



Exploring ways for a better interpretation of topic models

By
Sourav K Singh
Supervised by Professor Stoyan Tanev

A research project submitted in partial fulfilment of the requirements for the degree in
Master of Applied Business Analytics
In
Technology Innovation Management

Carleton University
Ottawa, Ontario
© 2023

Sourav K Singh

Abstract

Topic modelling is a technique used to automatically identify the underlying topics in a large collection of unstructured text data, such as news articles, social media posts, or customer reviews. This method has gained importance due to the increasing volume of unstructured data available on the internet. Topic modelling helps to identify the main themes or topics present in the text data without the need for manual reading and categorization. This can help to gain insights into the key issues or concerns that people are discussing, identify patterns in the data, and make data-driven decisions based on the insights obtained from the text data. The applications of topic modelling are wide-ranging, including social media analysis, market research, content recommendation, and customer feedback analysis. Overall, topic modelling is a valuable tool for gaining insights from large volumes of text data and can lead to more informed decision-making in various industries, but at the same time, it can be a complex process that requires domain knowledge, research, and additional text analytic tools. To extract meaningful insights from topic modelling, it is important to identify the main themes or topics in the text data and understand the nuances and relationships between them.

This project will explore processes that combine the results of the LDA topic model with automated topic labeling and text summarization algorithms to achieve a better interpretation of the topic model results. The process of combining topic modelling with text summarization can be challenging, as it requires the selection of appropriate summarization techniques and careful consideration of the impact of summarization on the underlying topics. Hence this project will also experiment with different approaches for selecting the associated documents within the topic, in order to identify the best possible ways of choosing the document for a better topic model interpretation.

Acknowledgement

I wish to express my utmost gratitude to Professor Stoyan Tanev, who was my supervisor throughout this project, for his unwavering faith in my abilities and continuous support. His exceptional attention to detail, dedication, and provision of the best possible assistance to all the students in the TIM program has been a constant source of inspiration to me. I am thankful for the knowledge and skills that I acquired during this journey under his mentorship.

Furthermore, I am deeply grateful to Professor Stoyan for providing me with the essential resources and data needed to conduct this project. His invaluable guidance and support were instrumental in helping me achieve my objectives.

I would also like to extend my appreciation to Professor Tony Bailetti and Michael Weiss for their invaluable support and guidance, particularly in the area of Chat GPT. Their knowledge-sharing session was instrumental in helping me successfully complete this project. Their contributions and dedication to promoting research in this field are highly commendable.

In addition, I would like to express my gratitude to Professors Steven Muegge, Mika Westerlund, and Eduardo Bailetti for their invaluable mentorship and guidance throughout my journey in the TIM program. Their expertise and support have greatly contributed to my growth and development in this field.

Lastly, I would like to thank the TIM faculty and administration for creating and providing such an exceptional practice-based program. I am also grateful to my fellow TIM students and friends for their unwavering support and assistance throughout my learning journey.

TABLE OF CONTENTS:

