop words list).

During the testing phase, we noticed that on some occasions several words would appear together as a single token after the texts had been cleaned and we sus-

pect this is due to bad quality of some of the original pdf:s. While it would be far too time-consuming to check the entire corpus manually for this we believe that this should very seldom occur in the data set we used for evaluation. This is because this data has been screened by researchers and only contains pdf:s published in 2005 or later.

6 Conclusions and Future Work

In this thesis we presented an approach for Automatic Topic Extraction which we call AuTopEx. This approach uses Natural Language Processing tools and techniques to pre-process the scientific articles of a corpus. Topics from this corpus are then extracted by training and querying a Latent Dirichlet Allocation (LDA) model. This model can be used to automatically classify the documents of the corpus (identifying which topics exist in which articles).

According to our results, Automatic Topic Extraction with Latent Dirichlet Allocation works better on full-text scientific articles than documents that consist of title, abstract and keywords. This is true both when querying the model for the most probable topic per article as well as when asking the model for the three most probable topics per article.

In our evaluation, the model's most probable topic was among the three relevant topics (according to the human annotators) in 34 % of the full-text documents evaluated. While the model is not as accurate than the human annotators it is important to note that this was the first evaluation of AuTopEx and perhaps most important of all: The model does this work in a couple of minutes while it takes humans many hours to perform the same task.

We believe that by refining this approach it will be possible to speed up topic extraction tremendously compared to manually reading and annotating papers.

Future work

One possible future experiment could be to allow the use of n-grams in the data set before performing the machine learning algorithm. If for example "autonomous vehicle" was considered a single word it could free up more space for other words to occur together with it in topics, possibly allowing for more meaningful interpretations by human readers. This process could also easily be automated. NLTK for example has the tool Collocations which performs n-gram analysis on documents.

Another idea that could possibly improve the results of our approach is to apply tf-idf on the text corpus

before training the model. Tfidf is the product of two statistics, term frequency and inverse document frequency. Term Frequency is the number of times a term occurs in a document. Inverse Document Frequency is a factor that diminishes the weight of terms that occur very frequently in the document set and increases the weight of terms that occur rarely. Thus a word like "the" will have a very low weight in tf-idf.

A high weight in tfidf is reached by a high term frequency (in the given document) and a low document frequency of the term in the whole collection of documents; the weights hence tend to filter out common terms. This could potentially be used for stop word removal.

The results would depend on how focused the language is in the different articles. An article which uses very broad language (using many synonyms for the same word) will produce different results than an article with very focused language. One idea could also be to duplicate the title of the paper a couple of times in each document before applying tf-idf. Seeing as how the title should reflect what the text is about this would help ensure that the most important words of the papers get a higher weight. Another experiment with tf-idf could be to give all nouns and verbs higher weight since they convey a lot of information about technologically-oriented topics.

A domain-specific lemmatizer for text preprocessing could be useful. This would however require a lot of work by several domain experts for a gold standard to be achieved and might be an unrealistic thing to wish for.

Automatization of the labeling stage could make the threat towards validity smaller while making the entire process quicker and easier to use, since there is less required input from the user. Such tools are already being applied[17].

Acknowledgements

We would like to thank Peter Ljunglöf of Chalmers University, Simon Dobnik of the University of Gothenburg and Victor Botev from Iris AI for their kind assistance. We would also like to thank our supervisors Alessia Knauss and Hang Yin of Chalmers University for their invaluable feedback.

References

[1] S. Bird, E. Klein, and L. Edward. *Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Sebastopol, California, United States, 2009.

[2] M. D. Blei. Introduction to probabilistic topic models. *Princeton University*, pages 1–16, 2011.

[3] M. D. Blei and et.al. Latent dirichlet allocation. *Journal of Machine Learning Research*, (3), 2003.

[4] J. Chang, J. Boyd-Graber, and et.al. Reading tea leaves: How humans interpret topic models. *Neural Information Processing Systems*, pages 1–9, 2009.

[5] R. Felizardo, Katia, N. Salleh, M. Martins, Rafael, E. Mendes, G. MacDonell, Stephen, Maldonado, and C. Jose. Using visual text mining to support the study selection activity in systematic literature reviews. *International Symposium on Empirical Software Engineering and Measurement*, pages 77–86, 2011.

[6] P. Fernando, V, C. Maria, O. F, and M. Rosane. The projection explorer: A flexible tool for projection-based multidimensional visualization. *Analytical and Bioanalytical Chemistry Volume 400, Number 4*, pages 1153 – 1159, 2011.

[7] R. Giuseppe and et.al. Semantic enrichment for recommendation of primary studies in a systematic literature review. *Digital Scholarship in the Humanities Advance Access*, pages 1–14, 2015.

[8] Henver, S. T. March, J. Park, and S. Ram. Design science in information systems research. *MIS Quarterly*, pages 1–14, 2004.

[9] A. Hindle and et.al. On the naturalness of software. *International Conference on Software Engineering (ICSE)*, pages 837–847, 2012.

[10] N. Jan van Eck and L. Waltman. Text mining and visualization using vosviewer. *ISSI newsletter*, pages 50–54, 2011.

[11] N. Jan van Eck and L. Waltman. Citnetexplorer: A new software tool for analyzing and visualizing citation networks. *Journal of Informetrics*, pages 802–823, 2014.

[12] N. Jan van Eck and L. Waltman. *Visualizing bibliometric networks*. In Y. Ding, R. Rousseau, & D. Wolfram (Eds.), *Measuring scholarly impact: Methods and practice*. Springer Publishing Company, 11 West 42nd Street, 15th Floor New York, NY 10036, 2014.

[13] A. G. Jivani. A comparative study of stemming algorithms. *International Journal of Computer Technology and Applications*, (Vol 2: Issue 6), 2011.

[14] M. Liakata, S. Dobnik, S. Saha, C. Batchelor, and D. Rebholz-Schuhmann. A discourse-driven content model for summarising scientific articles evaluated in a complex question answering task. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, page 747757, 2013.

[15] M. Liakata, S. Saha, S. Dobnik, B. Colin, and D. Rebholz-Schuhmann. Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics*, pages 991–1000, 2012.

[16] C. Marshall and P. Brereton. Tools to support systematic literature reviews in software engineering: A mapping study. *2013 ACM / IEEE International Symposium on Empirical Software Engineering and Measurement*, pages 296 – 299, 2013.

- [17] Q. Mei, X. Shen, and C. Zhai. Automatic labeling of multinomial topic models. pages 1–10, 2007.
- [18] K. Petersen, R. Feldt, S. Mujtaba, and M. Mattsson. Systematic mapping studies in software engineering. pages 1–10, 2008.
- [19] R. Rehurek and P. Sojka. Software framework for topic modelling with large corpora. *Natural Language Processing Laboratory Masaryk University*, pages 1–5, 2010.

