A topography in the field of audit research – a classification model by means of RapidMiner

Masterarbeit
zur Erlangung des akademischen Grades

Master of Science
im Masterstudium

Finance and Accounting
EIDESSTATTLICHE ERKLÄRUNG

Ich erkläre an Eides statt, dass ich die vorliegende Masterarbeit selbstständig und ohne fremde Hilfe verfasst, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt bzw. die wörtlich oder sinngemäß entnommenen Stellen als solche kenntlich gemacht habe.

Die vorliegende Masterarbeit ist mit dem elektronisch übermittelten Texdokument identisch.

Ort, Datum
Linz, 30.4.2019

Unterschrift

[Unterschrift]

[Signature]
# Table of Contents

1 Introduction .................................................................................................................. 2

1.1 Introductory words ................................................................................................. 2
1.2 Problem statement and research goals ................................................................... 3
1.3 Methodology ........................................................................................................... 4

2 Systematic Literature Review ..................................................................................... 5

2.1 Theoretical background ....................................................................................... 5
2.2 Database .................................................................................................................. 6
2.2.1 Journal and article sample .............................................................................. 6
2.2.2 Time limitation ............................................................................................... 6
2.3 Procedure ............................................................................................................... 7
2.4 Descriptive statistics ............................................................................................. 7

3 Thematic classification model by using RapidMiner .................................................. 9

3.1 Theoretical background LDA model ...................................................................... 9
3.2 RapidMiner ............................................................................................................. 9
3.3 Fundamental terms ............................................................................................... 10
3.3.1 Operators ........................................................................................................ 10
3.3.1.1 Process ...................................................................................................... 10
3.3.1.2 Subprocesses .......................................................................................... 11
3.3.1.3 Breakpoints ......................................................................................... 11
3.3.2 Parameters ....................................................................................................... 11
3.3.3 Ports ................................................................................................................ 11
3.3.4 Example set ..................................................................................................... 11
3.3.5 XML ............................................................................................................... 12
3.4 The RapidMiner process ...................................................................................... 12
3.5 The data import process ..................................................................................... 12
3.5.1 Conversion of database .................................................................................. 12
3.5.2 Import of database with Loop Files ............................................................... 13
3.6 Text preprocessing: the inner process of Loop Files .......................................... 14
3.6.1 Read Document .............................................................................................. 14
3.6.2 Tokenize .......................................................................................................... 15
3.6.3 Stem (Snowball) ........................................................................................... 15
3.6.4 Filter Stopwords (English) ............................................................................ 16
3.6.5 Intermediate preprocessing results ................................................................ 17
3.6.6 Filter Stopwords Dictionary .......................................................................... 18
3.7 Extract Topics from Documents ......................................................................... 19
3.7.1 Parameters ..................................................................................................... 20
3.8 Export of results ................................................................................................... 20
3.9 Parameter optimization ......................................................................................... 21
List of Figures

Figure 1: Growth of scientific publications 1980 – 2012 .......................................................... 3
Figure 2: Published articles over time ....................................................................................... 8
Figure 3: Publishing journal countries ..................................................................................... 8
Figure 4: The outside process ..................................................................................................... 13
Figure 5: The inner/sub process ............................................................................................... 14
Figure 6: Intermediate preprocessing result ............................................................................... 17
Figure 7: Stopwords Dictionary word cloud ............................................................................. 18
Figure 8: Extract Topics from Documents operator ................................................................. 19
Figure 9: Results export - Write Excel operator ....................................................................... 20
Figure 10: Operators building the Stopwords Dictionary .......................................................... 22
Figure 11: Parameters for Read Excel operator ......................................................................... 22
Figure 12: Topic classification results ....................................................................................... 23
Figure 13: Top words Auditor-client relationship and partner rotation .................................... 27
Figure 14: Top words Audit quality and auditor tenure .............................................................. 29
Figure 15: Top words Going concern ....................................................................................... 30
Figure 16: Top words Liability and litigation .......................................................................... 32
Figure 17: Top words Audit risk and fraud ............................................................................... 33
Figure 18: Top words Audit market and fees .......................................................................... 36
Figure 19: Top words Auditor's judgement .............................................................................. 38
Figure 20: Top words Corporate governance .......................................................................... 40
Figure 21: Top words Audit profession .................................................................................... 42
Figure 22: Top words Audit procedures .................................................................................. 44
Figure 23: Top words Audit report, financial standards, financial statements and its users ....... 46
Figure 24: Top words International regulations, regulatory reviews and SOX ......................... 47
List of Tables

Table 1: Topic confidence values in percent ................................................................. 24
Table 2: Typology with keywords ............................................................................ 49
### List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AC</td>
<td>Audit committee</td>
</tr>
<tr>
<td>CA</td>
<td>Continuous audit</td>
</tr>
<tr>
<td>CEO</td>
<td>Chief Executive Officer</td>
</tr>
<tr>
<td>cf.</td>
<td>compare</td>
</tr>
<tr>
<td>CSR</td>
<td>Corporate Social Responsibility</td>
</tr>
<tr>
<td>e.g.</td>
<td>for example</td>
</tr>
<tr>
<td>et al.</td>
<td>and others</td>
</tr>
<tr>
<td>etc.</td>
<td>et cetera</td>
</tr>
<tr>
<td>EU</td>
<td>European Union</td>
</tr>
<tr>
<td>ex.</td>
<td>example</td>
</tr>
<tr>
<td>exa</td>
<td>example set</td>
</tr>
<tr>
<td>GAAP</td>
<td>Generally Accepted Accounting Principles</td>
</tr>
<tr>
<td>ID</td>
<td>identity</td>
</tr>
<tr>
<td>IFRS</td>
<td>International Financial Reporting Standards</td>
</tr>
<tr>
<td>LDA</td>
<td>Latent Dirichlet Allocation</td>
</tr>
<tr>
<td>p.</td>
<td>page</td>
</tr>
<tr>
<td>pp.</td>
<td>pages</td>
</tr>
<tr>
<td>PCAOB</td>
<td>Public Company Accounting Oversight Board</td>
</tr>
<tr>
<td>PDF</td>
<td>portable document format</td>
</tr>
<tr>
<td>RPM</td>
<td>Remote Presentation Model</td>
</tr>
<tr>
<td>SOX</td>
<td>Sarbanes-Oxley Act</td>
</tr>
<tr>
<td>top</td>
<td>top words</td>
</tr>
<tr>
<td>txt</td>
<td>text file</td>
</tr>
<tr>
<td>UK</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>US</td>
<td>United States</td>
</tr>
<tr>
<td>Vol.</td>
<td>volume</td>
</tr>
<tr>
<td>XML</td>
<td>Extensible Markup Language</td>
</tr>
</tbody>
</table>
List of Annexes

Annex 1: List of included journals.........................................................................................................................................................59
Annex 2: Stopwords Dictionary..........................................................................................................................................................60
Annex 3: XML code RapidMiner process.............................................................................................................................................61
Abstract

This thesis presents a review of audit research published between 2002 and 2017. The constitution of the article database was performed by employees of the Institute for Corporate Accounting and Audit at the Johannes Kepler University in Linz by defining certain inclusion and exclusion criteria. These articles are the basis for the further proceedings. The subsequent review of the article database was done by conducting a systematic literature review. At first, the descriptive analysis of the research field by forming categories was performed manually together with several colleagues by using an Excel spreadsheet. Secondly, for the purpose of this master thesis, an automatized categorization process with the software RapidMiner was produced in order to present a new approach to the topic categorization process for systematic reviews. The classification model used is an LDA model which is a probability model for document grouping. Based on the topic classification model output the key emerging themes of every topic category are discussed and recent trends in the audit research field are outlined. The results of the review present a topography and an overview of the audit research field and offer a framework for researchers and students to categorize their research. Furthermore, the constructed RapidMiner classification model can be applied to other fields of research by minimally modifying the parameters.
1 Introduction

1.1 Introductory words

Nowadays we live in a world with an excess of data and the worldwide data volume is growing at a rapid rate. Depending on different factors, it can be seen as a curse or as a blessing. Naturally, the question of how to deal with this data flood arises. The answer to this question is most likely Big Data, which allows collecting and analyzing large data volumes by applying special algorithms.1 Whether at the trading floors, at online companies, in the natural sciences or healthcare – all these sectors are telling the same story: the worldwide data volume is increasing significantly, which sometimes not only exceeds the capacities of computers but also our imagination. There is no unified definition of the term Big Data. Originally it described an amount of information that was too large to process with a computer but nowadays is understood as the ability to use information or data in a way that new findings and knowledge of important value are gained. The larger the data scale it is applied on is, the more valuable the findings are.

A good example for the usefulness of Big Data is the model Google developers designed in order to predict the spread of the flu epidemic of H1N1 in the US in 2009. The health authorities and authorities for control of epidemics established a reporting obligation for new flu cases, but the thereby acquired information about the progression of the flu was always delayed by two weeks, which is a critical delay for a rapidly spreading epidemic. Google had stored an enormous amount of data regarding flu-specific online search terms of their customers, and they used it to develop a model which could determine the spread and progression of the flu. The difference to the reporting obligation system of the authorities was that they could determine the spread of the flu immediately and without delay. They used a model which analyzed the correlation between the frequency of flu-specific search terms and the spread of the flu by means of mathematical procedures. With this volume of data and the right model to analyze it, the authorities gained an important information advantage in the battle against the H1N1 epidemic in 2009, as it was a faster and more efficient indicator than the government statistics.2

As mentioned above, Big Data is of great importance for various industries in this day and age. In the field of audit Big Data can help to make more appropriate as well as timelier decisions using the potential of data mining for audit procedures.3 But also in the

---

1 Cf. LORENZ-MEYER, A. (2014)
audit research field Big Data can be beneficial, as the data volume in the academic research field is increasing steadily. Figure 1, from a study by Bornmann & Mutz, presents the growth of global scientific output regarding publications between 1980 and 2012. From this figure one can see that the volume of scientific publications doubled in 24 years.⁴

**Figure 1: Growth of scientific publications 1980 – 2012**  

Due to the continuous rise of academic research, getting an overview of the topics which have already been addressed (and to which extent) can be very challenging. This is of great importance to researchers in order to categorize their research work and to identify research gaps. This also applies to students working with academic literature.

This thesis addresses the research field audit research and is dealing with the stated problem of how to give an overview over the academic field of audit research with the help of Big Data or in this case more precisely text mining. Text mining analyzes large collections of documents and attempts to find useful information within the data sources by identifying and exploring relevant and interesting patterns.⁵ The goal of this thesis is to create a topography of the topics in the academic audit research landscape.

### 1.2 Problem statement and research goals

As discussed in chapter 1.1 the increasing volume of data worldwide also carries challenges for the audit research field. Academic research publications are rising, the

---

⁵ Cf. Verma, T. et al. (2014), p. 16
database gets broader and deeper and it can be hard to get an overview of the structure of existing research or research gaps. This thesis therefore aims to create a classification or rather categorization of existing audit research. With the help of the resulting classification model, researchers can get an overview of the existing academic audit research and its topics. The model shows which topics are covered the most in this field, which topics are interrelated and also helps researchers to classify their own research. Furthermore, it makes identifying research gaps easier and gives implications for future research. The scope of the audit research database is narrowed down to the highest rated journals and the time period in which the articles are published is limited to the most recent academic research from 2002 on, when SOX was passed (detailed explanation and breakdown of the database see chapter 2.2 Database).

Beyond the outcome of a topography of the existing audit research field, this classification model also provides a standard model for automatically classifying existing databases in any other research field with minimal effort. With just little changes in parameters (see chapter 3.8 Parameter optimization) the model can be applied, and the main topics can be identified. Furthermore, updating the classification of audit research is easy in the future, as the unmodified process can be readily applied to an extended database.