ABSTRACT.....	2
ACKNOWLEDGEMENT	3
1 INTRODUCTION.....	9
1.1 OBJECTIVE	11
1.2 DELIVERABLES	11
1.3 RELEVANCE.....	11
1.4 WHAT IS KNOWN?	11
1.5 WHAT IS UNKNOWN?	13
1.6 CONTRIBUTION.....	14
1.7 TOPIC LABEL INTERPRETATION METHOD.....	15
1.8 ORGANIZATION OF THE DOCUMENT.....	16
2 LITERATURE REVIEW	18
2.1 STREAM 1: TALENT MANAGEMENT.....	18
2.2 STREAM 2: TOPIC MODELLING APPROACH.....	19
2.3 STREAM 3: TOPIC LABELING AND INTERPRETATION	21
2.4 STREAM 4: TEXT SUMMARIZATION TECHNIQUES	25
1.1 STREAM 5: CHAT GPT	27
1.2 LESSONS LEARNED FROM THE LITERATURE.....	28
1.3 SUMMARY OF THIS CHAPTER	30
3 RESEARCH DESIGN AND METHOD	31
3.1 RESEARCH METHOD.....	31
3.2 TOPIC LABEL INTERPRETATION FRAMEWORK	33
3.3 IDENTIFIED LIST OF COMPANIES WITHIN THE TALENT MANAGEMENT DOMAIN.....	36
3.4 DATA COLLECTION.....	39
3.5 DATA ANALYSIS.....	41
3.6 IMPLEMENT AND APPLY THE TOPIC MODEL	43
<i>3.6.1 Identify the optimal number of topics using coherence score</i>	43
<i>3.6.2 Topic model.....</i>	45
<i>3.6.3 Data visualization</i>	45
3.7 COMPANIES REPRESENTING EACH TOPIC.....	48
<i>3.7.1 Cutting off document based on the topic percentage contribution</i>	49
3.8 SUMMARY OF THIS CHAPTER	50
4 SUMMARY OF RESULTS	51
4.1 BUILDING THE CORPUS OF TEXT DOCUMENTS.....	51
4.2 THE TOPIC MODEL.....	53
<i>4.2.1 Approach 1: Choosing the most associated documents from each company related to a topic</i>	58
4.2.1.1 Grouping companies based on the generated topic label.....	60
4.2.1.2 Automated multi-document text summarization for Approach 1.....	67
4.2.1.3 Incorporating Chat GPT with Approach: 1	69
<i>4.2.2 Approach 2: Choosing the 50 most associated documents from each company within a topic</i>	71
4.2.2.1 Grouping companies based on the generated label	73
4.2.2.2 Automated multi-document text summarization for Approach 2	76
4.2.2.3 Incorporating Chat GPT with Approach: 2	78
<i>4.2.3 Approach 3: Choosing all the documents from each company within a topic</i>	79
4.3 SUMMARY OF THIS CHAPTER	81

5 ANALYSIS AND DISCUSSION	82
5.1 PROBLEM AND GAP ADDRESSED	82
5.2 INTERPRETATION OF THE RESULTS.....	82
5.3 COMPARISON BETWEEN THE DIFFERENT APPROACHES USED FOR THE TOPIC INTERPRETATION.....	83
5.3.1 Approach 1.....	83
5.3.2 Approach 2.....	84
5.3.3 Approach 3.....	85
5.4 SUMMARY OF THE TOPIC MODEL INTERPRETATION APPROACHES.....	86
5.5 LINK TO THE LITERATURE.....	87
5.6 SUMMARY OF THIS CHAPTER	87
6 CONCLUSIONS, LIMITATIONS, AND FUTURE RESEARCH.....	89
6.1 CONCLUSIONS	89
6.2 LIMITATIONS	89
6.3 FUTURE RESEARCH	90
REFERENCE	92
APPENDIX	96
PYTHON LIBRARIES USED IN THIS PROJECT	96
PYTHON CODE FOR WEB SCRAPER	97
PYTHON CODE FOR STREAMLIT HOME PAGE DESIGN AND TEXT CLEANING.....	98
PYTHON CODE FOR TOPIC MODEL.....	98
PYTHON CODE FOR DATA VISUALIZATION	99
PYTHON CODE FOR GENERATING TOPIC LABEL.....	99
PYTHON CODE FOR TEXT SUMMARIZATION.....	100
PYTHON CODE FOR PLOTTING TF-IDF GRAPH	100
PYTHON CODE FOR PLOTTING WORD DISTRIBUTION	101
STOPWORDS TXT FILE SAMPLE	101
MULTI WORDS TXT FILE SAMPLE	102