title	true positive	false negative	false positive	true negative	total number of t	17 skräp	83 bra topics	
Fisheye optics for omnidirectional perception	2	1	16	64	18			
Data age based retransmission scheme for reliable control data exchange in platooning applications	2	0	19	62	21			89, 4, 16, 52,
Obstacle Avoidance in Real Time with Nonlinear Model Predictive Control of Autonomous Vehicles	3	0	15	65	18			
Intelligent Cruise Control Stop and Go with and without Communication	3	0	13	67	16			
Autonomous Navigation: Achievements in Complex Enviro	1	0	25	57	26			
Bayesian Network Based Collision Avoidance	2	1	17	63	19			
Experience, Results and Lessons Learned from Automated Driving on Germany's Highways	3	0	19	61	22			
Multi-Objective Path Planning using Spline Represent	3	0	20	60	23			
A Study on Autonomous Vehicle Development Process at University*	3	0	13	67	16			
Road Surface Recognition Using Laser Radar for Automatic Platooning	1	2	20	60	21			
Building a Prototype for Power-Aware Automatic Parking System	2	0	19	62	21			
A Computer Vision System for Detection and Avoidance for Automotive Vehicles	3	0	11	68	15			
Path Tracking of Autonomous Ground Vehicle Based on Fractional Order PID Controller Optimized by PSO	2	0	12	69	14			
Off-road Path Following using Region Classification and G Constraints*	3	0	19	61	22			
Self-Tuning PID Controller for Autonomous Car Tracking in Urban Traffic	2	1	18	62	20			
Shared Control of Autonomous Vehicles based on Velocity Space Optimization	2	1	19	61	21			
A 13,000 km Intercontinental Trip with Driverless Vehicles: Real-Time Coordination of Autonomous Vehicles	1	1	17	64	18			
Accurate and Efficient Traffic Sign Detection Using Discrim	3	0	16	64	19			
DeepDriving: Learning Affordance for Direct Perception in	2	1	16	64	18			
A Robust Algorithm for the Detection of Vehicle Turn Signa	2	0	5	76	7			
Constrained Global Path Optimization for Articulated Steeri	3	0	19	61	22			
360° detection and tracking algorithm of both pedestrian an using fisheye images	3	0	15	64	19			
State your position	2	0	18	63	20			
A robotic platform to evaluate autonomous driving systems	3	0	17	61	22			
Coordinated control of multiple vehicles with discrete-time periodic communications	1	3	17	63	17			89, 4, 16, 52,
Real-time Implementation of a Novel Safety Function for Pr	3	0	12	68	15			
Coordinated Path Following Control for a Group of Car-like	1	2	9	71	10			
A Combined Model- and Learning-Based Framework for In	2	1	17	63	19			
Towards a Framework for Testing Drivers' Interaction with	1	1	16	65	17			
Adopting WirelessHART for In-Vehicle-Networking	2	0	18	63	20			
Terrain Mapping for Off-road Autonomous Ground Vehicle	3	0	18	62	21			
Incremental Sampling-based Algorithm for Minimum-violation Motion Planning	1	2	21	58	23			
Vision-based Nighttime Vehicle Detection and Range Esti	3	0	19	62	21			
Design and Comparative Analysis of a Driveless LED light	0	3	4	76	4			
Local Path Planning for Off-Road Autonomous Driving With Avoidance of Static Obstacles	3	0	24	56	27			
HOG Based Multi-object Detection for Urban Navigation	2	1	17	63	19			
Genetic Algorithm Approach for Locating Automatic Vehicl Identification Readers	0	1	17	65	17			
Reliable Intersection Protocols Using Vehicular Networks	3	0	11	69	14			
INTELLIGENT TRAFFIC WITH CONNECTED VEHICLES	2	1	18	62	20			
MCMC Particle Filter for Real-Time Visual Tracking of Vehi	2	1	22	58	24			
Globally Asymptotically Stable Filter for Navigation aided b and Depth Measurements	0	3	22	58	22			
A Real-Time Multi-Sensor Fusion Platform for Automated	1	2	2	78	3			
A full-3D Voxel-based Dynamic Obstacle Detection for Urban Scenario using Stereo Vision	3	0	11	69	14			
A Real-Time Trajectory Control of Two Driving Mobile Rob	1	2	18	64	19			
Vehicle Automation in Cooperation with V2I and Nomadic Devices Communication	2	1	19	61	21			
Automatic vehicle classification and tracking method for ve movements at signalized intersections	3	0	19	61	22			
Multi-Target Tracking using a 3D-Lidar Sensor for Autono	1	2	19	61	20			
Traffic Sign Representation using Sparse-Representations	1	1	20	61	21			
Speed Profile Optimization for Vehicles Crossing an Intersection Under a Safety Constraint	3	0	14	66	17			
Sum	102	36	813	3198	918			
	true positive	false negative	false positive	true negative	total number of topics			
	Precision =	0,111						
	Recall =	0,739						
	F =	0,1930094118						

title	true positive	false negative	false positive	true negative	total number of topics	17 skräp	83 bra topics
Fisheye optics for omnidirectional perception	1	2	2	78	3		
Data age based retransmission scheme for reliable control data exchange in platooning applications	2	0	0	80	3		
Obstacle Avoidance in Real Time with Nonlinear Model Predictive Control of Autonomous Vehicles	0	3	3	77	3		
Intelligent Cruise Control Stop and Go with and without Communication	0	3	3	77	3		
Autonomous Navigation: Achievements in Complex Enviro	0	3	3	77	3		
Bayesian Network Based Collision Avoidance	1	2	2	78	3		
Experience, Results and Lessons Learned from Automated Driving on Germany's Highways	1	2	2	78	3		
Multi-Objective Path Planning using Spline Represent	1	2	2	78	3		
A Study on Autonomous Vehicle Development Process at University*	0	3	3	77	3		
Road Surface Recognition Using Laser Radar for Automatic Platooning	1	2	2	78	3		
Building a Prototype for Power-Aware Automatic Parking System	1	2	2	78	3		
A Computer Vision System for Detection and Avoidance for Automotive Vehicles	1	2	2	78	3		
Path Tracking of Autonomous Ground Vehicle Based on Fractional Order PID Controller Optimized by PSO	0	3	3	77	3		
Off-road Path Following using Region Classification and G Constraints*	1	2	2	78	3		
Self-Tuning PID Controller for Autonomous Car Tracking in Urban Traffic	1	2	2	78	3		
Shared Control of Autonomous Vehicles based on Velocity Space Optimization	1	2	2	78	3		
A 13,000 km Intercontinental Trip with Driverless Vehicles:	0	3	3	77	3		
Real-Time Coordination of Autonomous Vehicles	0	3	3	77	3		
Accurate and Efficient Traffic Sign Detection Using Discrim	0	3	3	77	3		
DeepDriving: Learning Affordance for Direct Perception in	1	2	2	78	3		
A Robust Algorithm for the Detection of Vehicle Turn Signa	2	1	1	79	3		
Constrained Global Path Optimization for Articulated Steeri	2	1	1	79	3		
360° detection and tracking algorithm of both pedestrian an using fisheye images	1	2	2	78	3		
State your position	1	2	2	78	3		
A robotic platform to evaluate autonomous driving systems	1	2	2	78	3		
Coordinated control of multiple vehicles with discrete-time periodic communications	2	1	1	79	3		
Real-time Implementation of a Novel Safety Function for Pr	1	2	2	78	3		
Coordinated Path Following Control for a Group of Car-like	1	2	2	78	3		
A Combined Model- and Learning-Based Framework for In	1	2	2	78	3		
Towards a Framework for Testing Drivers' Interaction with Adopting WirelessHART for In-Vehicle-Networking	0	3	3	77	3		
Terrain Mapping for Off-road Autonomous Ground Vehicle	3	0	0	80	3		
Incremental Sampling-based Algorithm for Minimum-violation Motion Planning	0	3	3	77	3		
Incremental Sampling-based Algorithm for Minimum-violation Motion Planning	1	2	2	78	3		
Vision-based Nighttime Vehicle Detection and Range Esti	2	1	1	79	3		
Design and Comparative Analysis of a Driveless LED light	0	3	3	77	3		
Local Path Planning for Off-Road Autonomous Driving With Avoidance of Static Obstacles	2	1	1	79	3		
HOG Based Multi-object Detection for Urban Navigation	0	3	3	77	3		
Genetic Algorithm Approach for Locating Automatic Vehicl Identification Readers	0	3	3	77	3		
Reliable Intersection Protocols Using Vehicular Networks	2	1	1	79	3		
INTELLIGENT TRAFFIC WITH CONNECTED VEHICLES	1	2	2	78	3		
MCMC Particle Filter for Real-Time Visual Tracking of Vehi	0	3	3	77	3		
Globally Asymptotically Stable Filter for Navigation aided b and Depth Measurements	0	3	3	77	3		
A Real-Time Multi-Sensor Fusion Platform for Automated	1	2	2	78	3		
A full-3D Voxel-based Dynamic Obstacle Detection for Urban Scenario using Stereo Vision	1	2	2	78	3		
A Real-Time Trajectory Control of Two Driving Mobile Rob	2	1	1	79	3		
Vehicle Automation in Cooperation with V2I and Nomadic Devices Communication	2	1	1	79	3		
Automatic vehicle classification and tracking method for ve movements at signalized intersections	1	2	2	78	3		
Multi-Target Tracking using a 3D-Lidar Sensor for Autono	1	2	2	78	3		
Traffic Sign Representation using Sparse-Representations	1	2	2	78	3		
Speed Profile Optimization for Vehicles Crossing an Intersection Under a Safety Constraint	1	2	2	78	3		
Sum	46	103	103	3897			
	true positive	false negative	false positive	true negative	total number of topics		
	Precision =	0,309					
	Recall =	0,309					
	F =	0,309					