1.3 Methodology

This thesis is a mix of the following two research methods. At first in chapter 2 the database of academic research papers about external audit is outlined and the Systematic Literature Review, which has been executed manually, is described. As mentioned previously, getting a proper overview of existing research can be very difficult and if it is possible, it is extremely time-consuming. This has also been confirmed throughout this thesis in chapter 2, as the effort was enormous, and several people had to be involved in the process.

In chapters 3 and 4, which represent the main part of this thesis, the same database is processed. The difference to the approach in chapter 2 is the fact that the articles are classified with the software RapidMiner, applying a mathematical three-level hierarchical Bayesian model called Latent Dirichlet allocation (further referred to as LDA), which is a probability model for document grouping.6 This software-based categorization model simplifies and automates the process of document classification and can therefore

---

massively facilitate the effort put into academic research categorization. Subsequent to the categorization of the database the key matters and emerging themes are revised. At the beginning of the chapters the theory of the Systematic Literature Review and the LDA model is briefly summarized in order to guarantee a better understanding of the following thesis.

2 Systematic Literature Review

2.1 Theoretical background

The goal of conducting a systematic literature review is structuring a respective research field for both policymaking and practice. It provides evidence in the form of knowledge being systematically managed, identifying emergent themes and pointing out important research gaps. Thus, systematic reviews contribute to research theory development by conducting a both quantitative and qualitative thematic analysis.7 Tranfield et al. (2003) define a systematic review process as synthesizing existing research in a systematic, reproducible and transparent way with the purpose of improving the knowledge base and informing practice and policy makers. The aim is to create a reliable stock of knowledge and identify key scientific contributions. The process of a systematic review is divided into the following 3 stages.8

→ **Stage 1**: Planning the review

The planning of a review starts with regular meetings of the respective participants (e.g. experts or practitioners). Thus, all difficulties or disputes regarding the exclusion or inclusion of studies can be resolved. The needs of the review are analyzed, a review proposal is prepared, and a review protocol is developed.

→ **Stage 2**: Conducting the review

The first step when actually conducting a review is the identification of search terms and keywords, which are derived from scoping the literature and discussions of the review team. Included studies which are stated in the review protocol are selected and extracted. As a result, the output of information searching is a list and database with all papers on which the systematic review is based.

→ **Stage 3**: Reporting and dissemination

In stage 3 a descriptive analysis of the research field is presented. This descriptive analysis is achieved by developing a set of categories divided by e.g. authors, journal origin, age profile, key orientations or research topic fields. Moreover, researchers are reporting to what extent there is consensus in literature, where contrary positions occur and key emerging themes.

The following chapters 2.2.1 and 2.2.2, which describe how the final database is composed, represent stages 1 and 2 of the systematic review process. Chapter 2.3, which describes the manual classification of the database, constitutes stage 3 of the review. The focus of this thesis, however, is the automatic classification and identification of key matters and emerging themes in chapters 3 and 4, representing another approach to perform stage 3 of the review process.

2.2 Database

2.2.1 Journal and article sample

When scanning for journals containing academic papers about audit research, only internationally recognized journals have been included. Furthermore, journals without or with restricted university access were excluded. From these journals, only the ones with a ranking of 2 and 3 were chosen. The resulting database has been narrowed down to English papers that deal with audit research exclusively. The aim was to have a database which solely contains external audit research, which is why articles dealing with topics like Environmental Audit, CSR Audit, Medical Audit, Internal Audit, Information System Audit, Project Audit, Security Audit and many more were excluded. The number of articles, which are listed in the final database, amounts to 1,717 articles. Annex 1 lists the 28 journals the articles have been extracted from. The database was constituted by employees of the Institute for Corporate Accounting and Audit at the Johannes Kepler University in Linz.

2.2.2 Time limitation

In addition to the inclusion and exclusion criteria mentioned in chapter 2.2.1, time was also a major criterion for the inclusion of articles into the database. The time frame was narrowed down to 2002 to 2017. The starting point of reviewing the literature was set in the year 2002, seeing as the US federal law SOX was passed in that year in order to improve the reliability of financial reporting. It was passed as a reaction to the financial reporting scandals at amongst others Enron, Peregrine Systems, Adelphia and
WorldCom, which resulted in huge losses for numerous investors when the share prices collapsed. SOX sets out changes in corporate governance and corporate financial reporting for public listed companies as a response to these financial scandals.  

2.3 Procedure

The manual literature review based on the database was carried out together with colleagues of the Institute for Corporate Accounting and Audit at the Johannes Kepler University in Linz. All 1,717 articles were looked at and classified according to predetermined criteria in an Excel spreadsheet and sample checks were performed. Every article ID was manually categorized according to research method, main topics, theories cited, keywords, number of observations/interviews/case studies etc., and time of data collection. Concerning the research method, the classification followed the scientific article of Brown/Jones. The topic-wise classification followed two different academic articles: KNECHEL et al. and LESAGE/WECHTLER. This is related to the automatic topic-wise classification with the help of the software RapidMiner, which will be discussed in chapter 3, and the identification of key emerging topics in chapter 4.1.3.

2.4 Descriptive statistics

The inclusion criteria for the database has already been laid out in chapter 2.2. This current chapter will explain the database in more detail.

Figure 2 shows the number articles in the final database with regard to their year of publication. As previously mentioned, the database starts with articles published in 2002 when the SOX regulations were implemented. This explains the rapid rise in publications on audit research in 2003, as well as the many audit research discussions in the following years. The rapid fall in 2017 can be dismissed, as it is due the fact that only articles up to the month of August of that year were included, which was when the database was composed. The other years represent full year coverage.

---

While Annex 1 shows the various different journals where the database was extracted from, Figure 3 displays the journal-country distribution. The majority of articles comes from US journals, followed by UK journals and only a tiny part of the database originates from journals from Australia and the Netherlands.

The most represented journals are the following ones, covering over 50 percent of the database.

- Auditing: A Journal of Practice & Theory (288 articles)
- Managerial Auditing Journal (285 articles)
- International Journal of Auditing (151 articles)
- Accounting Review (143 articles)
3  Thematic classification model by using RapidMiner

3.1  Theoretical background LDA model

Blei et al. (2003) define the LDA as “a generative probabilistic model for collections of discrete data such as text corpora”. Every item of a document collection is modeled as a finite mixture over topics and every topic is modeled as an infinite mixture over topic probabilities. Regarding text modeling, the topic probabilities offer a clear representation of every single document of the collection. In short this means that every item of the document collection is represented by a finite mixture of topic probabilities, representing every topic with a value between zero and hundred percent. Thus, every document is represented by every topic with a certain percentage value. The likelihood of belonging to one of all the topics is the highest, when the probability value is the highest.

3.2  RapidMiner

RapidMiner is one of the leading data science tools. It is a Java-based data science and analytics software with a quality graphical user interface, with which it is possible to perform advanced data analytics, including machine learning, predictive analytics, data mining, text analytics, business analytics, and a visualization of results – with little or no coding required. It integrates with several data source types including Excel, Access, Oracle, Microsoft SQL, dBASE, text files and word files. The RapidMiner data science platform includes several products, the one relevant for this thesis is RapidMiner Studio. It enables users to design their own code-free processes and models for data analysis. Consequently, high volumes of data sources can be accessed, transformed, analyzed and visualized.

RapidMiner also offers various extensions on the RapidMiner Marketplace, which extend its capabilities in areas such as text processing, web mining, Python scripting and many more. The extension used for the text mining process designed for this thesis is the text processing extension. Mining text is similar to data mining but focuses on text instead of structured forms of data, therefore making it possible to uncover similarities and associations in text data without guidance.

Regarding pricing, RapidMiner offers a free 30 day trial version and three different models for enterprise use ranging from 2,500 to 10,000 dollars per year and varying dependent on data rows, number of logical processors and productivity features.\textsuperscript{14} RapidMiner has an educational license program, which offers a free license for students and professors for one year. This is a version without data row limitation and includes the Auto Model and Turbo Preparation opportunities, which normally are just accessible when using the paid license.

3.3 Fundamental terms\textsuperscript{15}

3.3.1 Operators

Operators, which are grouped by function, are the building blocks which create a RapidMiner process. Every operator has an input and an output port and inbetween those ports the action of the operator is performed. The parameters of the operator control the action performed. RapidMiner has more than 1,500 operators available. If more than the basic operators are needed, extensions from the RapidMiner marketplace can be installed which contain additional operators. The operators are structured by the following functions: Process control, Utility, Repository Access, Import, Export, Data Transformation, Modeling and Evaluation. They can be selected and added to the desired place within the process view window by drag and drop and then connected by drawing a line between their input and output ports. If several operators are connected one speaks of a process.

3.3.1.1 Process

A process is the product of several interconnected operators and presents the analytical workflow. In the process every operator shapes and processes the data imported. At the beginning of the process it may just transform or convert the files, and in the end, it applies a model to the data set. A very important point in designing the process is the execution order in which the operators are placed. The data has to be prepared properly before applying a model to it, and even in the data preparation process certain operator order rules have to be observed. The process is shown in the design view, while the results of the process are shown in the results view.

\textsuperscript{14} Cf. Rapidminer, Inc. (2018) Pricing
3.3.1.2 Subprocesses

Operators can have subprocesses, which are shown by double clicking on the operator. These subprocesses are performed in the background and are the inner processes of the operators.

3.3.1.3 Breakpoints

So-called breakpoints can be set after the execution of an operator, in order to see the intermediate result of a process. For this the process execution stops and the result after the tagged operator is shown and can be virtually examined. This can be extremely helpful after the data preparation process in order to investigate whether all parameters are set correctly, whether the order of the operators is set right and whether the inner process works. Thereby it can be ensured that the data set is prepared correctly in order to apply a model to the data set.

3.3.2 Parameters

Parameters, whose values can be determined, are the settings which determine and control the execution of an operator. They are shown with a click on the operator on the right side of the user interface. Naturally, the parameters differ from operator to operator as the operators perform completely different actions and are required for correct functionality.

3.3.3 Ports

Input and output ports connect the operators and thereby fashion them into a functioning process. The help view on the right side of the user interface precisely states the type of input a certain operator requires and the output type it produces. This input can simply be a document, a collection of documents, a prepared example set, or a matrix. An output port can also deliver documents or example sets but also models which are applied to the data input.

3.3.4 Example set

An example set is a table created from the imported data with the examples in rows and the characteristics of the examples in the columns. It is the structured and prepared data set which can further be analyzed. The examples within the example set have concrete values and elements which can be compared to other examples of the example set.
3.3.5 XML

The created process is not only described by the graphic illustration in the process view where it is built, but also scripted with an XML code, which is generated by RapidMiner in the background, while users build the process graphically at the graphic user interface. This XML code can simply be copied and used by other RapidMiner users with the same operators, connections and parameters. XML is short for Extensible Markup Language, which is a textual format that is both machine-readable and readable for humans. Moreover, the process can also be exported as an RPM file, which is an architectural pattern in software development, and can be directly imported into RapidMiner from other users.

3.4 The RapidMiner process

The RapidMiner process built for the classification needs of this thesis is divided into three parts.