LIST OF FIGURES

FIGURE 1: LDA TOPIC MODELING (SINHA, 2020)	20
FIGURE 2: TOPIC LABEL INTERPRETATION FRAMEWORK.....	33
FIGURE 3: DATA SCRAPING TOOL.....	40
FIGURE 4: SAMPLE OF SCRAPED DATA	41
FIGURE 5: DATA CLEANING PROCESS.....	43
FIGURE 6: TOPIC MODEL COHERENCE GRAPH.....	44
FIGURE 7: WORD CLOUD.....	46
FIGURE 8: INTER-TOPIC VISUALIZATION MAP.....	47
FIGURE 9: SAMPLE TOPIC DOCUMENTS	48
FIGURE 10: TOPIC DOCUMENTS THRESHOLD	50
FIGURE 11: CLEANED TOKENIZED CORPUS	52
FIGURE 12: MOST FREQUENT WORDS IN THE CORPUS.....	53
FIGURE 13: TOPIC MODEL RESULT	54
FIGURE 14: TOPIC INTERPRETATION: APPROACH 1	58
FIGURE 15: AUTOMATIC TOPIC LABEL GENERATION	59
FIGURE 16: AUTOMATIC TOPIC LABEL GENERATION (CONTD...).....	60
FIGURE 17: COMPANIES GROUPED PER TOPIC LABEL	61
FIGURE 18: COMPANIES GROUPED ON TOPIC	62
FIGURE 19: COMPANIES GROUPED ON TOPIC	64
FIGURE 20: COMPANIES GROUPED ON TOPIC	66
FIGURE 21: SAMPLE OF MOST ASSOCIATED COMPANY DOCUMENT.....	68
FIGURE 22: SHORT SUMMARY OF TOPIC 1: MOST ASSOCIATED DOC FROM EACH COMPANY	68
FIGURE 23: INSIGHTS FROM CHAT GPT FOR TOPIC 1.....	70
FIGURE 24: TOPIC INTERPRETATION: APPROACH 2.....	71
FIGURE 25: GENERATED TOPIC LABEL: APPROACH 2.....	72
FIGURE 26: GENERATED TOPIC LABEL: APPROACH 2 (CONTD...)	73
FIGURE 27: COMPANIES GROUPED BASED ON THE TOPIC LABEL: APPROACH 2.....	74
FIGURE 28: COMPANIES REPRESENTING TOPIC: ONLINE CODING TEST	75
FIGURE 29: SAMPLE OF THE TOP 50 ASSOCIATED COMPANY DOCUMENT.....	77
FIGURE 30: SHORT SUMMARY: TOP 50 ASSOCIATED DOCS FROM EACH COMPANY	77
FIGURE 31: CHAT GPT INSIGHTS FOR THE TOPIC	79
FIGURE 32: COMPANIES GROUPED ON A TOPIC	80

LIST OF TABLES

TABLE 1: RESEARCH METHODOLOGY	31
TABLE 2: LIST OF TALENT MANAGEMENT COMPANIES.....	39

1 Introduction

The rise in technology has led to the extensive use of Machine Learning (ML) in various industries. ML helps organizations perform tasks that require problem-solving, pattern identification, learning, reasoning, and planning, among others. One of the applications of ML that have become popular is text analytics, which helps organizations understand large amounts of data to extract meaningful insights.

The field of topic modeling has become increasingly important, as it is an unsupervised machine learning approach that organizes text information to identify related pieces of text. This can be achieved without a predetermined set of tags or manually categorized training data, making it an unsupervised learning method. Topic modeling can quickly and easily analyze data without the need for training, but the accuracy of the results cannot be guaranteed.

To comprehend the way in which the meaning or context of a word is influenced by the larger context within which it is used in natural language, several text analysis methods employ techniques known as "bag analysis group word analysis." These methods focus on examining groups of words collectively instead of counting them individually. One of these techniques is topic modelling, while others include cluster analysis and latent semantic analysis, all of which can be utilized to identify grouping in texts. Each of these methods offers valuable insights and information.

Topic modeling is a valuable tool for extracting insights and identifying patterns in large volumes of unstructured text data. It has the potential to provide information with a deeper understanding of the industry by identifying common themes and topics within any domain. This information can help organizations stay up to date with the latest trends and developments in the industry, and make informed decisions on how to best position their services. By using topic modelling, a firm can gain a competitive edge by having a better understanding of its market and its competitors. Additionally, topic modelling can help an organization identify new opportunities for growth and expansion, such as targeting specific industries or niches that are emerging within the talent management field.

Overall, topic modelling can be a powerful tool for CBDI to enhance their market intelligence and make data-driven decisions for the future of its business.