title	true positive	false negative	false positive	true negative	total number of topics		40 skröp	60 bra topics
A_B_P_C_S_D_M_C_	1	2	14	43	15			
A_B_S_D_M_P_P_P_	2	1	9	48	11			
A_B-N_C_Grand_2012_	3	0	11	46	14			
A_C_C_D_Gillet_2014_S	3	0	12	45	15			
A_C_L_N_S_M_M_N_	0	3	8	49	8			
B_B_H_Giese_2008_Incr	1	0	15	44	16			
B_W_A_K_M_P_T_A_	0	2	16	42	16			
B-M_S_Chung_Jin-Woo;	0	3	12	45	12			
C_C_J_Liu_2010_A_Rein	3	0	13	44	16			
C_C_Y_H_F_G_C_B_	2	1	12	45	14			
C_L_B_N_T_M_C_S_	2	1	14	43	16			
C_W_Axelrod_2015_Enfor	0	1	16	43	16			
D_B_W_M_I_Posner_20	3	0	13	44	16			
D_C_S_D_B_P_Stone_	3	0	8	49	11			
G_A_J_I_N_E_M_Neb	2	1	13	44	15			
H_x_E_C_ne;T_Sattler;	2	1	14	43	16			
J_A_C_S_A_Pascoal_2	3	0	10	47	13			
J_C_S_U_B_L_M_Mau	3	0	11	46	14			
J_S_B_P_H_H_Chen_2	3	0	12	45	15			
K_B_H_M_A_Zell_2012	1	1	9	49	10			
K_C_F_J_T_R_S_J_B	2	1	14	43	16			
L_C_A_F_L_Pallottino_2	1	1	15	43	16			
L_x_F_J_M_Alvarez;F_	2	1	10	47	12			
M_A_A_R_M_J_M_Ekl	2	1	9	48	11			
M_A_J_M_Dolan_2011_	1	2	11	46	12			
M_A_P_F_C_O_J_Sjo	1	2	13	44	14			
M_A-M_W_S_M_Y_W	1	2	9	48	10			
M_B_C_H_A_L_M_A_	2	1	15	43	16			
M_B_Z_G_P_Z_M_B_	0	3	11	46	11			
M_C_D_P_M_Pasquier_	0	3	15	42	15			
M_H_Ang_2015_Achievin	0	0	14	46	14			
M_J_B_C_M_Veth_201	1	1	11	47	12			
M_J_H_Berg;R_Olsson;	1	2	10	47	11			
M_x_E_A_yr;x00E;;M_	0	1	5	54	5			
N_C-B_A_M_J_R_M_	3	0	13	44	16			
N_T_Atsumuro_Yamaguchi	2	1	10	47	12			
P_B_D_K_C_B_J_Dick	3	0	13	44	16			
P_V_K_B_S_Vidas_201	2	1	14	43	16			
P_V_M_E_O_J_R_d_	2	1	11	46	13			
Q_B_M_P_C_Laugier_2	1	2	5	52	6			
S_A_G_B_R_R_P_Mu	2	1	11	46	13			
S_A_S_C_Y_S_Alj_201	2	1	9	48	11			
S_B_M_M-P_R_M-P_	2	1	14	43	16			
S_D_B_B_E_A_Speran	1	2	15	42	16			
S_J_A_S_B_K_K_I_J	1	2	6	51	7			
S_P_B_R_W_Sadowski	1	2	15	42	16			
T_A_M_M_M_Ali_2015_	2	1	14	43	16			
Y_A_P_P_F_P_A_Burr	2	1	14	43	16			
Z_B_J_J_N_Y_S_Linc	3	0	10	47	13			
Z_K_x_E_T_Akg;x00Fc,	3	0	13	44	16			
Sum	83	54	591	2273	673			
	true positive	false negative	false positive	true negative	total number of topics			
	Precision =	0,123						
	Recall =	0,606						
	F =	0,2044938272						

title	true positive	false negative	false positive	true negative	total number of topics		40 skr�p	60 bra topics
A_B_P_C_S_D_M_C	0	3	3	54	3			
A_B_S_D_M_P_P_P	1	2	2	55	3			
A_B-N_C_Grand_2012	1	2	2	55	3			
A_C_C_D_Gillet_2014	0	3	3	54	3			
A_C_L_N_S_M_M_N	1	2	2	55	3			
B_B_H_Giese_2008_Inc	0	3	3	54	3			
B_W_A_K_M_P_T_A	1	2	2	55	3			
B-M_S_Chung_Jin-Woo	1	2	2	55	3			
C_C_J_Liu_2010_A_Rei	2	1	1	56	3			
C_C_Y_H_F_G_C_B	1	2	2	55	3			
C_L_B_N_T_M_C_S	0	3	3	54	3			
C_W_Axelrod_2015_Enf	0	3	3	54	3			
D_B_W_M_I_Posner_2	1	2	2	55	3			
D_C_S_D_B_P_Stone	0	3	3	54	3			
G_A_J_I_N_E_M_Ne	1	2	2	55	3			
H_x_E_C_ne;T_Sattle	2	1	1	56	3			
J_A_C_S_A_Pascoal	0	3	3	54	3			
J_C_S_U_B_L_M_Ma	3	0	0	57	3			
J_S_B_P_H_H_Chen	0	3	3	54	3			
K_B_H_M_A_Zell_201	1	2	2	55	3			
K_C_F_J_T_R_S_J	1	2	2	55	3			
L_C_A_F_L_Pallottino	1	2	2	55	3			
L_x_F_J_M_Alvarez;F	1	2	2	55	3			
M_A_A_R_M_J_M_E	0	3	3	54	3			
M_A_J_M_Dolan_2011	0	3	3	54	3			
M_A_P_F_C_O_J_Sj	0	3	3	54	3			
M_A-M_W_S_M_Y	1	2	2	55	3			
M_B_C_H_A_L_M_A	0	3	3	54	3			
M_B_Z_G_P_Z_M_B	0	3	3	54	3			
M_C_D_P_M_Pasquier	0	3	3	54	3			
M_H_Ang_2015_Achievi	1	2	2	55	3			
M_J_B_C_M_Veth_20	1	2	2	55	3			
M_J_H_Berg;R_Olsson;	0	3	3	54	3			
M_x_E_A_yr;x00E;M	2	1	1	56	3			
N_C-B_A_M_J_R_M	2	1	1	56	3			
N_T_Atsumuro_Yamaguc	0	3	3	54	3			
P_B_D_K_C_B_J_Dic	1	2	2	55	3			
P_V_K_B_S_Vidas_20	1	2	2	55	3			
P_V_M_E_O_J_R_d	0	3	3	54	3			
Q_B_M_P_C_Laugier	0	3	3	54	3			
S_A_G_B_R_R_P_M	1	2	2	55	3			
S_A_S_C_Y_S_Alj_20	1	2	2	55	3			
S_B_M_M-P_R_M-P	0	3	3	54	3			
S_D_B_B_E_A_Spera	0	3	3	54	3			
S_J_A_S_B_K_K_I	1	2	2	55	3			
S_P_B_R_W_Sadowsk	0	3	3	54	3			
T_A_M_M_M_Ali_2015	1	2	2	55	3			
Y_A_P_P_F_P_A_Bu	0	3	3	54	3			
Z_B_J_J_N_Y_S_Lin	2	1	1	56	3			
Z_K_x_E_T_Akg;x00F	1	2	2	55	3			
SUM:	35	115	115	2735				
	true positive	false negative	false positive	true negative	total number of topics			
	Precision =	0,233						
	Recall =	0,233						
	F =	0,233						