→ The data import process (with the Loop Files Operator)
→ The text preprocessing procedure (by tokenizing, stemming and filtering the database)
→ The Extract Topics from Documents operator (applying the LDA model on the preprocessed database)

The outside process is built with two different operators: The Loop Files operator, which imports the whole database, and the Extract Topics from Documents operator, which applies the LDA model to the preprocessed database.\(^\text{16}\) The inside process is placed inside the Loop Files operator and can be opened by double-clicking on this operator.\(^\text{17}\)

3.5 The data import process

3.5.1 Conversion of database

As pointed out in chapter 3.2 RapidMiner is processing various file types. The database relevant to this thesis described in chapter 2 consists of PDF files, which is not ideal for RapidMiner processing. Therefore, the 1,717 PDF files were converted into text files. Seeing as not all PDF files were convertible into text files, the final database consisted of 1,695 articles, which corresponds to a rejection rate of 1.2 percent.

\(^\text{16}\) See Figure 4 – page 20
\(^\text{17}\) See Figure 5 – page 21
3.5.2 Import of database with Loop Files

The repository section of RapidMiner is only suitable for single files like Excel spreadsheets for example, so an operator which is able to import several files at once had to be used for the import. This was done by selecting the Loop Files operator. Its path can be determined to various directories in the parameters section. Additionally, the parameter file type, which specifies how to filter the file names, is set to “global” and the box “enable parallel execution” is ticked, which enables the parallel execution of the inner processes in the Loop Files operator. Furthermore, the box “enable macros” is ticked, which includes the following three macros in the process for every file: file name, file type and folder name.

The Loop Files operator executes the inner process (subprocess within the Loop Files operator, see chapter 3.6) tasks on every single selected file. The end result then is propagated to the outside process, and by connecting the output port of the Loop Files operator to any other input port, the data output this operator produces can be used further.

Loop Files is the first part of the outside process. In the second part the classification model is applied by using a specific operator using the LDA model. Figure 4 illustrates the whole outside process in the design view of the graphic user interface. As mentioned before, the inner process of data preparation (text preprocessing procedure) is hidden within the Loop Files operator.

Figure 4: The outside process (own compilation)
3.6 Text preprocessing: the inner process of Loop Files

In this chapter the process of organizing and structuring the imported data collection is described. The preprocessing phase is a crucial one for the efficiency and success of the text mining process. It is the inner/sub process of the Loop Files operator. The data is processed through four different operators which read, tokenize, stem, and filter the individual files for English stopwords. Thereby, the database is structured in a way that makes it processable and valuable information can be extracted. All these four operators are part of the text processing extension available at RapidMiner marketplace. Figure 5 illustrates the inner process of Loop Files on the graphic user interface, which is shown by double clicking on the Loop Files operator.

As previously mentioned, the execution order in which the operators are placed are of great importance to the functionality of the process. At first, the whole data collection has to be read by an appropriate operator. Secondly, it has to be tokenized before it can be stemmed, and it has to be stemmed before complete words are filtered out. In the next four subchapters the four operators, with the help of which the data collection is prepared for analysis, are explained in more detail.

Figure 5: The inner/sub process (own compilation)

3.6.1 Read Document

The operator Read Document reads texts from a collection of multiple files. The parameter “content type” is set to “txt” in order to specify the file type. The encoding used for reading the files is set to “SYSTEM”. Moreover, the box “extract text only” has to be ticked, in order for structural information to be ignored. The input port is connected
to the beginning of the subprocess, which is the output of the Loop Files operator, where the data collection was imported. The output port is connected to the next operator for data preparation.

### 3.6.2 Tokenize

Tokenization is a very important component in the beginning of every analytical task. Tokenizing texts means breaking the whole text into words, symbols, phrases, and meaningful elements, which are called tokens.\(^\text{18}\) In the case of this thesis, every single document of the data collection is thus split into a sequence of various tokens. The aim of tokenization is to explore the words in a sentence and break it down into parts called tokens. There are different options of how to define the splitting points – the data can be split by linguistic sentences, regular expressions, specified characters, and non-letters (which represents the default settings of RapidMiner).\(^\text{19}\) In the case of this thesis the parameter “mode” is set to “non-letters”, consequently splitting the text documents into tokens which consist of single words. This is important for upcoming steps, seeing as it makes the frequency of occurrence of the words countable. By using the non-letters mode, the operator searches for symbols like spaces or hyphen characters and at this point splits the document into a new token. Therefore, the document is split into non-letter tokens only.

### 3.6.3 Stem (Snowball)

Stemming, in linguistics also called lemmatization, means reducing the words to their stems, roots, or base. This stemming process is very useful because words having the same stem are therefore transformed into the same base for further processing of the data. This way words such as *likely, like, liking*, and *unlike* are all reduced to their stem *like*.\(^\text{20}\) Word stemming cannot be applied to all languages but for most of them a pattern of word structure appears. This reduction of various grammatical word forms such as adverbs, verbs, nouns, adjectives etc. to their base form is the main purpose of stemming texts. Thereby, related words or forms of words are reduced to their joint base form which is very important, seeing as the meaning of the words can be considered the same for information retrieval systems. In order to identify the base forms of the words, stemming algorithms are applied. The context in which the words occur is neglected when carrying

out stemming, and the stem does not have to be an existing word from the dictionary which is morphologically correct. After these stemming algorithms have been applied, the texts consist of the word stems rather than the original words. This reduces the number of original words and by that also the processing time required afterwards.\textsuperscript{21}

As in this thesis the topic identification model is based on word frequencies and word probabilities, the stemming process is very important. If the words are not stemmed properly, words having the same word stem and therefore being very similar or even equivalent to each other, would be counted separately and this would distort the topic classification.

Word stemming cannot be applied to all languages, but for most of them a pattern of word structure appears. While for the Chinese language it is impossible to stem words, for the English language and most other Western European languages it is. Regarding this thesis, the parameter “language” is set to the English language, as the article database contains English articles only.

3.6.4 Filter Stopwords (English)

Filtering stopwords is important in order to remove common words from the remaining stemmed tokens, which do not benefit text mining purposes. Such stopwords are usually articles, prepositions, and pronouns. As every document has these stopwords in it, they are not necessary for the mining and classification of texts and can thus be eliminated. A positive side effect of this stopword elimination is the reduction of the dataset and by that a reduced processing time.\textsuperscript{22} The operator Filter Stopwords (English) removes words that are equal to words from a built-in word list from a document.\textsuperscript{23}

Stopwords make up for an extremely large fraction of a text’s size, which makes them unnecessary for information retrieval objectives as they are very common and appear in every document. Accordingly, the stopwords have no real purpose for describing the content of the documents. They are used too frequently to be able to differentiate between non-keyword and keyword terms and can thus be ignored.\textsuperscript{24}

Regarding this thesis, of course the operator Filter Stopwords (English) is used, as all articles of the database which are processed are English ones. By applying this operator, all English stopwords included in the predefined stopword list are removed from the data.

\textsuperscript{22} Cf. VERMA, T. et al. (2014), p. 17.
In order for this operator to function properly, the documents have to be tokenized first (see 3.6.2). As this operator refers to a built-in list in the background, no parameters have to be set.

3.6.5 Intermediate preprocessing results

As already mentioned, breakpoints can be set before and after every operator in order to control and analyze intermediate results. After the preprocessing described in chapter 3.6 this can be very useful for determining if the process is built correctly and works. When setting a breakpoint after the last preprocessing operator Filter Stopwords and running the process, it shows the results for the process up to the breakpoint. In this case, every preprocessed document is shown, one after the other, by clicking the run button again and again. Below, every document is displayed unprocessed without being tokenized, stemmed and filtered for comparison purposes. Figure 6 illustrates an example of one arbitrary text document, which is tokenized, stemmed and filtered by stop words.

Figure 6: Intermediate preprocessing result (own compilation)
3.6.6 Filter Stopwords Dictionary

The Filter Stopwords Dictionary is a user-defined stopwords list, which is filtered from every single document, like the stopwords from the operator Filter Stopwords (English) described in 3.6.4, with the difference of not using a built-in stopwords list but defining the stopwords manually. Thereby, words which are not characteristic of the thematic classification on their basis, are eliminated from the text and are therefore not distorting the classification result. This can concern words which are characteristic of the whole research topic, like in this thesis audit, auditor, percent, or rate for example – and are thus appearing in every single article of the database. Secondly, it can also refer to words which appear in almost every document because they are typically indicated in the publication details, such as e.g. download, journal, kepler, universitaet, vol, paper, etc. Also, words which relate to the research method of the articles, like for instance survey, interview or interviewee, are excluded. The full Stopwords Dictionary list can be found in Annex 2. Figure 7 illustrates the defined collection of stopwords with the help of a word cloud.

Figure 7: Stopwords Dictionary word cloud (own compilation)
3.7 Extract Topics from Documents

The operator Extract Topics from Documents applies an LDA model, which allows the identification of topics in documents, on the preprocessed database. The input port requires to be connected to a preprocessed collection of documents. Therefore, the output port of the Loop Files operator has to be connected to the input port of the Extract Topics from Documents operator. The preprocessing of the data collection within the Loop Files operator has already been described in chapter 3.6.

The operator has four different output ports, which can be connected with the result ports, if the result is desired.

- The output port **exa** delivers an example set with all document IDs and the topic classification results per document with the corresponding confidence values in every single topic.
- The output port **top** delivers details on each topic by providing the top words per topic.
- The output port **mod** delivers the topic model, which can be applied to new collections of documents.
- The output port **per** delivers the log likelihood value of the fit, which can be used for optimization assessments.

Figure 8 illustrates the Extract Topics from Documents operator with its output ports and the setting of its parameters on the right side.

**Figure 8: Extract Topics from Documents operator (own compilation)**
3.7.1 Parameters

The parameters number of topics and top words per topic have to be set manually. The parameter number of topics is set to 12 in this thesis. The optimization regarding this setting is highly important for an optimal classification result and is described in chapter 5.8 Parameter optimization. The parameter top words per topic does not influence the outcome, as it simply sets the top words which are exported in the results. The parameter iterations describes the multiple repetition of the process regarding the goal of convergence towards a solution.

3.8 Export of results

The classification results can be processed to the results view of RapidMiner. For analysis and optimization of parameters purposes, the results can also be exported in various data formats like e.g. Access, Excel, CSV, ARFF or XRFF. For the purpose of this thesis the export format Excel was chosen. Figure 9 illustrates the adapted export process. The Write Excel operator is used two times, in order to export the topic classifications (export port exa) and the top words per topic (export port top). Both ports are connected to a Write Excel operator and the export location has to be determined in the parameter section.

Figure 9: Results export - Write Excel operator (own compilation)
3.9 Parameter optimization

The classification results are optimized by varying the following two parameters and subsequently analyzing the different outcomes, which is the subjective interference in the classification process.

- Number of topics
- Stopword Dictionary

3.9.1 Number of topics

For the optimization of the number of topics parameter, it is essential that one has a general understanding of the underlying database, in the case of this thesis the rough audit research issues. Depending on the scope of the database, the results are exported several times with various different settings of the parameter number of topics and the best result is chosen as the parameter number of topics. This explains why a rough knowledge about the database is essential for subjectively influencing the classification with parameters. Regarding this thesis the results were exported with the parameter set from 8 to 25, and the classification result was optimal with the parameter set to 12. The decision, if the number of topics set is optimal, was based on the mix of top 10 words assigned to each topic. If the top 10 words of each topic made sense to be regularly associated with each other – and distinct topics which are explainable by these top words are emerging – the number of topics parameter is set ideally.

3.9.2 Stopword Dictionary

Simultaneously to the optimization of the parameter number of topics, the Stopwords Dictionary has to be optimized. As already mentioned in chapter 3.6.6, the words assigned to the Stopword Dictionary are excluded from the text and therefore also from the top words, which means that they do not influence the classification. The assigned top words are combed looking for words not being characteristic of the topics. After having excluded such words from the top words by updating the Stopwords Dictionary, the process has to be run again. Then, the top 10 words are again combed for words not being characteristic of the classification and once again the Stopwords Dictionary has to be updated. This process has to be repeated over and over again, until no uncharacteristic words are included in the top words of the classification result. Figure 10 shows the inside and outside process, with the two red circled operators Read Excel and Filter Tokens.
using ExampleSet building the Stopwords Dictionary and integrating it into the whole RapidMiner process.