The Centre for Cross-border, Digital and Inclusive Entrepreneurship (CBDI) is an organization dedicated to helping businesses thrive in today's ever-changing economic landscape. With a focus on cross-border trade, digital innovation, and inclusivity, CBDI provides valuable consulting services to companies looking to develop effective strategies for growth and success. This project will introduce a more rigorous analytical approach to consulting, helping companies to identify and capitalize on new opportunities in today's dynamic business landscape. In today's data-driven world, businesses are increasingly turning to advanced analytic techniques to extract insights and gain a competitive edge. One area that has seen growth in recent years is text analytics, which involves the application of algorithms and statistical models to extract insights from unstructured text data. By applying advanced analytic techniques to the results of topic modelling, we can unlock valuable insights that might otherwise have gone unnoticed. Ultimately, this will help our client (the value proposition lab) to make more informed decisions, improve their operations, and stay ahead of the competition in today's rapidly evolving business landscape.

The main objective of this project is to improve the interpretation of topic labels by utilizing topic modelling techniques in combination with advanced text analytic algorithms such as grouping the topic documents based on companies, text-to-text transformers, LDA genism summarization and others. Combining these techniques aims to create a more comprehensive understanding of the topics and their underlying meaning. The use of text summarization algorithms will help to extract the most important information from the relevant documents, allowing us to identify the key themes and concepts associated with each topic. Clustering techniques will enable us to group related topics together, providing additional insights into the patterns and relationships that exist within the data. The resulting interpretation of topic labels will provide a deeper level of insight and understanding. Ultimately, this research project has the potential to significantly improve the effectiveness of topic modelling and text analytics and to provide valuable insights.

1.1 Objective

Applying text analytic algorithms on topic modelling results to extract meaningful insights from the extracted documents about the interpretation of the topics.

1.2 Deliverables

This TIM project has four deliverables,

- A corpus of text documents extracted from 75 talent management companies
- Topic modelling results on the corpus of scraped web pages from the above-selected Talent Management company's websites
- A process to interpret each of the topic models based on the summarization of the associated documents and the application of Chat GPT to derive summarized information from the generated topic model results
- A refined process that could generate additional topic interpretation insights based on the companies associated with each topic

1.3 Relevance

This project will help my client the Value Proposition Lab of the Centre for Cross-border, Digital and Inclusive Entrepreneurship to strengthen the analytical approach used to provide consultation to companies. This is a TIM Project supporting the needs of the Centre for CBDI entrepreneurship.

1.4 What is known?

Topic modeling is a powerful technique for identifying and extracting relevant topics from a large corpus of textual data. However, interpreting and labeling these topics can be challenging,

particularly when dealing with large and complex datasets. To address this issue, researchers have explored a variety of approaches that leverage different computational techniques, including LDA topic models, word vectors, letter trigram vectors, seq2seq algorithms, and multiple text summarization algorithms such as Sum Basic and MapReduce.

LDA (Latent Dirichlet Allocation) is a popular topic modeling algorithm that assigns each word in a corpus to a particular topic based on its probability of occurrence. This approach enables us to identify the most relevant topics in a given corpus, but it can be difficult to interpret the results without further analysis. To address this issue, researchers have explored techniques that incorporate additional information, such as word vectors and letter trigram vectors.

Word vectors are a type of representation that captures the semantic meaning of words based on their context within a given corpus. By using word vectors to analyze the output of an LDA model, researchers can identify the most relevant words for each topic and group them together based on their similarities. This approach enables us to create more accurate and meaningful labels for the topics identified by the model, which can help us to better understand the underlying patterns in the data.

Letter trigram vectors are another type of representation that captures the patterns of letter sequences within words. By using letter trigram vectors to analyze the output of an LDA model, researchers can identify the most relevant letter sequences for each topic and group them together based on their similarity. This approach can help us to identify more specific and nuanced topics, particularly in cases where the underlying patterns are more complex or difficult to discern.

Seq2seq algorithms are a type of neural network architecture that can be used to generate summaries of text documents. By training a seq2seq model on a corpus of text and then using it to summarize new documents, researchers can extract the most important information from a given text and use it to label the topics identified by an LDA model. This approach can help us to create more concise and informative labels for the topics, which can improve our understanding of the underlying patterns in the data.