Fisheye optics f	96, 70, 99			(11, 0.1390023	, (32, 0.084519	(99, 0.10121410958494333)		
Data age based control data exc	86, 44			(38, 0.0680274	(44, 0.2029934	(86, 0.073215505908526407),		
Obstacle Avoid Model Predictiv	78, 82, 57			(12, 0.0740088	(52, 0.1608838	(62, 0.1970401075172302)		
Intelligent Cruis Stop and Go wit	52, 55, 44			(42, 0.1047802	(44, 0.2432348	(82, 0.11078935942794398)		
Autonomous Na	87			(48, 0.0769147	(82, 0.0935574	(98, 0.10503959543129056)		
Bayesian Netwo	95, number 4, 5			(38, 0.1212474	(70, 0.0789667	(82, 0.227441332992600527)		
Experience, Re Lessons Learne Automated Drivi Germany's High	37, 55, 62			(55, 0.0847935	(82, 0.1893622	(98, 0.10634841700615484)		
Multi-Objective	62 ,5, 57			(5, 0.10523515343964221), (62, 0.1230114569017088), (84, 0.095175123030142555)				
A Study on Auto University*	No relevant topic			(34, 0.0832615	(74, 0.1271243	(82, 0.22852976265870359),		
Road Surface R Automatic Plato	35, 18, 54			(25, 0.1109076	(82, 0.0807853	, (94, 0.14401170820600007),		
Building a Proto Parking System	47, 62			(47, 0.1143939	(66, 0.1301513	(82, 0.096545128076018782),		
A Computer Visi Avoidance for A	5, 2, 99			(82, 0.2103425	(94, 0.1302130	(99, 0.1686118185386129)		
Path Tracking o Fractional Order	62, vehicle control			(16, 0.1431923	(23, 0.0979858	(89, 0.064092231910171715),		
Off-road Path F Constraints*	48, 28, 63			(18, 0.1338110	(82, 0.1087812	(94, 0.11068233022525438),		
Self-Tuning PID Autonomous Ca	37, 95, Vehicle control			(24, 0.1157130	(52, 0.1390555	(60, 0.10808760689065738),		
Shared Control based on Veloci	45, vehicle control, 78			(17, 0.0792136	(52, 0.0732116	(62, 0.23565339508091404),		
A 13,000 km Int	95, vehicle control, 82			(66, 0.0681289	(74, 0.4081724	(82, 0.18704383461036819)		
Real-Time Coor	35, 86			(32, 0.0783232	(44, 0.1282214	(55, 0.085426293809304457)		
Accurate and Ef	96 ,85, 2			(18, 0.1082158	(55, 0.0854262	(93, 0.082951350685758013)		
DeepDriving: Le	77, 88, 2			(18, 0.1216196	(66, 0.1038493	(82, 0.18960915856709309)		
A Robust Algorit	33, 88			(12, 0.0261620	(33, 0.8143677	(95, 0.020885043865360702)		
Constrained Gio	62, 78, 10			(10, 0.1476177	(57, 0.0813935	(62, 0.16430836004450927)		

360° detection a using fisheye Im	47, 96, 99				(82, 0.1103107	(96, 0.0795476	(99, 0.16454368461705091)		
State your positi	1, 48,				(1, 0.25972098	(15, 0.1332502	(52, 0.077168595128774123)		
A robotic plattform	Vehicle control, 55, 60				(11, 0.1056438	(55, 0.0618856	(66, 0.081648855621690927)		
Coordinated co discrete-time pe	77, 82, 86				(16, 0.2589560	(17, 0.0772982	(62, 0.065248220945653648)		
Real-time Imple	Vehicle control, 24, 25				(11, 0.0762797	(25, 0.1102223	(52, 0.27489935471687632)		
Coordinated Pat	82, 86, vehicle control				(8, 0.52763602	(16, 0.0910834	(62, 0.075266478492064609)		
A Combined Mo	35, 92, vehicle control				(35, 0.0605185	(62, 0.1552549	(76, 0.16840433276106981)		
Towards a Fra	55, 45				(49, 0.0351974	(55, 0.5331159	(82, 0.110881759229887201)		
Adopting Wirele	86, 61				(66, 0.1851605	(74, 0.0803295	(82, 0.086750851051705574)		
Terrain Mappin	96, 99 18,				(18, 0.1373032	(96, 0.0928738	(99, 0.097444633362412547)		
Incremental Sa Minimum-violati	60, Vehicle control, 78				(15, 0.1016502	(56, 0.1132531	(62, 0.12246169949335441)		
Vision-based Ni	51, 96, 55				(12, 0.0442678	(33, 0.2048777	(82, 0.18723811097421242)		
Design and Co	0, 33, 9				(29, 0.0346098	(37, 0.6986309	(55, 0.19485987878501795)		
Local Path Plan Driving With Av	62, 78, 99,				(10, 0.0552967	(35, 0.0468949	(78, 0.054222311030339045)		
HOG Based Mu	98, 51, 87				(18, 0.1317220	(82, 0.1025475	(94, 0.13515502755201253),		
Genetic Algorith Identification Re	36				(15, 0.1440173	(18, 0.1953752	(27, 0.10887306087309488)		
Reliable Interse	38, 44, 42				(38, 0.3940938	(42, 0.1518469	(76, 0.076593369408331127)		
INTELLIGENT VEHICLES	44, 49, 20,				(44, 0.1111465	(66, 0.0837658	(82, 0.20864493379938656)		
MCMC Particle	33, 88, 57				(12, 0.0842157	(18, 0.1213754	(98, 0.083557085210943058)		
Globally Asympt and Depth Mea	91, 87, 61				(15, 0.0640450	(16, 0.1830467	(70, 0.06206132784724458)		
A Real-Time Mu	37, 71, 70				(37, 0.8187293	(55, 0.0787838	(78, 0.034186934006349756)		
A full-3D Voxel-Urban Scenario	99, 98, 28				(11, 0.1127621	(12, 0.2129591	(96, 0.10500542994600785)		
A Real-Time Tr	vehicle control, 78, 99				(11, 0.1131325	(52, 0.1852250	(87, 0.1906831953200861)		
Vehicle Automa Nomadic Devic	55,99,76				(11, 0.0868970	(55, 0.0991782	(82, 0.21023791869611239)		
Automatic vehic movements at s	8, 42, 54				(8, 0.07249082	(82, 0.1718510	(94, 0.071540369671543758)		