**Figure 10: Operators building the Stopwords Dictionary (own compilation)**

The parameter “attribute” within the Filter Tokens using ExampleSet operator has to be set to “word” because the words we are excluding or rather filtering are single words. Figure 11 indicates the parameter “data set meta data information” of the operator Read Excel, which has to be set for the process to successfully integrate the Stopwords Dictionary into the process. Furthermore, the path of the operator has to be set to the Stopwords Dictionary excel file location. The excel file simply lists one excluded stop word after another in the first column with the first row remaining empty.

**Figure 11: Parameters for Read Excel operator (own compilation)**

### 3.9.3 Top words per topic

The parameter top words per topic, in contrast to the parameters number of topics and Stopwords Dictionary, does not affect the classification result. This parameter just sets the number of top words which are characteristic of each topic and are shown in the result. At this point it is extremely important to emphasize that the top words are only characteristic of a specific topic combined with each other, especially when it comes to comparatively general words.
4 Classification model results

4.1 Topic predictions

4.1.1 Overview topics

The processed database of 1.695 articles was classified into the 12 topics in Figure 12. The figure represents the number of documents assigned to each topic. The output of RapidMiner itself is an example set with each document ID and its topic predictions from topic 0-11. Consequently, the topics have to be named manually. In this thesis, it was decided to name them based on the characteristic top 10 words of the classification output, in parts similar to the Inductive Typology of Auditing Research by Lesage and Wechtler.

Figure 12: Topic classification results (own compilation)

The LDA working behind this RapidMiner classification model assigns confidence values (topic probabilities) for each of the 12 topics to every single document ID of the database. More precisely, this means that a document can deal with various topics but is assigned to the topic it covers the most. The principle of the confidence values is further specified and explained in the following chapter 4.1.2.

4.1.2 Topic confidence values

The documents of the database are allocated to one of the 12 topics (topic 0 – topic 11) in the end, even if they concern more than one topic. More specifically, they are assigned to the topic with the highest confidence value/topic probability.
However, the topic confidence values/topic probabilities do not only have the sole purpose of being able to assign every document to the most probable or rather most covered topic. The great advantage of the LDA model with its topic probabilities is the possibility of showing a document as a mixture over topics, where the topic probabilities represent the distribution over the various topics. This represents the difference to and advantage over classical clustering techniques, where the topic assignment is restricted to one topic/cluster only. Table 1 presents a sample of four individual documents, on which a closer look is taken in order to explain the topic confidence value principle.

### Table 1: Topic confidence values in percent (own compilation)

<table>
<thead>
<tr>
<th>Document ID</th>
<th>Topic 0</th>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic 3</th>
<th>Topic 4</th>
<th>Topic 5</th>
<th>Topic 6</th>
<th>Topic 7</th>
<th>Topic 8</th>
<th>Topic 9</th>
<th>Topic 10</th>
<th>Topic 11</th>
</tr>
</thead>
<tbody>
<tr>
<td>ex. 1</td>
<td>996</td>
<td>3.24</td>
<td>65.38</td>
<td>0</td>
<td>0.57</td>
<td>4.42</td>
<td>13.95</td>
<td>0.57</td>
<td>27.50</td>
<td>2.67</td>
<td>0.06</td>
<td>1.83</td>
</tr>
<tr>
<td>ex. 2</td>
<td>1165</td>
<td>0.14</td>
<td>0.02</td>
<td>1.8</td>
<td>1.76</td>
<td>39.71</td>
<td>13.22</td>
<td>0.27</td>
<td>10.38</td>
<td>0</td>
<td>23.03</td>
<td>0.0</td>
</tr>
<tr>
<td>ex. 3</td>
<td>1647</td>
<td>0.71</td>
<td>0</td>
<td>23.86</td>
<td>61.46</td>
<td>0</td>
<td>0</td>
<td>12.69</td>
<td>0</td>
<td>0</td>
<td>1.21</td>
<td>0.06</td>
</tr>
<tr>
<td>ex. 4</td>
<td>1017</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>94.93</td>
<td>0</td>
<td>0.01</td>
<td>15.41</td>
<td>6.10</td>
<td>0.04</td>
<td>2.46</td>
</tr>
</tbody>
</table>

#### 4.1.2.1 Article 996 – The Relationship between Auditor Tenure and Audit Quality Implied by Going Concern Opinions

As the title implies this article deals with the relationship between auditor tenure and audit quality by examining audit failures, earnings management, and issued audit opinions. The sample for the analysis includes stressed bankrupt companies and stressed non-bankrupt companies. With the likelihoods of auditors issuing going concern opinions as an indicator for audit quality, the assumption of negative effects of auditor tenure on audit quality is tested. As can be seen from Table 1, the article is mostly allocated to topic 1 (Going concern) and topic 7 (Audit quality and auditor tenure). From this example we can clearly see a weak point in this model. While the subjectively central topic of this article is audit quality and auditor tenure, the article is eventually assigned to the topic going concern. This is because the whole content of the article is analyzed and most of the content is about testing the going concern opinions issued by auditors for the companies included in the sample. At this point it is of great importance to point out that this does not indicate an incorrect result. Contentwise the article covers the topic going
concern at least as much as the topic audit quality and auditor tenure. The LDA correctly assigns the article to the two contentwise main topics with high confidence levels.

4.1.2.2 Article 1165 – Audit fee determinants and the large auditor premium

Article 1165 covers the determination of audit fees looking at both historical as well as today’s company data. In doing so the complexity and size of the audit, the industrial sectors, and at start-up costs in the first years of new audit engagements are being considered. This explains why the article is not only assigned to topic 4 (Audit market and fees) but also to topic 9 (Audit profession) with a high confidence value. The results of the study show that the determinants of audit fees are to a great extent associated with the effort required by the audit engagement, which also relates to the topic audit profession.

4.1.2.3 Article 1647 – De-composition of Fraud-Risk Assessments

Article 1647 is about an auditor’s sensitivity to fraud-risk assessments when management’s attitude suggests low risk. From Table 1 one can see that the article is in the first place correctly assigned to topic 3 (Audit risk and fraud). However, looking at the confidence value for topic 2 (Liability and litigation), the connection to the topic liability and litigation is striking. Working on the topic of risk-fraud assessments, the liability and litigation issue is naturally an important component.

4.1.2.4 Article 1017 – Costs to Comply with SOX

Article 1017 deals with the rising costs of companies to comply with SOX regulations, including external consulting and technology expenses, internal labour costs, as well as auditor attestation charges. The article is primarily assigned to topic 11 (International regulations, regulatory reviews and SOX), followed by topic 4 (Audit market and fees), as it not only covers the SOX regulations per se but also the associated rising audit costs.

4.1.3 Top words per topic and topic descriptions

The Extract Topics from Documents operator builds topics based on frequently recurring words which are characteristic of a topic and often occur together. Figures 13-24 illustrate the top 10 words per topic, which are further explained in the following subchapters. The number after each top word states how often the word occurs in the articles assigned to the respective topic. It is important to interpret the top words collectively and not
individually as the algorithm assigns articles to the topic in which the top words occur often and together with each other. In addition to the top words per topic, the key emerging themes of each topic are explained in the following subchapters.

4.1.3.1 Auditor-client relationship and partner rotation

The calls for mandatory auditor rotations in the current environment substantiate the concerns that audit quality is reduced by longer auditor tenure or rather by a longer auditor-client relationship. While mandatory partner rotation rules exist in both the EU and the US, mandatory auditor rotation rules only exist in the EU, seeing as the US decided to only retain partner rotation rules. However, so far empirical studies have not been able to prove that the benefits of compulsory auditor rotation rules outweigh the costs associated with them.\footnote{Cf. CAMERAN, M. et al. (2015), p. 1.} Moreover, empirical studies associate longer auditor-client relationships with auditors placing higher constraints on extreme reporting decisions of financial performance by the management.\footnote{Cf. MYERS, J. et al. (2003), p. 1.}

When taking a closer look at Figure 13, the connection between the top words\footnote{It shall be noted that top words are written in italics, as they only represent word stems but no full words.} client, relationship, and engagement cannot be negated. The words are appearing from over 4,000 up to over 10,000 times in the 111 allocated articles in a collective manner. Furthermore, a linkage between the top words partner, rotation and mandatory can be observed, as these words appear collectively from over 2,000 times to nearly 20,000 times. Clearly the words quality, independence and pressure in connection with the other top words also make up an important part of the top words section of this topic. The topic category auditor-client relationship and partner rotation not only refers to the duration of the auditor-client relationship (also referred to as auditor tenure) but also to the whole auditor-client relationship itself. Furthermore, in most cases the auditor-client relationship aspect is linked to the rotation aspect. Articles assigned to this topic not only regard various aspects of the auditor-client relationship and the independence issue of the audit profession, but also audit engagement length and partner rotation aspects.
4.1.3.2 Audit quality and auditor tenure

In response to audit failures various critics debate the issue of long auditor tenure leading to a decline in audit quality and subsequently lower financial reporting quality, which is not in the public’s and investor’s interest. Arguments suggest that over time the auditor’s incentives are moving toward profiting from and maintaining the client, while issues like audit quality and risk as well as liability and litigation are paid less attention to. Thus, auditors can become less objective and less strived to detect material misstatements. On the other hand, in the early years of audit engagements the client-specific knowledge, which is critical for the detection of material misstatements, is to a great extent missing. This absence of client-specific knowledge can be partly compensated by industry knowledge, excellent general knowledge, and additional effort put on the engagement - but results in high engagement start-up costs.

There are arguments for and against extended auditor-client relationships. Arguments for long auditor tenure imply that over time auditors develop a better understanding of the client as well as the industry. Additionally, there is empirical evidence that audit failures and litigation are higher within the early years of audit engagements. Furthermore, it is argued that with a longer auditor-client relationship the auditor’s effort to develop and

---

28 See chapter 4.1.3.1
maintain a reputation in the industry increases, which gives the auditor strong incentives to improve and maintain audit quality. Arguments against extended auditor tenure criticize that auditors who become overfamiliar with a client develop blind spots, become less independent, less objective, less skeptical, and become more concerned about maintaining the relationship with the client in order to profit from it. In summary, longer auditor tenure results in both increased client-specific auditor expertise and increased efforts to protect the auditor’s reputation, but also increased incentives for the auditor to please the auditee.\(^\text{30}\)

Looking at Figure 14, one can observe that the top words *quality* and *client*, which are already top words of the topic auditor-client relationship and partner rotation, are also top words for this topic category. This already indicates the similarity of the two topics. Of course, the collective appearance with other top words of Figure 14 has to be considered as well. The difference between the two topics is that the topic *audit quality and auditor tenure* is focused on the duration of the auditor-client relationship (also referred to as auditor tenure) and is not about all facets of the relationship, which is the case with the previous topic *auditor-client relationship and partner rotation*. In general, it can be said that the two topics are quite similar, with one focusing on all aspects of the auditor-client relationship and rotation issues, and the other one focusing on audit quality and the duration of the auditor-client relationship. The topic *audit quality and auditor tenure* also includes more specific articles with regard to earnings conservatism, earnings management, audit sampling, non-audit fees, earnings reporting, the Big 4, industry specialization, accrual accounting, auditor size, auditor reputation etc. - all in relation to audit quality and auditor tenure. These aspects of the topic are reflected in the top words in Figure 14, which are collectively appearing between over 7,100 and over 21,000 times. Moreover, a lot of SOX related articles fall into this topic category, as the financial scandals which resulted in the implementation of the SOX regulations, are discussed to have been provoked by poor financial reporting quality.