Finally, multiple text summarization algorithms, such as Sum Basic and MapReduce, can be used to extract the most important information from a given text document. These algorithms work by identifying the most salient sentences or phrases in a text and summarizing them into a

shorter, more concise form. By using these algorithms to summarize the text associated with each topic identified by an LDA model, researchers can extract the most important information and use it to create more informative and accurate labels for the topics.

Overall, by combining the above techniques, we can create a comprehensive approach to topic modeling interpretation that enables us to identify and label topics automatically while extracting the most important information from relevant text documents. This approach can help us to create more accurate and meaningful labels for the topics identified by an LDA model, which can improve our understanding of the underlying patterns in the data and enable us to make more informed decisions based on the insights we gain.

1.5 What is unknown?

Interpreting topics from text documents extracted from company websites can be challenging due to the large amount of data and the need for accurate and meaningful interpretation. To address this issue, a systematic approach is needed that can combine multiple computational techniques, including topic modelling, word vectors, letter trigram vectors, seq2seq algorithms, and text summarization algorithms such as Sum Basic and MapReduce.

The first step in this approach would be to extract text documents from company websites and preprocess them to remove noise and irrelevant information. Next, topic modelling algorithms such as LDA can be used to identify the most relevant topics in the corpus. However, simply identifying topics is not enough; we need a process to interpret and label them accurately.

To achieve this, we can leverage the power of word vectors and letter trigram vectors to analyze the output of the topic modelling algorithms. Word vectors capture the semantic meaning of words based on their context within a given corpus, while letter trigram vectors capture the patterns of letter sequences within words. By using these techniques to analyze the topics identified by the model, we can identify the most relevant words and letter sequences for each topic and group them together based on their similarity.

Seq2seq algorithms can be used to generate summaries of the text documents associated with each topic. This can help us to extract the most important information and use it to label the

topics accurately. We can also use text summarization algorithms such as Sum Basic and MapReduce to extract the most salient sentences or phrases in the text documents and summarize them into a shorter, more concise form.

By combining these techniques in a systematic process, we can create a comprehensive approach to interpreting topics from text documents extracted from company websites. This approach can help us to identify and label topics automatically while extracting the most important information from relevant text documents. It can also improve our understanding of the underlying patterns in the data and enable us to make more informed decisions based on the insights we gain.

1.6 Contribution

The contribution of this project consists of the integration of multiple techniques and algorithms for the sake of performing an important task in the topic modeling process. By combining different techniques, we can improve the effectiveness and accuracy of topic modelling and interpretation, leading to better decision-making and insights for businesses and organizations across various industries.

One of the key benefits of this approach is its potential to significantly improve the efficiency of topic modelling. By automating the process of interpreting topics, we can save time and resources while improving the quality and accuracy of the results. This can be especially valuable for businesses and organizations that need to analyze large amounts of text data on a regular basis.

Furthermore, the use of this integrated approach can help to promote further research and development in the field of text analytics and natural language processing. As researchers and practitioners continue to explore the possibilities of combining different techniques and algorithms, we can expect to see new and innovative approaches to analyzing and interpreting text data.

In conclusion, the integration of multiple techniques and algorithms in this approach represents a significant advancement in the field of text analytics and natural language processing. By leveraging the power of topic modelling, word vectors, letter trigram vectors, seq2seq

algorithms, and text summarization algorithms such as Sum Basic and MapReduce, we can improve the efficiency and accuracy of our analyses and gain valuable insights into the underlying patterns in text data.

1.7 Topic label interpretation method

- Review the following literature streams:

- **Talent management:**

Talent management refers to the process of attracting, developing, and retaining skilled employees. It involves various HR practices such as recruiting training, performance management, and succession planning. Companies use talent management strategies to ensure they have a skilled workforce that can meet their business needs and drive growth.

- **Topic modeling approach:**

Topic modelling is a technique used to analyze large collections of text data and identify underlying themes or topics. It involves the use of machine learning algorithms to group words or phrases that frequently appear together in a given corpus of text. This can help to identify patterns and insights that might not be immediately obvious from the manual analysis.

- **Topic labeling and interpretation:**

Topic labelling refers to the process of assigning a name or label to a topic identified by a topic modelling algorithm. Interpretation involves analyzing the output of the algorithm to gain insights into the underlying patterns and themes in the data. Effective topic labelling and interpretation can help to extract meaningful insights from large volumes of text data.