Multi-Target Tra	15, 77, 98			(6, 0.09598793	(52, 0.0773815	(98, 0.27334997661984639)		
Traffic Sign Rep	2, number 20			(2, 0.10946451	(18, 0.1408045	(63, 0.1056367450161169)		
Speed Profile O Intersection Un	38, 42, vehicle control			(38, 0.1000932	(59, 0.0869065	(62, 0.16169786391830407)		

Fisheye optics f	96, 70, 99			(11, 0.13900239476104639)		
Data age based control data exc	86, 44			(44, 0.20299343908608297)		
Obstacle Avoid Model Predictiv	78, 82, 57				(62, 0.1970401075172302)	
Intelligent Cruis Stop and Go wit	52, 55, 44			(44, 0.24323481593723542),		
Autonomous Na	87				(98, 0.10503959543129056)	
Bayesian Netwo	95, number 4, 5				(82, 0.22744132992600527)	
Experience, Re Lessons Learne Automated Drivi Germany's High	37, 55, 62			(82, 0.18936225101487586)		
Multi-Objective	62, 5, 57			(62, 0.1230114569017088)		
A Study on Auto University*	No relevant topi				(82, 0.22852976265870359),	
Road Surface R Automatic Plato	35, 18, 54				, (94, 0.14401170820600007),	
Building a Proto Parking System	47, 62			(66, 0.13015139933796271)		
A Computer Visi Avoidance for A	5, 2, 99			(82, 0.21034250761433265),		
Path Tracking o Fractional Order	62, vehicle cont			(16, 0.14319230305794198),		
Off-road Path F Constraints*	48, 28, 63			(18, 0.13381102708044132)		
Self-Tuning PID Autonomous Ca	37, 95, Vehicle			(52, 0.13905556918544504)		
Shared Control based on Veloci	45, vehicle cont				(62, 0.23565339508091404),	
A 13,000 km Int	95, vehicle cont				(74, 0.40817247415538177)	
Real-Time Coor	35, 86			(44, 0.12822145200088794)		
Accurate and Ef	96, 85, 2			(18, 0.10821584373327547)		
DeepDriving: Le	77, 88, 2				(82, 0.18960915856709309)	
A Robust Algorit	33, 88			(33, 0.81436774856870142)		

Constrained Glo	62, 78, 10					(62, 0.16430836004450927)
360° detection a using fisheye im	47, 96, 99					(99, 0.16454368461705091)
State your positi	1, 48,		(1, 0.25972098026667584)			
A robotic platfor	Vehicle control,		(11, 0.10564387259839503)			
Coordinated co discrete-time pe	77, 82, 86		(16, 0.25895607465347326)			
Real-time Imple	Vehicle control,					(52, 0.27489935471687632)
Coordinated Pat	82, 86, vehicle		(8, 0.52763602635487761)			
A Combined Mo	35, 92, vehicle				(76, 0.16840433276106981)	
Towards a Fra	55, 45			(55, 0.53311597367673769)		
Adopting Wirele	86, 61		(66, 0.18516054221053074)			
Terrain Mappin	96, 99 18,		(18, 0.13730328783108575)			
Incremental Sa Minimum-violati	60, Vehicle cont					(62, 0.12246169949335441)
Vision-based Ni	51, 96, 55			(33, 0.20487778194549938)		
Design and Co	0, 33, 9			(37, 0.69863093814294153)		
Local Path Plan Driving With Av	62, 78, 99,		(10, 0.055296772486137021)			
HOG Based Mu	98, 51, 87					(94, 0.13515502755201253),
Genetic Algorith Identification Re	36			(18, 0.19537523540597748),		
Reliable Interse	38, 44, 42		(38, 0.39409381753496925)			
INTELLIGENT VEHICLES	44, 49, 20,					(82, 0.20864493379938656)
MCMC Particle	33, 88, 57			(18, 0.12137547648943883)		
Globally Asympt and Depth Mea	91, 87, 61			(16, 0.18304670757585845)		
A Real-Time Mu	37, 71, 70		(37, 0.81872937773255205)			
A full-3D Voxel-Urban Scenario	99, 98, 28			(12, 0.2129591221670441)		
A Real-Time Tr	vehicle control,					(87, 0.1906831953200861)