4.1.3.3 Going concern

If auditors are doubting a company’s ability to continue its existence/operations without the threat of liquidation in the future, they are obligated to issue a going concern opinion. If a company goes bankrupt and the auditor has not issued a going concern opinion, it is considered an audit failure. However, the going concern assessment is among the most difficult audit tasks, which has been confirmed by the history of audit failures in the past, with the result of auditors being questioned after the collapses. In those cases, auditors sometimes identified financial distress and a going concern problem but subsequently did not issue a going concern opinion. This may be due to the fact that auditors can significantly harm auditees by informing the outside world about the financial problems of a company, or in some cases it can also be justified by the fear of losing the audit assignment if not complying with the client’s desires. Nonetheless, issuing going concern opinions in the audit reports is important in order to minimize or avoid the risk of liability when it comes to lawsuits and the damage of the auditor’s reputation if an upcoming bankruptcy of the client is not identified in advance.31

Looking at the financial scandals at companies like Enron and WorldCom etc. ending in bankruptcies, it was found that Enron and 95 other companies out of 228 bankrupt companies received clean audit opinions, meaning that no going concern opinions were issued.

---

issued, the year before their bankruptcy. The consequences of collapses like these scandals emphasize the importance of going concern assessments and opinions.\textsuperscript{32}

Looking closely at Figure 15 it can be seen that amongst some rather general top words, the top words \textit{opinion}, \textit{concern} and \textit{bankruptcy} stand out. Again, interpreting their collective appearance together with the other top words from over 2,600 to over 8,000 times in 98 articles, the allocation to the topic \textbf{going concern} can be regarded as correct. The articles assigned to this topic deal with plenty of auditor-client relationship issues in the going concern assessment process, reporting accuracy, going concern opinion quality in Big 4 companies and non-Big 4 firms, investors reactions to going concern opinions, and many more.

\textbf{Figure 15: Top words Going concern (own compilation)}

\begin{figure}
\centering
\includegraphics[width=\textwidth]{going_concern_top_words}
\caption{Top words Going concern (own compilation)}
\end{figure}

\subsection*{4.1.3.4 Liability and litigation}

Important aspects for auditors are the auditor’s liability and litigation issues. Third parties like investors for instance are allowed to sue audit companies for damages. These damage compensations may result in huge amounts of legal costs for auditors. The liability risk and risk of litigation exposure can be reduced by stating a going concern opinion when auditing a client who is financially distressed.\textsuperscript{33} Therefore, an auditor shall issue a going concern report when there is doubt about the auditee’s capability to remain a going

\textsuperscript{32} Cf. TUCKER, R. et al. (2003), p. 402.
\textsuperscript{33} See chapter 4.1.3.3
concern in the foreseeable future.\textsuperscript{34} When it comes to litigation, it is common that an auditor is named as a defendant when it comes to bankruptcy, undetected fraud or a restatement of the audited financial statements of the client. When these matters are results of materially misstated and audited financial reports, auditors are likely to have performed negligently and may be held liable in litigation.\textsuperscript{35}

Due to the liability of auditors for audit failures, audit firms tend to be unwilling to accept audit engagements from risky clients. The willingness of auditors to perform audit services for risky clients depends a fair amount on the valid legal liability regimes. This raises concerns as it can for instance prevent start-up companies from raising external capital and subsequently from growing and creating jobs. Concerned critics therefore argue that auditors can be compensated for the higher liability risk by higher audit fees, as the liability provided by the auditor represents an implicit insurance for outside investors.\textsuperscript{36}

Linking the aspects above to Figure 16, it can be seen that the top words \textit{litig, client, legal, insur, report, qualiti, liabil, neglig, etc.} represent the key matters dealt with in academic literature about the topic \textbf{liability and litigation}. This topic clearly deals with financial reporting quality in connection with liability and litigation, client risk issues, the legal environment of auditors, the implicit insurance provided by auditors’ liability and negligent behaviour of auditors. The top words appear from over 1,400 to over 5,000 times in 73 articles in a collective manner, which confirms the allocation of the articles to this topic by the RapidMiner LDA model. The topic is without a doubt very much related to the topic \textbf{going concern}, which is also evident by the two identical top words \textit{client} and \textit{report} of the two topics. The connection is obvious, as going concern opinions represent a rather good protection against the litigation risk.

\textsuperscript{36} Cf. LAUX, V./NEWMAN D. (2010), p. 262.
4.1.3.5 Audit risk and fraud

Practitioners and audit policy makers often have concerns about auditors failing to discover or rather detect fraud. Auditors are required to perform separate evaluations of fraud risk. These fraud risk assessments have an important influence on the subsequent audit process, including evaluations of audit evidence and auditor’s judgement. If high levels of fraud risk are determined, auditors later on assess audit evidence more carefully, than if no or low levels of fraud risk are assessed.\textsuperscript{37}

Professional skepticism and the appropriate audit procedures are critical to the fraud evaluations of auditors. Professional skepticism is defined as the attitude of questioning and critically assessing audit evidence, which is essential for the detection of fraud, by auditors. Critics have cited a lack of professional skepticism and thus are emphasizing the need for increased efforts regarding skepticism and additional audit procedures/testing, in order to uncover fraud and material misstatements. The emphasis of audit partners on professional skepticism is essential for an efficient and effective identification of fraud risk factors and appropriate audit procedures for this purpose.\textsuperscript{38}

By taking a look at Figure 17, it can be seen that the terms most commonly used in academic literature about the topic \textbf{audit risk and fraud}, all appear in the top words output of the RapidMiner classification. Again, the allocation of the articles to this topic

can be confirmed as correct, considering the top words appearing collectively from over 5,000 to over 36,000 times in 127 articles. The combination of the top words *risk*, *fraud*, and *assess* is standing out, as it represents word combinations like “fraud risk assessments” or “assess risk”. Further frequent word combinations derived from the top words rating are “material misstatement”, “detect fraud”, “detect errors”, “assess materiality”, “assess evidence”, etc. The topic **audit risk and fraud** clearly deals with articles about audit risk and fraud testing, critically recognizing audit evidence, audit sampling and testing with regard to risk, fraud and misstatements, audit risk models, efficiency and effectiveness of fraud and risk assessments, accounting errors, materiality uncertainty and fraudulent management.

**Figure 17: Top words Audit risk and fraud (own compilation)**

<table>
<thead>
<tr>
<th>Audit risk and fraud</th>
<th>materi, 8489</th>
<th>test, 6896</th>
</tr>
</thead>
<tbody>
<tr>
<td>fraud, 20977</td>
<td>misstat, 6501</td>
<td>error, 6033</td>
</tr>
<tr>
<td>assess, 11474</td>
<td>controll, 5569</td>
<td>detect, 5236</td>
</tr>
<tr>
<td>risk, 36207</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 4.1.3.6 Audit market and fees

The audit market is a segmented one consisting of numerous smaller audit firms and a small number of large ones who dominate the market. Critics have raised concerns regarding this nature of competition and the market power of the dominating audit firms. The decrease of the largest audit firms (from Big 6 to Big 4) has substantiated the concerns further. Following the decrease in major audit service suppliers and their

---

40. The Big 6 became Big 5 due to the merger of Lybrand & Price Waterhouse and Coopers in order to form PricewaterhouseCoopers in 1998. Through the collapse of Arthur Andersen, the Big 5 have been reduced to Big 4 in 2002. The Big 4 consist of Deloitte Touche, KPMG, Ernst & Young and PricewaterhouseCoopers.
increase in market power, there has been an increase in audit fees paid to these dominating audit firms in the past.\textsuperscript{41} Nowadays in developed countries the Big 4 are auditors of more than 75 percent of listed companies, which proves that the concerns of standard-setting bodies about this supplier concentration are justified.\textsuperscript{42} Another issue related to market competition and concentration are the industry specializations audit firms have developed. Such industry specializations can result in cost savings, product differentiation, higher expertise, and sometimes even a higher premium by charging a specialist fee premium for the auditor, but they can also carry an additional risk factor due to the concentration of risk. For the clients and their stakeholders being audited by a specialist auditor means a higher quality of earnings and disclosure and in few cases a lower audit fee derived as a result of economies of scale, but also carries the risk of information transfer across other clients of the same industry.\textsuperscript{43} Regarding audit fee pricing, prior research has reported higher premiums for Big 4 companies than for smaller auditors. These high audit premiums can be connected to the previously addressed market power and domination. Examining audit pricing in general, most of academic research is based on a model of audit fees by Simunic (1980). This model interprets audit fees as a function of litigation risk and audit effort. The audit premium in turn is depicted as an interaction of various factors including expertise, independency, audit quality, audit production technology and financial resources to satisfy legal claims, concerning which in most cases Big 4 auditors are better positioned.\textsuperscript{44} The audit premium therefore represents an incremental audit fee charge, which tends to be higher for big audit firms in the oligopolistic segment of the audit market.\textsuperscript{45} Another important aspect of audit pricing, which is often addressed in academic audit research, is the performance of non-audit services for audit clients. The relationship between non-audit and audit fees is rather complex, which is confirmed by the divergent opinions about the thereby impeded audit quality, independency and arising knowledge spillovers. Regarding audit quality, academic research finds some evidence for non-audit services being linked to reduced independency and preferential treatment. However, most researchers find no impaired audit quality linked to the provision of non-audit services.\textsuperscript{46} Negative consequences concerning auditor independence can result from both economic

\textsuperscript{41} Cf. CARSON, E. et al. (2012), p. 48.
\textsuperscript{44} Cf. CARSON, E. et al. (2012), p. 50.
\textsuperscript{45} Cf. EBRAHIM, A. (2010), p. 106.
and social bonding between auditor and client in the case of joint performance of non-audit and audit services. While economic bonding relates to the financial dependence on the client and the handling of conflicts in a beneficial way for the auditee, social bonding concerns the knowledge-based trust resulting from repeated and constant interaction with the client by performing non-audit services.\textsuperscript{47} Corresponding to research opinions about non-audit services impeding audit quality and independence of the auditor, it is not clarified if the combined provision of non-audit and audit services improves the efficiency and effectiveness of an audit by the emergence of knowledge spillovers. In cases where knowledge spillovers exist, they only enhance audit efficiency and effectiveness if they do not impair audit quality.\textsuperscript{48} Documented evidence of negative effects of non-audit services on audit quality is conflicted, however, one of the most considerable reforms of SOX was the prohibition of a series of non-audit services to audit clients.\textsuperscript{49} Taking a closer look at Figure 18, the top words industri, big, specialist and market can be identified related to the audit market and competition issue by appearing collectively from over 5,900 to over 19,000 times in the 189 articles allocated to the topic. The top words fee, price and premium are characteristic of articles concerning audit pricing/fees and are found from over 5,000 to over 54,000 times in the 189 articles. The top word service is characteristic of the non-audit services and fees aspect of the topic, as the words non and audit have been removed from the database by being integrated in the Filter Stopwords Dictionary. Articles which are assigned to the topic audit market and fees cover issues such as audit market segmentation/concentration, audit market competition, audit firm industry specialization, the Big 4 in relation to market power and audit fees, auditor resignation and the determinants of audit pricing including audit fees, the audit premium, and non-audit fees.