- **Text summarization techniques:**

Text summarization involves the process of automatically generating a shorter version of a text document that retains the most important information. There are various techniques for text summarization, including extractive summarization, abstractive summarization, and query-based summarization. Text summarization can be a useful tool for identifying key

insights and patterns in large volumes of text data.

- **Chat GPT:**

Chat GPT is a language model developed by Open AI that uses deep learning algorithms to generate human-like responses to text-based questions or prompts. It has been used in various applications, including chatbots, virtual assistants, and customer service.

- Build a list of companies in the Talent Management domain
- Web scrap to create the corpus data of 75 Talent management companies
- Execute and utilize the topic model on the collection of documents generated by extracting data from the above companies' websites
- Implement a topic model interpretation based on the summarization of the most highly associated documents
- Incorporating insights from text documents of most highly associated companies to enhance the topic interpretation based on summarization
- Provide a refined interpretation integrating both types of interpretation and follow-up Chat GPT analysis
- Finally documenting the process and finalizing the result

1.8 Organization of the document

The project consists of six chapters that cover a range of topics related to Talent Management, Topic Modelling, Text Summarization, and topic label interpretation.

Chapter 1, "Introduction," provides an overview of the project's goals, deliverables, relevance, and contribution, as well as a summary of the methodology used. The chapter highlights what is known and unknown about the topic and outlines the problem addressed by the project.

Chapter 2, "Literature Review," provides a comprehensive discussion of the relevant literature streams. It covers Talent Management, Topic Modelling approach, Topic labelling and interpretation, Text Summarization technique, and ChatGPT. This chapter sets the foundation for the project and provides insights into the state of research on these topics.

Chapter 3, "Method," outlines the steps taken to complete the project. The chapter discusses the data collection, pre-processing, and modelling techniques used to create the Topic Model. It also discusses the evaluation criteria used to determine the best model.

Chapter 4, "Results," categorizes the results of the project's data analysis. The chapter presents the top topics generated by the Topic Model and the associated keywords. The chapter also includes a comparison of different models based on their evaluation metrics.

Chapter 5, "Analysis of Results," presents a detailed analysis of the results obtained in the project. The chapter discusses the problem addressed by the project, new insights discovered through research, and the relationship between the literature and the project's findings. The chapter also highlights the limitations of the study and provides suggestions for future research.

Finally, in **Chapter 6**, "Conclusions and Recommendations," the report concludes with a summary of the findings and recommendations for further study. The chapter presents the key takeaways from the project and discusses the implications of the findings for Talent Management. It also identifies areas for further research and suggests potential improvements to the methodology used in the project.

Overall, the project provides a valuable contribution to the field of Talent Management by applying Topic Modelling and Text Summarization techniques to analyze a large corpus of data. The project's results provide insights into the most important topics and themes in the field, which can help organizations make better decisions regarding Talent Management.

2 Literature review

It is important to have a thorough understanding of topic modeling to develop effective machine-learning models. Conducting a literature review can provide an overview of the various methods used to achieve topic modelling, which can serve as a basis for further study. With the right guidance and assistance, this understanding can be leveraged to benefit organizations such as CBDI by enabling them to gain insights from large amounts of unstructured data through the use of topic modelling techniques. By identifying clusters of related words and concepts, topic modelling can assist in discovering patterns and trends within the data that may not be immediately apparent through manual analysis. This can provide valuable insights for decision-making and can help CBDI stay ahead of the competition.

The literature review in this project covers five main topics: Talent Management, Topic Modelling Approach, Topic Labelling and Interpretation, Text Summarization Techniques, and Chat GPT. This review was conducted to gather relevant information and insights into these areas that could aid in the completion of the project. By exploring existing research and literature, the project aimed to build upon previous work and contribute to the understanding and development of these fields. Overall, the literature review serves as a foundation for the project, providing a framework for the methods used and the results obtained.