clean_C_C_J.	(64, 0.3030772	(61, 0.2103519	(23, 0.091408064252341062)	(36, 0.18485644806255622)	27	24	46
clean_G_A_J.	(36, 0.2382778	(37, 0.1099536	(94, 0.091980057358372974)	(72, 0.255996012169338933)	autonomous ve	3	37
clean_Z_B_J.	(54, 0.1524063	(91, 0.1294819	(37, 0.109381222093028534)	(2, 0.2275136900047077)	autonomous ve	8	18
clean_M_A_P.	(22, 0.2103915	(37, 0.2065201	(14, 0.08843138212265389)	(8, 0.18093794262797613)	8	81	10
clean_S_B_M.	(35, 0.1151748	(14, 0.0856002	(6, 0.0816938863744275481)	(8, 0.5258245421157947)	autonomous ve	14	24
clean_D_C_S.	(0, 0.28780847	(14, 0.1804275	(67, 0.16098908890118968)	(95, 0.23475704115663953)	95		
clean_A_B_N.	(2, 0.22751369	(8, 0.13359666	(98, 0.114482204054275112)	(2, 0.11209784927366824)	62	31	
clean_D_B_W.	(64, 0.1208752	(10, 0.0981494	(35, 0.085743075471516106)	(37, 0.17786710077087728)	3	2	37
clean_T_A_M.	(86, 0.1014034	(67, 0.0901782	(15, 0.08420761338234857)	(64, 0.30307722290019357)	25	90	46
clean_K_C_F.	(14, 0.1592115	(54, 0.1507850	(32, 0.074856650682069967)	(36, 0.15072481588128858)	49	2	56
clean_J_C_S.	(14, 0.1694132	(16, 0.1556911	(61, 0.13212397282221669)	(97, 0.18982659068292873)	2	97	46
clean_M_J_B.	(56, 0.1906900	(6, 0.17118230	(2, 0.14453992095615442)	(31, 0.30094220805326038)	ABOUT INTERNET OF THINGS		39
clean_L_X_F.	(91, 0.1982367	(64, 0.1295936	(61, 0.12567340234222596)	(64, 0.12087527306715326)	32	46	91
clean_Z_K_x.	(14, 0.1791856	(76, 0.1176699	(2, 0.1108253691003747)	(0, 0.28780847536848242)	94	autonomous ve	27
clean_J_A_C.	(8, 0.16329881	(0, 0.15599647	(34, 0.11921252479198195)	(36, 0.23827786532158918)	3	85	autonomous vehicle control
clean_J_S_B.	(86, 0.1676208	(34, 0.1203641	(62, 0.098210500234343259)	(64, 0.1820666842132069)	16	2	23
clean_A_B_S.	(72, 0.2559601	(61, 0.1129838	(76, 0.10728386724429928)	(8, 0.1632988121483552)	31	autonomous ve	62
clean_M_B_Z.	(37, 0.1686570	(67, 0.1659653	(86, 0.13212604834057004)	(14, 0.16941322421587424)	32	58	68
clean_M_x_E.	(76, 0.6515046	(64, 0.1347841	(80, 0.11052799511617664)	(86, 0.16762081225685336)	86	31	62
clean_S_A_S.	(38, 0.2684245	(46, 0.1710797	(72, 0.12607618268822365)	(14, 0.16798481714197125)	autonomous ve	ABOUT ROBOT	39
clean_M_J_H.	(61, 0.3349081	(8, 0.24586743	(3, 0.10115871457693873)	(14, 0.15921153738421318)	54	70	46
clean_S_J_A.	(78, 0.8130703	(2, 0.04279595	(8, 0.041522056163905016)	(3, 0.20696931793586407)	78	34	autonomous vehicle control
clean_K_B_H.	(14, 0.1679848	(16, 0.1519493	(2, 0.12432099890414264)	(91, 0.19823679216460691)	64	46	54
clean_P_V_K.	(46, 0.1449212	(31, 0.0869137	(6, 0.081995218719289387)	(8, 0.18235012515513513)	10	autonomous ve	14
clean_B_B_H.	(95, 0.2347570	(85, 0.1075159	(36, 0.10197158258058026)	(36, 0.20043286904297364)	16	81	autonomous vehicle control
clean_P_B_D.	(64, 0.1508724	(3, 0.12064299	(16, 0.11892074962130053)	(22, 0.21039159482527384)	81	autonomous ve	10
clean_M_A_J.	(36, 0.2004328	(85, 0.1446369	(3, 0.13290219339100076)	(67, 0.32417530572754122)	29	51	94
clean_A_C_L.	(8, 0.52582454	(35, 0.1053866	(83, 0.099992708227210195)	(78, 0.22281829026120953)	36	autonomous ve	3
clean_S_P_B.	(61, 0.1677124	(16, 0.1457492	(42, 0.089127651321651818)	(37, 0.16865701426790394)	3	81	24
clean_M_H_A.	(81, 0.1769540	(2, 0.16601571	(91, 0.11885317918223637)	(37, 0.19873329096901887)	51	81	3
clean_P_V_M.	(8, 0.32051546	(6, 0.14205713	(3, 0.13477592829968937)	(81, 0.17695400107396794)	Very lightweight paper, should produce low probabilities		
clean_S_D_B.	(10, 0.1479399	(86, 0.1079217	(14, 0.08591718688928264)	(56, 0.19069006158388141)	2	46	
clean_N_T_At	(81, 0.5448869	(20, 0.1467503	(38, 0.066431417501460852)	(61, 0.33490816986681493)	autonomous ve	14	3
clean_S_A_G.	(67, 0.4859769	(37, 0.0879113	(94, 0.052500121435727472)	(76, 0.66150463233416078)	46		
clean_C_C_Y.	(36, 0.1507248	(2, 0.13043038	(56, 0.12716253801145527)	(32, 0.16486208522831469)	97	68	46
clean_A_C_C.	(8, 0.18093794	(36, 0.1405603	(16, 0.11095466067959121)	(81, 0.54488699580868405)	20	38	24
clean_C_L_B.	(97, 0.1898263	(64, 0.1845176	(78, 0.097066065651726055)	(64, 0.1508724259221636)	17	98	42
clean_M_B_C.	(78, 0.2228182	(36, 0.1871256	(37, 0.10862600579458387)	(46, 0.14492128773661103)	46	42	70
clean_B-M_S.	(37, 0.1778671	(98, 0.1415837	(95, 0.11439256411777463)	(8, 0.32051546792194785)	autonomous ve	8	14

clean_H_x_E.	(64, 0.1820686	(23, 0.1043574	(62, 0.10328175892908135)			(64, 0.3092776500292902)						68	98	2	
clean_B_w_A.	(2, 0.11209784	(0, 0.10982743	(78, 0.10974667443858584)			(67, 0.48597694012599374)						31	94	72	
clean_M_C_D.	(37, 0.1987332	(67, 0.1021661	(61, 0.092271868290538586)			(38, 0.2684245399469421)						38	autonomous ve	16	
clean_L_C_A.	(3, 0.20696931	(78, 0.1675713	(52, 0.10858727166328266)			(35, 0.11517488141378281)						31	autonomous ve	34	
clean_Q_B_M.	(64, 0.3092776	(14, 0.2421471	(94, 0.16943680061267874)			(10, 0.14793998316307688)						16	97	42	
clean_A_B_P.	(36, 0.1848564	(29, 0.1298889	(76, 0.11872437549516186)			(78, 0.81307037047720154)		autonomous ve				autonomous ve	81	78	
clean_M_A-M.	(67, 0.3241753	(86, 0.2248693	(31, 0.12382616448151035)			(61, 0.16771243527230448)						46	70	autonomous vehicle control	
clean_M_A_A.	(8, 0.18235012	(78, 0.1422352	(14, 0.10409280196304362)			(86, 0.10140347723875423)						86	autonomous ve	62	
clean_C_W_A	(31, 0.3009422	(62, 0.0918108	(0, 0.080914136703779838)			(64, 0.1485523219220494)						91	32		
clean_N_C-B.	(32, 0.1648620	(97, 0.1284760	(46, 0.1179398406423926)			(54, 0.15240639953675808)						46	54	91	
clean_Y_A_P.	(64, 0.1485523	(78, 0.1357089	(98, 0.075125971005098865)			(14, 0.17918560023085695)						3	37	autonomous vehicle control	
												13 st			
												37 st			