\textsuperscript{49} Cf. KNECHEL, W. et al. (2013), p. 401.
4.1.3.7 Auditor’s judgement

Prior research has shown that errors in auditor’s judgement can be observed in the decision-making process of audit engagements due to biases like anchoring, adjustment or knowledge/experience. These biases are present when auditors focus on initial unaudited values, assessments and conditions or simply adjust values in order to arrive at a specific judgement. Regarding auditor’s judgement, knowledge can even be a curse as for example outcome knowledge may bias the auditor’s evaluations of client conditions or internal control assessment knowledge may bias the auditor’s evaluations of internal control.\(^50\) However, usually auditors’ expertise and knowledge is regarded to mitigate reporting biases and conflicts of interest arising in their decision-making process.\(^51\) Therefore, auditor performance and judgement are commonly enhanced by task- or industry-based knowledge and experience.\(^52\) Other factors influencing auditor’s judgement are pressures regarding time, deadline and budget. Time-budget pressures tend to result in tradeoffs of audit efficiency and effectiveness.\(^53\)

Furthermore, the audit review mode can also influence the judgement of auditors in charge of workpaper documentation. This indicates in turn that the judgement of the reviewer is influenced by the audit review mode, as reviewers are oriented towards

\(^{50}\) Cf. KNECHEL, W. et al. (2013), p. 393.
prepared workpapers. The audit review is the sequential and hierarchical process of interactions between subordinate preparers and reviewers, which concludes with the judgement of the reviewing auditor, typically an experienced auditor. The audit review can be conducted by discussions in person, where preparer and auditor are in the same location, or electronically. In practice electronic reviews are more common, where reviewers and subordinates interact online by discussing and reviewing notes. However, studies have shown that preparers of workpapers devote more pre-review effort to reviews with in-person interaction, which eventually results in a higher quality of judgement and decision-making for face-to-face compared to electronic audit reviews. Another important issue which influences the judgement of audit reviewers is the familiarity of preparer and reviewer. On the one hand, familiarity between team members can reduce uncertainty but on the other hand, it can pose a risk to the objectivity of auditor’s judgements.

Another factor influencing auditor’s judgement are the credibility evaluations, which determine the competence and trustworthiness of the auditee. These evaluations affect multiple subsequent judgements, evidence assessments, and audit effort planning amongst others and are therefore vital to the audit effectiveness and efficiency. The credibility evaluations are strongly determined by familiarity with the auditee and task difficulty.

Reflecting upon Figure 19, we can identify the top words judgment and review, representing key words like auditor’s judgement and audit review, both appearing over 10,000 times in the 188 articles assigned to the topic auditor’s judgement. Furthermore, the top words experi, knowledg, perform, decis, evalu, assess, task and team (emerging collectively from over 4,000 times to over 9,000 times) can be linked to the key matters of the topic, such as experience and knowledge improving audit performance and decision-making, judgement in evaluations and assessments, task complexity, team assessments, and audit review teams. In addition to these key matters articles assigned to this topic cover workpaper documentation methods and quality, performance feedbacks, professional skepticism issues in regard to judgement, fair value measurements, materiality judgements, engagement challenges, electronic audit, interpersonal affairs in view of judgement, the timing impact and auditor’s trust in clients.

4.1.3.8 Corporate governance

Corporate governance refers to rules, policies, resolutions, and controls implemented in order to influence and specify corporate behaviour. In addition to shareholders and advisors, the primary stakeholders who affect governance are members of the board of directors, who are appointed to protect the interests of shareholders by monitoring the management.  

Audit committees (further referred to as AC), which are in charge of overseeing disclosure and financial reporting, are associated with potentially mitigating weaknesses of corporate governance. ACs have become a common part of corporate governance in the last decades, especially since many regulatory bodies have recommended their adoption or have even made them obligatory after high-profile corporate failures/collapses since 2000. There are a number of potential effects of ACs on corporate governance, one of them being the influence on the company’s financial reporting approach, disclosure and compliance with standards. ACs are expected to monitor the compliance with legal and ethical standards as well as the accounting process and

preventive fraud controls of the company, representing the effect of audit committees on financial reporting. Another potential impact is the positive effect on corporate performance, representing an advantage for investors or rather shareholders. By improving control and management practices through ACs, corporate performance can be enhanced. A third potentially influenced dimension is the audit function. ACs can affect the external and internal audit function through the appointment, remuneration and removal of auditors, the settlement of disputes between executive directors and auditors as well as overseeing internal audit regarding risk assessments and financial or compliance goals.\footnote{Cf. TURLEY, S./ZAMAN, M. (2004), p. 307.} In addition to the afore mentioned effects of ACs, an important consequence of their adoption is the discharge of execution and accountability of the director’s responsibilities, which leads to a balancing of accountability and power. However, the implied benefits of ACs depend on various factors such as the level of independence and expertise of the members.\footnote{Cf. ALZEBAN, A./SAWAN, N. (2015), p. 61.}

Not only ACs but also board independence is an important factor regarding corporate governance. After the accounting scandals in 2002, the subsequent regulations require listed companies to have more than half of the directors on boards to be independent ones. While these regulations improve independence within the board of directors, their effectiveness is questioned by critics. This is due to the fact that the independent directors often have limited or inferior information in contrast to corporate insiders, which can give rise to difficulties concerning their responsibility of monitoring and controlling managers and agency problems.\footnote{Cf. ZHANG, J./YU, Y. (2016), p. 794.}

Academic literature documents that firms with CEOs with financial expertise are likely to benefit with regard to better disclosure practices and improved financial policies. In particular since the accounting failures and implementation of SOX, there is an upward trend for appointing CEOs with a background in finance. In literature such CEOs are referred to as financial experts. This rise of CEOs with a financial background is very plausible, as it creates the advantage of an increased focus on disclosure policies and financial reporting in addition to a better capability of handling and using financial resources more efficiently. The advantages of appointing financial experts as CEOs are confirmed by the association of decreasing auditors’ engagement risk and auditees who are managed by CEOs possessing financial expertise.\footnote{Cf. KALELKAR, R./KHAN, S. (2016), p. 326.}
Looking at Figure 20 the top words *corpor*, *govern*, *committe*, *sharehold*, *report* and *control* are reflecting the key matters of this topic: the control mechanism of corporate governance and audit committees for the purpose of protection of shareholders’ interests and the monitoring of financial reporting. Furthermore, the top words *board*, *director*, *ceo* and *expert* relate to the issue of board of directors and the impact of CEOs or rather, management. The addressed top words appear from over 3,000 to over 21,000 times in a collective manner in 89 articles assigned to the topic *corporate governance*. Articles allocated to this topic do not only concern the general discourse of corporate governance bodies but are frequently linked to post-SOX issues, equity incentives of management, and internal audit. Further specific issues that are covered are AC compensation, stock options and incentive compensation of ACs, independence of AC members, AC diligence, auditor choice, shareholder voting on auditor selection, and shareholder dissatisfaction towards corporate governance and auditors.

**Figure 20: Top words Corporate governance (own compilation)**

<table>
<thead>
<tr>
<th>Corporate governance</th>
</tr>
</thead>
<tbody>
<tr>
<td>expert, 7594</td>
</tr>
<tr>
<td>board, 7459</td>
</tr>
<tr>
<td>report, 6661</td>
</tr>
<tr>
<td>ceo, 4781</td>
</tr>
<tr>
<td>govern, 3849</td>
</tr>
<tr>
<td>control, 3205</td>
</tr>
<tr>
<td>corpor, 4157</td>
</tr>
<tr>
<td>director, 7572</td>
</tr>
<tr>
<td>committe, 21228</td>
</tr>
<tr>
<td>sharehold, 3025</td>
</tr>
</tbody>
</table>

**4.1.3.9 Audit profession**

Financial corporate scandals have led to much debate concerning the auditing profession and its aim as well as scope. Auditors must have a body of complex knowledge and a high level of proficiency in order to perform the audit process in compliance with auditing

---

standards to pass regulatory inspections successfully and to keep up with their social and public responsibility. The performance of a financial audit has to verify and assure the validity of the accounting procedures and the annual report for the users of financial statements. Therefore, the audit profession requires presenting a true and fair view of accounts under the compliance with accounting standards.

The audit profession is an important component in the corporate governance system with an ethic responsibility, which is essential not only for investors, but also for the capital markets to work. The financial collapses in the early 2000s have cast a shadow on the audit profession and its professional ethos. In order to reorient the audit profession, the Sarbanes-Oxley Act implemented various regulations for example, which included the prohibition of the performance of certain consulting services and the creation of the PCAOB, which is a standard-setting, monitoring and inspection body for listed public company audits. Before the new regulatory establishments like the PCAOB were established after the high-profile bankruptcies in the early 2000s, the audit profession was largely self-regulated with being subject to peer review mechanisms. The audit firms being subject to independent monitoring institutions nowadays represents a very substantial and fundamental change in the history of public auditing and accounting.

A crucial aspect for the credibility of the audit profession is auditor independence, hence being critical and independent from both the audited company and its management. The auditor is obliged to report the truthfulness and credibility of the financial statements for its users, which is representing an ethical social contract between the auditor and society.

Reflecting on Figure 21 the top words profession, profess, process and perform stand out. These top words concern the articles covering the audit profession itself, the performance of the audit financial reporting process and the essentiality of professional knowledge. The top words independ and critic can clearly be associated with the discussion on independence in the thematic field of audit profession. Furthermore, the top word social can be linked to the social responsibility aspect of the audit profession. The top words occur frequently and collectively from over 2,800 to over 6,200 times in 200 assigned articles. As can be seen by comparing Figure 21 and 22, the topic audit profession is closely related to the topic audit procedures. Out of 10 top words per topic three of them

---

are the same for both topics (perform, process and time). This is perfectly logical due to the similarity of the two topics. However, while the audit profession topic covers the performance of the financial audit in relation to the audit profession within its legal environment and its social responsibility, the topic audit procedures focuses on the production process of the audit. The articles assigned to the topic audit profession predominantly cover credibility and independence of auditors, the legislative framework of the auditing profession, public sector auditing and public interest, performance audits of government institutions, general legitimacy in the audit field and new roles for auditors.

Figure 21: Top words Audit profession (own compilation)

4.1.3.10 Audit procedures

An elemental component of the audit process are analytical audit procedures which affect risk assessments as well as auditor’s judgement. Decision support systems and standard audit programs can facilitate audit procedures and the decision-making process, but also have potential to reduce judgement quality.68 Advances in technology have not only impacted accounting discipline but also the audit profession and its procedures. While only about 25 percent of information was stored in a digital form in the year 2000, nowadays almost every stored information is electronic.

This changing environment requires auditors to adapt audit procedures and techniques and make use of electronic data processing as well as audit automation.\(^6\)\(^9\)

With respect to the increase of electronization and digitalization of business, there is the need and opportunity for the implementation of continuous audit practices. Continuous audit (further referred to as CA) is different to traditional audit regarding its frequency and its focus on automated procedures and processes. CA is a process which examines accounting practices, compliance, business procedures and risk controls on a regular basis, in order to automate error detection in a continuous system which creates alarm triggers in case of errors or anomalies.\(^7\)\(^0\) CA is not to be confused with online auditing or rather computer-aided auditing, as it is simply about a system being run automatically at regular intervals, whereas online auditing means that the auditor is completing a periodic audit assisted by technology. Online auditing is an advanced form of audit processes with significant effects on results and decision making, where the auditor has direct access to the clients’ business data or various other data pools by the use of computer networks. A related concept is the remote audit. The remote audit enables the reduction of the location requirements of audit procedures. The audit team members can communicate through networked equipment and perform audit procedures, obtaining electronic evidence and interact with the client, independent of their physical location. Both remote audit and online audit use information and communication technology.\(^7\)\(^1\)

Looking at Figure 22 the top words perform, process and extern are indicative of the discussion of performing external audit processes and procedures and appear between about 5,100 and 7,100 times in the 179 allocated articles. Again, the collective appearance of the top words with other top words of this category has to be taken into account, especially because three top words of the topic category audit procedures are also indicative for the topic audit profession. Three further important top words appearing from over 4,600 to over 12,900 times are system, control and technolog, which are representative of aspects such as the technological advance and Big Data impacts on audit, standard audit systems and programs, decision support systems, online collaborative audit systems and CA systems. The top word service can be explained by the frequently occurring word combination auditing service providers. The top word model is characteristic of this topic category because it indicates the discussion of audit process modeling. The top word time not only refers to the timeliness of periodic audits,

\(^7\)\(^1\) Cf. ENI, L. (2016), p. 67.
but also to the CA systems providing results in almost real time. In addition to the key matters derived from the top words of this topic, the assigned articles also cover audit procedure automation, various types of audit systems, IT audits, internal audit control audits, Big Data and information analytics effects on audit, audit effort and lags and types of audit software.