2.1 Stream 1: Talent management

In recent decades, companies have encountered difficulties in retaining valuable employees due to intense competition in the job market and a scarcity of talented individuals. As a result, organizational leaders have been compelled to enhance their human resource strategies. Instead of fostering talent inclusively, many organizations tend to rely on recruiting exclusive talent for development purposes (Kaliannan et al. 2023). However, the organization faces multiple challenges when it comes to developing and retaining talent in emerging market economies, these challenges include the lack of suitable infrastructure, limited investment in education and training, low wages, and intense competition for skilled workers.

Pereira et al. (2022) suggest that companies operating in these markets must address these challenges in order to effectively manage their talent and remain competitive. Additionally, they note that there are cultural and institutional factors that must be considered when

designing talent management strategies for emerging markets. Effective talent management is critical for achieving competitive advantage and improving organizational performance. Organizations should adopt a strategic and holistic approach to talent management, which involves identifying and nurturing talent, providing training and development opportunities, and offering competitive compensation packages (Anon 2018). Talent management is a complex process and organizations must be proactive and innovative in addressing the challenges associated with attracting and retaining talented employees (Bošjančič & Slana 2018).

2.2 Stream 2: Topic modelling approach

Topic modelling is an unsupervised machine learning technique that's capable of scanning a set of documents, detecting word and phrase patterns within them, and automatically clustering word groups and similar expressions that best characterize a set of documents.

The most commonly used methods for performing topic modeling are Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF).

LDA is a probabilistic model that assumes each document is a mixture of a small number of topics, and each topic is a probability distribution over words. LDA has been widely used in various fields such as natural language processing, computer vision, and social network analysis. It has been shown to be effective in discovering topics in large collections of text data. It is widely used to identify topics within a large corpus of text data and has already been applied in various fields, such as natural language processing and information retrieval (Jelodar et al. 2019). Before applying a Latent Dirichlet Allocation (LDA) model to a dataset, it is important to clean and preprocess the data to ensure that it is in a suitable format for analysis. This may involve tasks such as removing stop words, stemming, or lemmatizing words and converting text to lowercase. Proper cleaning and preprocessing of the dataset can improve the efficiency and accuracy of the LDA model and help to produce more meaningful results (Abdulla et al. 2021).

Alkhodair et al. (2018) discuss that while topic modeling is a powerful tool for analyzing large amounts of textual data, it can be challenging to interpret the results, especially in the case of short and informal text such as microblogs. To address this issue, the authors propose a novel

approach that combines topic modeling with sentiment analysis, which allows for a more nuanced understanding of the topics and themes present in microblog data. They demonstrate the effectiveness of this approach through experiments on a large dataset of microblogs from Twitter.

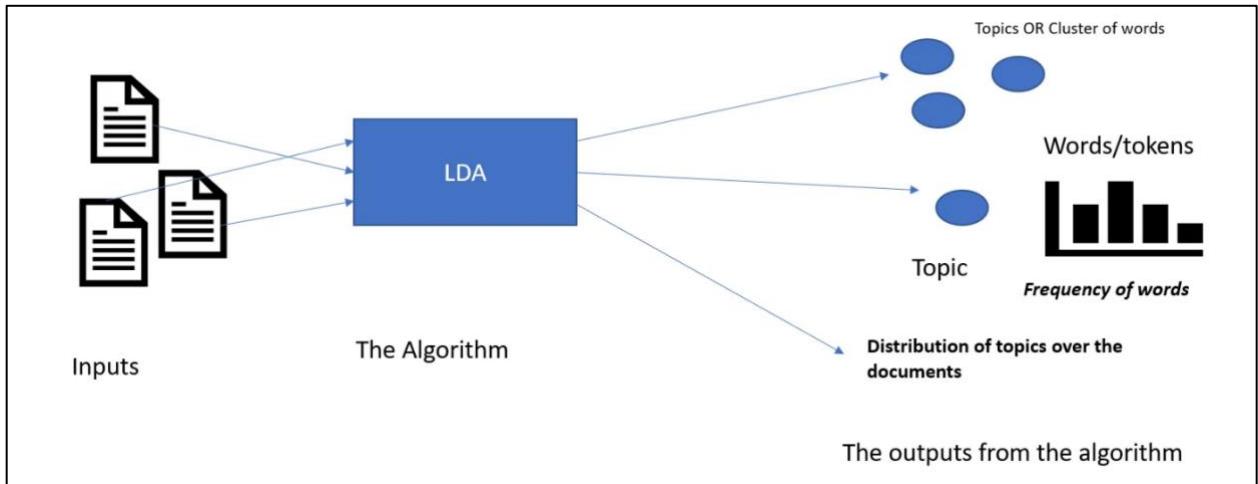


Figure 1: LDA topic modeling (Sinha, 2020)