image	Word #1	Word #2	Word #3	Word #4	Word #5	Word #6	Word #7	Label	Acronyms/ word explanations
0	agent	coil	power	circuit	policy	aorta	turn	Power management	
1	uncertainty	parameter	estimation	covariance	method	matrix	system	Probability estimation	
2	image	recognition	sample	vehicle	model	mask	logo	Image recognition	
3	oscillation	natural	system	vo	velocity	locomotion	value		
4	vehicle	speed	control	lateral	reference	profile	strategy	Vehicle Control	
5	obstacle	vehicle	task	avoidance	control	path	velocity	Obstacle avoidance management	
6	vehicle	utility	target	feature	eye	assignment	based	Vehicle Utility	
7	ve	olarak	filtre	ign	br	ekbho	bu		
8	vehicle	task	method	control	background	detection	video	Vehicle detection (video feed?)	
9	sensor	market	imu	system	cost	fusion	data	IMU market	imu = inertial measurement unit, measures crafts velocity orientation and gravitational forces
10	path	curve	vehicle	bézier	point			Path prediction	
11	system	vehicle	image	obstacle	used	decision	detection	Obstacle detection	
12	vehicle	set	time	approach	algorithm	system		Vehicle software	
13	arduous	narikiyo	centred	parametrize	hiriko	emphified	pean		
14	rate	ber	vocoder	channel	amr	ecall	mode		
15	problem	bound	path	constraint	solution	state	approach	Pathfinding	
16	control	vehicle	tracking	trajectory	controller	following	time	Vehicle Control	
17	vehicle	set	trajectory	reachable	constraint	tk	system	Vehicle limitations	
18	road	model	color	algorithm	point	method	image	Road modelling	
19	arduous	narikiyo	centred	parametrize	hiriko	emphified	pean		
20	image	character	plate	network	neural	value	recognition	Image input to machine learning	
21	arduous	narikiyo	centred	parametrize	hiriko	emphified	pean		
22	arduous	narikiyo	centred	parametrize	hiriko	emphified	pean		
23	particle	measurement	fastslam	filter	problem	position	based	position measur	fastslam is an algorithm for localising a robot and mapping it's surroundings
24	vehicle	threat	figure	position	landmark	avg	time	Vehicle threat assesment	
25	vehicle	control	wheel	tire	friction	road	longitudinal	Wheel control	
26	arduous	narikiyo	centred	parametrize	hiriko	emphified	pean		
27	attack	vehicle	risk	severity	problem	car	number	Security aspects	
28	cost	disparity	cloud	time	census	pixel	weight	Image processi	disparity could be related to image recognition
29	agent	ri	source	signal	ci	formation	algorithm		
30	process	group	request	pi	node	quorum	message	Obstacle management	
31	image	obstacle	top	used	system	view	path	Obstacle management	
32	system	data	obstacle	tentacle	mobility	information	vehicle	Obstacle management	
33	light	vehicle	detection	tracking	algorithm	frame	signal	Vehicle detection	
34	control	brake	system	vehicle	design	stopping	model	Brake control/management	
35	vehicle	state	time	model	fleet	based	road	Model for vehicle fleets	
36	research	technology	curve	patent	stage	vehicle		Vehicle research	
37	vehicle	component	driving	architecture	platform	system	control	System architecture	
38	vehicle	intersection	figure	type	lane	system	output	Intersection handling	
39	av	ri	leader	time	follower	one			
40	arduous	narikiyo	centred	parametrize	hiriko	emphified	pean		
41	tag	reader	antenna	function	rra	position	positioning		tag might be related to gps technology
42	intersection	vehicle	traffic	time	group	car	light	Intersection handling	
43	arduous	narikiyo	centred	parametrize	hiriko	emphified	pean		
44	vehicle	communication	cooperative	attack	cacc	stream		Vehicle coopera	cacc could be correlated active clause coverage, A Logic Coverage Criterion from Software Testing
45	trajectory	passenger	based	test	boarding	alighting	tracking	System test with	alighting might refer to getting off a vehicle
46	demand	policy	vehicle	condition	tmhp	time	stability		
47	parking	vehicle	space	pedestrian	task	state	system	Parking	
48	image	based	figure	time	navigation	algorithm	environment	Image-based navigation	
49	car	vehicle	algorithm	caravan	time	aid	three	Vehicle cooperation	
50	information	platform	service	self	driving	tourist	content	Information service for tourists?	
51	vehicle	detection	cluster	algorithm	ve	time	FALSE	Vehicle detection?	
52	vehicle	control	angle	model	lateral	dynamic	system	Vehicle Control	
53	arduous	narikiyo	centred	parametrize	hiriko	emphified	pean		
54	task	vehicle	tracking	target	model	platoon	measurement	Vehicle tracking (platoon implies several vehicles)	
55	driving	driver	automated	vehicle	wa	speed	distance	Driving	
56	graph	pattern	rule	system	time	set	model	Data representation	
57	trajectory	vehicle	trim	point	using	time	maneuver	Maneuvering	
58	task	cbba	bundle	assignment	agent	bid	ij		
59	agent	task	node	algorithm	cbba	assignment	time		
60	system	driving	change	vehicle	development	software	simulation	Software development for vehicles in simulation	
61	vehicle	prt	system	wheel	consumption	fuel	campus	Fuel Consumption	
62	vehicle	path	planning	constraint	trajectory	problem	time	Path planning (Planera körning?)	
63	road	image	network	feature	learning	pixel	segmentation	Image recognition for roads	
64	oscillation	vo	system	natural	velocity	locomotion	control	??	
65	ve	ign	filtre	olarak	ekbho	bir	bu		
66	car	system	model	control	driven	ha	vehicle	System for vehicle control	
67	fault	bg	system	element	set	model			
68	arduous	narikiyo	centred	parametrize	hiriko	emphified	pean		
69	frequency	stimulus	signal	autonomous	car	ss		Signal interpretation from sensors?	
70	camera	state	imu	estimate	vehicle	model	behavior	Hardware beha	(IMU is a chip that works along with cameras)
71	system	sensor	vehicle	control	team	rascal	figure	Hardware relation	
72	arduous	narikiyo	centred	parametrize	hiriko	emphified	pean		
73	signal	turbulence	sequence	time	set	system	coherence	Signal data behavior	
74	vehicle	system	data	wa	sensor	test	gps	Vehicle localization/vehicle navigation	
75	arduous	narikiyo	centred	parametrize	hiriko	emphified	pean		
76	vehicle	road	time	future	position	forwarding	driving	Maneuver planning (Planera körning?)	
77	road	lane	semantic	node	map	based	image	Lane following	
78	vehicle	trajectory	obstacle	car	driving	self	road	Obstacle avoidance	
79	driving	event	feature	data	driver	speed	classification	Driving data handling	
80	plate	image	character	license	correction	value	neural	Image recognition of license plates	
81	arduous	narikiyo	centred	parametrize	hiriko	emphified	pean		
82	vehicle	system	lane	line	detection	sensor	wa	Lane following	
83	potential	control	vehicle	field	sin	co		?	
84	segment	point	track	tolerance	curvature	curve	arc	Curve assessment	
85	video	tracking	frame	system	tdt	sequence	proposed	Video tracking	
86	uxv	node	flow	routing	network	packet	path	Network	
87	robot	mobile	control	navigation	environment	obstacle	motion	Navigation	
88	camera	image	vehicle	feature	map	track	method	Vehicle tracking through image processing	
89	car	driver	control	system	track	program	model	Vehicle Control	
90	prt	vehicle	fuel	system	consumption	campus	amd	Fuel Consumption	
91	observer	fx	fy	longitudinal	sensor	sliding	fault	Navigation failure management	
92	probability	input	traffic	markov	chain	state	participant	Traffic predictio	markov = probability mathematician
93	traffic	light	prior	detection	image	location	score	Traffic Light detection	
94	road	image	system	edge	method	detection	algorithm	Lane detection	
95	follower	sensor	velocity	leader	vehicle	test	heading	Follow other vehicle	
96	image	camera	point	frame	estimation	motion	method	Image processing	
97	image	visual	database	localization	feature	solution	infrared	Image processing	

98	object	grid	moving	detection	laser	motion	data	Object detection					
99	obstacle	camera	depth	image	car	map	plane	Obstacle detection					