Figure 22: Top words Audit procedures (own compilation)

4.1.3.11 Audit report, financial standards, financial statements and its users

An audit report is published along with the annual report in line with the financial statements of a company. It includes the set of followed accounting standards and the audit opinion and is the ultimate product of the audit process. Globalization and the increase in international capital flows have augmented the necessity of a cross-country comprehension of financial reports and audit reports, in order to improve the reader’s understanding. Harmonization of financial standards and audit practice supports the comparability of annual reports and financial statements across countries and is therefore considered to be an appropriate approach for dealing with the previously addressed need for cross-country comprehension.

which are meant to eliminate inconsistencies across the financial statements of companies of various countries, are implemented in more than 100 countries worldwide (including the EU), the US are still holding on to the GAAP rules. Other capital markets without an IFRS mandate are Japan and China.\(^{75}\)

With reference to the already addressed comparability of financial statements, the relevance of financial statements as well as audit reports for financial statement users (financial markets and investors) has to be highlighted.\(^{76}\)

Taking a look at Figure 23 one can identify the top words *report, opinion, company* and *firm* appearing from over 6,900 to over 15,200 times in 136 articles. These top words can be associated with the word combinations financial reports of companies/firms, audit reports and audit opinion. Further important top words are *statement, standard, ifr* and *legal* representing the thematization of accounting and auditing standards, financial statements, IFRS standards and the legal enforcement of standards. The quoted top words appear from over 2,800 to over 4,800 times. The top word *tax* may seem confusing in the context of this rather general topic but represents discussions about tax services provided by the auditor which may diminish tax avoidance strategies by management and therefore improve the quality of audited financial reports. The articles allocated to this topic **audit report, financial standards, financial statements and its users** more specifically discuss IFRS adoption aspects, various types of accounting standards, various types of reports (sustainability, modified, voluntary, mandatory, etc.), global audit firm networks, audit and accounting harmonization issues and different formats of audit reports.

\(^{75}\) Cf. PwC (2018): CFOdirect

4.1.3.12 International regulations, regulatory reviews and SOX

The accounting scandals in the early 2000s led to new laws in the US in order to improve financial reporting and disclosure as well as corporate governance. The Sarbanes-Oxley Act was an important part of the changes, requiring firms and auditors to provide an appropriate internal control structure and procedure for financial reporting including assessments of this structure and procedure. Moreover, SOX resulted in a major structural change of the audit industry, as the new requirements to the audit process have led to increased audit hours, audit cost and subsequently higher audit fees for clients. In addition, auditors are prohibited from providing clients with certain non-audit services. Another requirement originating from the SOX passage is the audit firm inspection by the Public Company Accounting Oversight Board (further referred to as PCAOB), which is an independent non-profit organization regulating auditors of publicly traded companies in order to assure the reliability of audits and to minimize audit risk. Before these radical changes, auditors had only been subject to self-regulation by peer-reviewing of their audits. Just like there were critical voices with regard to self-regulation of

---

auditors, there are now critics who doubt the sufficiency of the PCAOB inspector’s expertise to regulate audit firms properly. 

Reflecting on Figure 24 the top words *sox, pcaob, inspect, disclosur and control* – appearing between over 6,500 and over 14,900 times in a collective manner in the 123 articles – stand out and can be linked to the key matter of the SOX passage and the related PCAOB inspections in order to control financial reporting, disclosure and audits. As there are rather general top words like *report, firm and compani* within the top words of the topic category *international regulations, regulatory reviews and SOX*, it is again of utmost importance to consider them in collective appearance with other, more characteristic top words. The top word *qualiti* indicates the frequent discussion of the impact of international regulations and SOX on audit quality. Articles assigned to this topic category more specifically deal with subject matters such as costs associated with compliance to SOX, proposed regulatory changes to audit reporting standards, the effectiveness of peer reviews, PCAOB inspection control and effectiveness, PCAOB and SOX impacts on audit quality and internal audit, auditor attestation under SOX and PCAOB exemptions.

**Figure 24: Top words International regulations, regulatory reviews and SOX (own compilation)**

4.2 Typology of audit research topics with keywords

Resulting from the topic classification results and its top words per topic, a typology of the audit research topics was generated. This typology indicates the top word stems per topic, which are subsequently transitioned into full words representing the topic. The transitioned words represent keywords for every specific topic. Therefore, these keywords are indicative of the audit research topics, of course with the condition of appearing collectively with other keywords of the respective topic category. Table 2 presents the constructed typology with the top word stems as well as keywords per topic. In some cases, top words outside the top 10, which have not been included in the results chapter, have been included and transitioned into keywords. These included keywords are similarly essential for the topic categorization, but simply do not fall into the top 10 range. The generated typology demonstrates the most representative word stems and keywords for every topic of the audit research field falling under the defined criteria (time horizon 2002-2017, external audit, university access, journal ranking 2 or 3). With the help of this typology the great amount of specialized literature is roughly presented in one single framework.

As the RapidMiner topic model was set to produce an output of 12 topic categories, the following typology is a twelve-category typology. It represents a useful tool for researchers and students to categorize audit research papers.

Comparing this typology with the inductive typology of auditing research of LESAGE/WECHTLER79 of 2012, in the typology of this thesis the absence of the topic category education and tax audit can be noted. This is due to the fact that in the topic model of this thesis there is no single category for these aspects and the articles discussing tax audit are mostly allocated to the general topic audit report, financial standards, financial statements and its users. For the aspect of auditor education and training the RapidMiner topic model has not generated a single topic category, as the articles apparently did not cover it extensively enough. Another important difference is that the typology of Lesage surprisingly did not include the keywords SOX or PCAOB in the topic international regulation, whereas in our comparable topic category the words appear between over 6,000 and over 10,000 times. This fact is surprising, as SOX was already passed in 2002 and the paper of Lesage/Wechtler was published only ten years later.

<table>
<thead>
<tr>
<th>TOPIC</th>
<th>%</th>
<th>TOP WORD STEMS</th>
<th>KEYWORDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Auditor-client relationship and partner rotation</td>
<td>6,55%</td>
<td>partner; client; rotat; profession; engag; relationship; qualiti; pressur; mandator; independ</td>
<td>partner; client; relationship; rotation; rotate; engagement; quality; independence; independent; pressure; mandatory; time; firm; behaviour;</td>
</tr>
<tr>
<td>Going concern</td>
<td>5,78%</td>
<td>firm; opinion; report; concern; compani; client; influenc; effect; big; bankruptci</td>
<td>firm; opinion; report; reporting; concern, company; client; influence; effect; Big 4; bankruptcy; control;</td>
</tr>
<tr>
<td>Liability and litigation</td>
<td>4,31%</td>
<td>report; liabl; qualiti; client; neglig; insur; litig; legal; penalti; damag</td>
<td>report; reporting; liability; quality; client; negligence; insurance; litigation; legal; law; penalty; damage;</td>
</tr>
<tr>
<td>Audit risk and fraud</td>
<td>7,49%</td>
<td>risk; fraud; assess; materi; test; misstat; error; control; detect; evid</td>
<td>risk; fraud; assess; assessment; materiality; test; testing; misstate; misstatement; error; control; detect; detection; evidence; indicator;</td>
</tr>
<tr>
<td>Audit market and fees</td>
<td>11,15%</td>
<td>fee; client; firm; industri; big; servic; market; price; specialist; premium</td>
<td>fee; client; firm; industry; Big 4; non-audit; service; market; price; pricing; specialist; specialization; premium; competition;</td>
</tr>
<tr>
<td>Audit report, financial standards, financial statements and its users</td>
<td>8,02%</td>
<td>compani; report; tax; firm; opinion; bank; statement; standard; ifr; legal</td>
<td>company; report; reporting; tax; firm; opinion; bank; statement; standard; IFRS; legal; law; adopt; market; disclosure; disclose;</td>
</tr>
<tr>
<td>Auditor's judgement</td>
<td>11,09%</td>
<td>review; judgement; experi; perform; task; decis; evalu; knowldg; assess; team</td>
<td>review; judgement; judgment; experience; perform; performance; task; decision; evaluate; evaluation; knowledge; assess; assessment; test; testing; team; interact; behaviour;</td>
</tr>
<tr>
<td>Category</td>
<td>Percentage</td>
<td>Relevant Terms</td>
<td></td>
</tr>
<tr>
<td>-------------------------------------------------------------</td>
<td>------------</td>
<td>-------------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>Audit quality and auditor tenure</td>
<td>10.74%</td>
<td>firm; earn; quality; accrual; sample; Big 4; client; estimate; estimation; report; tenure;</td>
<td></td>
</tr>
<tr>
<td>Corporate governance</td>
<td>5.25%</td>
<td>committee; director; board; report; CEO; corpor; expert; govern; control; shareholder;</td>
<td></td>
</tr>
<tr>
<td>Audit profession</td>
<td>11.80%</td>
<td>profession; report; profess; govern; process; social; perform; time; critic;</td>
<td></td>
</tr>
<tr>
<td>Audit procedures</td>
<td>10.56%</td>
<td>system; control; extern; process; perform; technology; service; model; standard; time;</td>
<td></td>
</tr>
<tr>
<td>International regulation, regulatory reviews and SOX</td>
<td>7.26%</td>
<td>report; firm; control; PCAOB; inspect; inspection; company; restate; restatement; quality; disclose; disclosure; SOX; deficiencies; opinion; review; international; harmonization;</td>
<td></td>
</tr>
</tbody>
</table>
5 Discussion

5.1 Conclusion

In conclusion it can be said that the RapidMiner classification model developed in this thesis, with the help of the Extract Topics from Documents operator, delivers exceptionally excellent and reliable results for broad topic classifications. Almost automatically, it provides a vast overview of topics in large document collections and even lets the user extract the results in various data formats for further analysis. Therefore, the results cannot only be visualized in suitable ways but can also easily be examined more closely. This possibility of a detailed investigation refers to the topic probabilities/topic confidence values, which are quoted for every single article and describe the extent to which every article also discusses topic categories other than the one topic category it is allocated to (detailed explanation in chapter 6.1.2).

Reviewing existing academic literature is a complex and time-consuming task but getting a general overview of topics can certainly be facilitated with the text mining process of this thesis. Thus, the categorization process of systematic literature reviews (cf. chapter 2.1 - stage 3 of the review process) can be greatly simplified. The designed process can be used for a new classification of audit research in the future by updating the database, running the process again and evaluating the results. Furthermore, the process can be used for the topic classification of other academic research fields. However, in this case the parameters have to be set according to the respective research field and the Stopwords Dictionary has to be adapted, in order to obtain correct results.