0	system	vehicle	control	driver	controller	traffic	autonomous	autonomous vehicle c	40, 76			
1	re	maximum	cubic	interpolation	manoeuvre	phase	parametrized					
2	vehicle	system	autonomous	control	intelligent	path	navigation	autonomous vehicle navigation				
3	vehicle	model	driving	road	autonomous	driver	based	autonomous driving				
4	re	maximum	cubic	interpolation	manoeuvre	phase	parametrized					
5	re	maximum	cubic	interpolation	manoeuvre	phase	parametrized					
6	measurement	vehicle	inertial	noise	path	controller	ground					
7	re	maximum	cubic	interpolation	manoeuvre	phase	parametrized					
8	control	vehicle	lateral	speed	autonomous	cruise	based	autonomous vehicle speed control		cruise lateral och speed refererar till speed		
9	re	maximum	cubic	interpolation	manoeuvre	phase	parametrized					
10	vehicle	obstacle	uncertainty	probabilistic	planning	electric	linear	vehicle obstacle avoidance				
11	re	maximum	cubic	interpolation	manoeuvre	phase	parametrized					
12	re	maximum	cubic	interpolation	manoeuvre	phase	parametrized					
13	re	maximum	cubic	interpolation	manoeuvre	phase	parametrized					
14	vehicle	autonomous	system	simulation	sensor	real	time	autonomous vehicle simulation				
15	vehicle	dc	lane	battery	system	converter	control	vehicle power management				
16	vehicle	trajectory	path	optimization	control	planning	method	Vehicle path planning	Samma som 78			
17	radar	automotive	sige	bicmos	technology	packaging	ghz	Radar technology				
18	sliding	mode	vehicle	skid	control	observer	force	Sliding mode				
19	vision	robot	road	vehicle	obstacle	dynamic	static					
20	vehicle	interface	mobile	human	user	automotive	interaction	Human-vehicle interface/interaction				
21	re	maximum	cubic	interpolation	manoeuvre	phase	parametrized					
22	vehicle	system	autonomous	data	tracking	tire	signal					
23	vehicle	detection	obstacle	algorithm	system	camera	autonomous	obstacle detection				
24	pedestrian	feature	behavior	traffic	road	estimation	relevance	traffic behavior				
25	road	cue	detection	level	method	low	vision	road detection	samma som 64			
26	re	maximum	cubic	interpolation	manoeuvre	phase	parametrized					
27	road	software	challenge	robot	urban	traffic	system	urban traffic software				
28	re	maximum	cubic	interpolation	manoeuvre	phase	parametrized					
29	system	lighting	autonomous	vehicle	car	intelligent	led	vehicle lights/lighting				
30	re	maximum	cubic	interpolation	manoeuvre	phase	parametrized					
31	vehicle	sensor	system	fleet	network	automated	management	Networked vehicles				
32	object	detection	data	vehicle	sensor	fusion	classification	Object detection & classification				
33	re	maximum	cubic	interpolation	manoeuvre	phase	parametrized					
34	vehicle	tracking	distributed	network	system	topology	dynamic	Networked vehicle mapping/tracking				
35	vehicle	automated	parking	human	autonomous	system	detection	Automated parking				
36	road	maneuver	autonomous	vehicle	driving	prediction	model	Maneuver prediction				
37	control	system	driving	vehicle	road	gps	driver	Driving				
38	car	system	parking	automated	android	existing	human		35			
39	cell	sram	write	tfet	characteristic	circuit	noise	Hardware				
40	control	system	vehicle	autonomous	integrator	robot	dynamic		0			
41	software	component	algorithm	robotic	advanced	robot	system	robot software				
42	sensor	system	market	calibration	autonomous	fusion	parameter	Sensor system				
43	control	warehouse	system	vehicle	net	petri	generalized	vehicle-warehouse system		petri net used for graphical modeling of formal system		
44	transport	autonomous	vehicle	model	agent	based	future	Autonomous transport				
45	re	maximum	cubic	interpolation	manoeuvre	phase	parametrized					
46	camera	vehicle	image	road	vision	system	autonomous	Image processing for autonomous vehicles				
47	re	maximum	cubic	interpolation	manoeuvre	phase	parametrized					
48	re	maximum	cubic	interpolation	manoeuvre	phase	parametrized					
49	system	vehicle	autonomous	forest	control	terrain	ground	Off road vehicle control				
50	tracking	driving	autonomous	activity	driver	recognition	classification					
51	neural	network	system	artificial	braking	car	labview	artificial intelligence				
52	robot	vehicle	autonomous	state	position	control	industrial	?				
53	re	maximum	cubic	interpolation	manoeuvre	phase	parametrized					
54	vehicle	video	detection	tracking	traffic	automatic	method	Vehicle detection/tracking	lik 58, 66			
55	vehicle	skew	plate	system	recognition	number	correction	Vehicle number plate				
56	semantic	autonomous	mapping	bridge	scale	large	map	Map/mapping				
57	re	maximum	cubic	interpolation	manoeuvre	phase	parametrized					
58	vehicle	obstacle	lane	tracking	system	camera	vision	vehicle/obstacle tracking	lik 54, 66			
59	locomotion	system	natural	oscillation	matrix	damping	body					
60	re	maximum	cubic	interpolation	manoeuvre	phase	parametrized					
61	system	road	lane	detection	vision	vehicle	based	lane detection				
62	driving	network	self	communication	vehicular	car	vehicle	vehicle communication	obs samma som 72 och 99			
63	fuzzy	control	decision	logic	system	set	paper					
64	image	detection	estimation	vision	stereo	road	based	Road detection	samma som 25			
65	re	maximum	cubic	interpolation	manoeuvre	phase	parametrized					
66	vehicle	tracking	particle	method	time	real	filter	vehicle tracking	lik 54, 58			
67	intersection	traffic	autonomous	vehicle	transportation	intelligent	road	?				
68	fusion	sensor	track	filter	information	data	environment	environment information				
69	re	maximum	cubic	interpolation	manoeuvre	phase	parametrized					
70	vehicle	vision	road	algorithm	hough	lane	ransac	image analysis		hough = smart computer guy who does computer vision stuff		
71	graph	control	vehicle	theory	system	dimensional	rigid	?				
72	vehicle	driving	automated	communication	sensor	technology	infrastructure	Vehicles communication	samma som 62 och 99			
73	re	maximum	cubic	interpolation	manoeuvre	phase	parametrized					
74	learning	terrain	classification	mechanical	visual	supervision	automatic	Terrain classification				
75	re	maximum	cubic	interpolation	manoeuvre	phase	parametrized					
76	robot	vehicle	control	sensor	image	autonomous	mobile		0			
77	re	maximum	cubic	interpolation	manoeuvre	phase	parametrized					
78	vehicle	planning	path	control	autonomous	approach	based	path planning	samma som 16			
79	re	maximum	cubic	interpolation	manoeuvre	phase	parametrized					
80	control	vehicle	sensor	tool	rist	robot	guided	guided vehicle				
81	vehicle	system	collision	safety	service	avoidance	sensor	safety management				
82	re	maximum	cubic	interpolation	manoeuvre	phase	parametrized					
83	vehicle	driving	personalization	autopilot	safety	automated	control	automated driving				
84	re	maximum	cubic	interpolation	manoeuvre	phase	parametrized					
85	safety	safety	vehicle	dynamic	participant	verification	markov	safe decision making				
86	autonomous	vehicle	system	car	decision	communication	algorithm	decision making				
87	controller	hierarchical	intelligent	vehicle	control	model	level	intelligent vehicle controller				
88	re	maximum	cubic	interpolation	manoeuvre	phase	parametrized					
89	re	maximum	cubic	interpolation	manoeuvre	phase	parametrized					
90	trajectory	profile	generation	velocity	curvature	curve	vehicle	Vehicle curve handling				
91	road	detection	representation	pedestrian	image	geometry	classification	image classification in traffic				
92	re	maximum	cubic	interpolation	manoeuvre	phase	parametrized					
93	control	vehicle	field	potential	aircraft	avoidance	spline	POTENTIAL AIRCRAFT AVOIDANCE				
94	vehicle	information	behavior	tracking	intersection	position	system	intersection handling				
95	vehicle	system	autonomous	time	real	utility	coordination	autonomous vehicle coordination				
96	re	maximum	cubic	interpolation	manoeuvre	phase	parametrized					
97	vehicle	localization	visual	method	feature	route	robot	localization/navigation				
98	control	vehicle	motion	autonomous	system	dynamic	tracking	autonomous vehicle motion tracking				
99	vehicle	communication	system	wireless	intelligent	highway	roadside	vehicle communication	obs samma som 62 och 72			