5.2 Limitations

Naturally, every research paper has its limitations. In the case of this thesis, the restriction of the topic model is the final allocation of the articles to one single topic category. The documents of the database have to be assigned to one single topic in order to make it possible to draw conclusions and be able to analyze the composition of the audit research field. However, the handling of this matter is not that simple. Audit research papers largely do not deal with only one topic. Often, they are a mix of several topics, or analyze

---

80 The process is published at the RapidMiner Community Sample Repository under the following link: https://community.rapidminer.com
By using the hyperlink at the website, the RapidMiner Studio is opened automatically, the process is directly created and ready to use.
Alternatively, the XML code of Annex 3 can be copied and inserted into the XML code section in the RapidMiner Studio, in order to use the designed process.
the effect of one topic on another. The overlapping of topics as a result cannot be negated. The constructed model represents one approach to deal with this issue. It assigns every article to the most debated topic and the user is then able to examine the mix of topics of every single article in more detail. The shortcoming of this approach is that the topic allocation can be slightly distorted by the top words approach. For instance, when an article analyzes the relationship between auditor tenure and audit quality by using the going concern opinion issuance as an indicator (see table 1 – ex. 1), the article is allocated to the going concern topic if the content of the article mainly contains top words of the going concern topic. This applies for this example, as the greatest part of the article is the regression model with all its explanations, which deals with going concern opinions of various firms included in the sample. This allocation to the going concern topic is not incorrect, however, if a person would categorize the article manually, the article would go straight into the topic category audit quality and auditor tenure, as the abstract is referring to this topic as the research focus. Using the RapidMiner classification model, the entire text of the document, and not only the intended research outcome which predomnates the abstract, is considered. This leads the downside of possibly creating distortions by not identifying the subjectively most essential topic because the topic which is covered the most is chosen as a topic category. Nevertheless, it carries the advantage of not disregarding certain topics within the articles, which are discussed extensively but are not in the abstract or listed as a primary research question. Because of the mixture of topics in the majority of the articles a clear delimitation is simply not possible.

5.3 Recent trends

The main recent trend in audit research publications of the last decade is clearly SOX, along with the related PCAOB inspections. While the peer reviewing effectiveness by auditor self-regulation has always been discussed and still is, the major part of recent academic literature on international audit regulation are the debates over the SOX passage, the PCAOB and its impact, effectiveness, consequences, enforcements, and disciplinary actions. Additionally, it is not an isolated topic, as the regulations affect various other audit matters. The issue of SOX increasing audit costs and subsequently audit fees is also heavily discussed, which causes an overlap with the topic audit market and fees. Another topic, which is many times discussed along with SOX is audit quality, seeing as apart from minimizing audit risk, a major aim of SOX is the improvement of audit quality.
Another recent trend in audit research are the increasingly addressed innovative audit procedures and processes due to the increasing electronization and digitalization of the audit environment. Due to globalization and rapid technological advances, matters like online auditing, auditing software, analytical tools for the analysis of large data volumes, information processing possibilities and weaknesses as well as information systems have been discussed heavily in recent academic audit literature. The research covering audit procedures is often linked to discussions about the audit profession, as the topic audit profession concerns the performance of the audit by the auditor which requires audit procedures to be performed. Furthermore, audit procedures are often discussed along with the topic auditor’s judgement, since audit procedures lay the foundation for auditor decision making and judgement.

5.4 Further research options and needs

As pointed out in the previous chapter, text mining has its limitations. An option for further research would be applying the underlying classification model on the abstracts of the articles and to then compare the results with the classification results of this thesis. Another possibility would be to apply artificial intelligence machine learning applications on the database for classification purposes.
List of References


Annex 1: List of included journals

<table>
<thead>
<tr>
<th></th>
<th>Journal Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Accounting Review</td>
</tr>
<tr>
<td>2</td>
<td>Journal of Accounting Research</td>
</tr>
<tr>
<td>3</td>
<td>Accounting, Organizations and Society</td>
</tr>
<tr>
<td>4</td>
<td>Journal of Accounting and Economics</td>
</tr>
<tr>
<td>5</td>
<td>Contemporary Accounting Research</td>
</tr>
<tr>
<td>6</td>
<td>Review of Accounting Studies</td>
</tr>
<tr>
<td>7</td>
<td>Abacus</td>
</tr>
<tr>
<td>8</td>
<td>Accounting and Business Research</td>
</tr>
<tr>
<td>9</td>
<td>Accounting Forum</td>
</tr>
<tr>
<td>10</td>
<td>Accounting Horizons</td>
</tr>
<tr>
<td>11</td>
<td>Auditing: A Journal of Practice and Theroy</td>
</tr>
<tr>
<td>12</td>
<td>Behavioral Research in Accounting</td>
</tr>
<tr>
<td>13</td>
<td>Journal of Accounting, Auditing and Finance</td>
</tr>
<tr>
<td>14</td>
<td>Journal of International Accounting, Auditing and Taxation</td>
</tr>
<tr>
<td>15</td>
<td>British Accounting Review</td>
</tr>
<tr>
<td>16</td>
<td>European Accounting Review</td>
</tr>
<tr>
<td>17</td>
<td>Accounting, Auditing and Accountability Journal</td>
</tr>
<tr>
<td>18</td>
<td>Critical Perspectives on Accounting</td>
</tr>
<tr>
<td>19</td>
<td>International Journal of Accounting</td>
</tr>
<tr>
<td>20</td>
<td>Journal of Accounting Literature</td>
</tr>
<tr>
<td>21</td>
<td>Journal of Accounting and Public Policy</td>
</tr>
<tr>
<td>22</td>
<td>Journal of Business Finance and Accounting</td>
</tr>
<tr>
<td>23</td>
<td>International Journal of Auditing</td>
</tr>
<tr>
<td>24</td>
<td>Current Issues in Auditing</td>
</tr>
<tr>
<td>25</td>
<td>Journal of International Financial Management and Accounting</td>
</tr>
<tr>
<td>26</td>
<td>Managerial Auditing Journal</td>
</tr>
<tr>
<td>27</td>
<td>Advances in Accounting</td>
</tr>
<tr>
<td>28</td>
<td>International Journal of Accounting Information Systems</td>
</tr>
</tbody>
</table>
## Annex 2: Stopwords Dictionary

<table>
<thead>
<tr>
<th>account</th>
<th>higher</th>
<th>rate</th>
<th>regim</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>agress</td>
<td>iaf</td>
<td>receiv</td>
<td>iia</td>
<td></td>
</tr>
<tr>
<td>april</td>
<td>icreas</td>
<td>relat</td>
<td>interviewe</td>
<td>see</td>
</tr>
<tr>
<td>arl</td>
<td>individu</td>
<td>require</td>
<td></td>
<td></td>
</tr>
<tr>
<td>asset</td>
<td>influenc</td>
<td>research</td>
<td></td>
<td></td>
</tr>
<tr>
<td>assur</td>
<td>inform</td>
<td>respond</td>
<td></td>
<td></td>
</tr>
<tr>
<td>audit</td>
<td>interest</td>
<td>respons</td>
<td></td>
<td></td>
</tr>
<tr>
<td>auditor</td>
<td>intern</td>
<td>result</td>
<td></td>
<td></td>
</tr>
<tr>
<td>base</td>
<td>interview</td>
<td>role</td>
<td></td>
<td></td>
</tr>
<tr>
<td>because</td>
<td>issu</td>
<td>sec</td>
<td></td>
<td></td>
</tr>
<tr>
<td>busi</td>
<td>johann</td>
<td>set</td>
<td></td>
<td></td>
</tr>
<tr>
<td>car</td>
<td>journal</td>
<td>size</td>
<td></td>
<td></td>
</tr>
<tr>
<td>case</td>
<td>juror</td>
<td>societi</td>
<td></td>
<td></td>
</tr>
<tr>
<td>chang</td>
<td>kepler</td>
<td>state</td>
<td></td>
<td></td>
</tr>
<tr>
<td>choice</td>
<td>les</td>
<td>studi</td>
<td></td>
<td></td>
</tr>
<tr>
<td>continu</td>
<td>level</td>
<td>survey</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cost</td>
<td>linz</td>
<td>tabl</td>
<td></td>
<td></td>
</tr>
<tr>
<td>data</td>
<td>list</td>
<td>total</td>
<td></td>
<td></td>
</tr>
<tr>
<td>des</td>
<td>loan</td>
<td>type</td>
<td></td>
<td></td>
</tr>
<tr>
<td>develop</td>
<td>make</td>
<td>universitaet</td>
<td></td>
<td></td>
</tr>
<tr>
<td>differ</td>
<td>manag</td>
<td>use</td>
<td></td>
<td></td>
</tr>
<tr>
<td>download</td>
<td>member</td>
<td>valu</td>
<td></td>
<td></td>
</tr>
<tr>
<td>earn</td>
<td>nas</td>
<td>variabl</td>
<td></td>
<td></td>
</tr>
<tr>
<td>effect</td>
<td>non</td>
<td>vol</td>
<td></td>
<td></td>
</tr>
<tr>
<td>effort</td>
<td>particip</td>
<td>weak</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ethic</td>
<td>perceive</td>
<td>work</td>
<td></td>
<td></td>
</tr>
<tr>
<td>expect</td>
<td>plan</td>
<td>year</td>
<td></td>
<td></td>
</tr>
<tr>
<td>factor</td>
<td>posit</td>
<td>implement</td>
<td></td>
<td></td>
</tr>
<tr>
<td>finance</td>
<td>practic</td>
<td>condit</td>
<td></td>
<td></td>
</tr>
<tr>
<td>first</td>
<td>probabl</td>
<td>famili</td>
<td></td>
<td></td>
</tr>
<tr>
<td>function</td>
<td>provid</td>
<td>share</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Annex 3: XML code RapidMiner process

```xml
<?xml version="1.0" encoding="UTF-8"?><process version="9.0.003">
<context>
<input/>
<output/>
_macros/>
</context>
<operator activated="true" class="process" compatibility="9.0.003" expanded="true" name="Process">
<process expanded="true">
<operator activated="true" class="read_excel" compatibility="9.0.003" expanded="true" height="68" name="Read Excel" width="90" x="112" y="238">
<parameter key="excel_file" value="/Users/ninaploetzl/Dropbox/JKU/5. Semester/Masterarbeit/RapidMiner/stop words list dictionary.xlsx"/>
<parameter key="first_row_as_names" value="false"/>
<list key="annotations"/>
<list key="data_set_meta_data_information">
<parameter key="0" value="word.true.polynominal.attribute"/>
</list>
</operator>
</operator>
<operator activated="true" class="concurrency:loop_files" compatibility="9.0.003" expanded="true" height="82" name="Loop Files" width="90" x="179" y="85">
<parameter key="directory" value="/Users/ninaploetzl/Documents/txt Dateien konvertiert gesammelt"/>
<parameter key="enable_macros" value="true"/>
</operator>
<operator activated="true" class="text:read_document" compatibility="8.1.000" expanded="true" height="68" name="Read Document" width="90" x="112" y="187"/>
<operator activated="true" class="text:tokenize" compatibility="8.1.000" expanded="true" height="68" name="Tokenize" width="90" x="246" y="187"/>
<operator activated="true" class="text:stem_snowball" compatibility="8.1.000" expanded="true" height="68" name="Stem (Snowball)" width="90" x="380" y="187"/>
<operator activated="true" class="text:filter_stopwords_english" compatibility="8.1.000" expanded="true" height="68" name="Filter Stopwords (English)" width="90" x="514" y="187"/>
<operator activated="true" class="operator_toolbox:filter_tokens_using_exampleset" compatibility="1.7.000" expanded="true" height="82" name="Filter Tokens Using ExampleSet" width="90" x="447" y="289"/>
<connect from_port="file object" to_op="Read Document" to_port="file"/>
</process>
</process>
```