NMF is a matrix factorization technique that seeks to find non-negative basis vectors that can be used to reconstruct the original data. NMF has been applied to various tasks such as image processing, speech analysis, and text mining. NMF is particularly useful when dealing with sparse data and has been shown to be effective in topic modeling.

Other methods for performing topic modeling include Probabilistic Latent Semantic Analysis (PLSA), Hierarchical Dirichlet Process (HDP), and Correlated Topic Model (CTM). These methods have their own strengths and weaknesses and are suitable for different types of data and applications. Therefore, it is important to choose the right method for the specific task at hand.

Limitations of existing topic modeling algorithms:

Topic modeling is a powerful tool for analyzing large amounts of text data, but it has its limitations. One of the primary challenges is that the generated topics can be difficult to interpret. While the algorithm generates a list of words associated with each topic, these words may not accurately capture the main topic of the document. As a result, it can be challenging to understand the true essence of the topic, and it may require manual effort to fully interpret the

results.

Another issue with topic modeling is that sometimes the generated topics may contain repeated words, which can make it difficult to determine which topic is most relevant to the document. This can also lead to confusion and a less accurate interpretation of the results.

Additionally, it's important to keep in mind that topic modeling results should not be solely relied upon to make crucial decisions. The results can give us an idea of the most common themes within the data, but they do not provide a complete understanding of the underlying meaning. This means that we need to exercise caution while interpreting the results and not rely solely on them to make important decisions.

Overall, while topic modeling is a powerful tool for analyzing text data, it should be used in conjunction with other methods and approaches to gain a more comprehensive understanding of the data. By being aware of the limitations and potential pitfalls of topic modeling, we can more effectively interpret the results and make more informed decisions.

2.3 Stream 3: Topic labeling and interpretation

Topic labeling refers to the process of assigning a descriptive label or name to a topic generated by a topic modeling algorithm. The label should accurately capture the main theme or concept represented by the topic, making it easier for users to understand and interpret the results. This labeling process can be done manually or through automated methods, such as using the most frequently occurring words in the topic or analyzing the associated documents. Effective topic labeling is important for making the results of topic modeling more accessible and useful for analysis and decision-making.

Topic labeling can be done manually by reviewing a sample of documents within a particular topic and then assigning a label that accurately reflects the underlying theme or subject matter of those documents. This process involves human judgement and expertise to determine the most appropriate label for each topic.

However, manual topic labeling can be a time-consuming and resource-intensive task, especially when dealing with large datasets. It is also subjective, meaning that different people

may assign different labels to the same topic based on their interpretation of the documents. This can result in inconsistent or unreliable labeling, leading to inaccurate analysis and decision-making.

To address these issues, automated methods have been developed for topic labeling using natural language processing and machine learning techniques. These methods can provide more efficient and consistent labeling of topics, reducing the manual effort required and increasing the reliability and accuracy of the analysis. Automated topic labeling or automated topic label generation can save time and resources compared to manual labelling, especially for large datasets. In addition, it can reduce potential biases introduced by human labelers and increase the consistency of labelling across different datasets. Automated topic labelling also provides a way to summarize the main themes and concepts present in large text datasets, which can be useful for tasks such as information retrieval and topic analysis. Automated labelling algorithms typically use statistical or machine learning techniques to assign labels to topics based on the most frequently occurring words or phrases within each topic.

There have been multiple approaches by which the labeling of topics has been performed by automatic approaches. One approach which is highly popular among all topic modeling methods is finding appropriate topic labels with the help of finding top high-ranking terms based on the highest probability of occurrence. In this, each term is assessed and compared against the highest-ranked term in the same topic by using coherence measures such as point-wise mutual information or other feature measure based on conditional probabilities. This paper uses multiple feature selection methods such as conditional probabilities, PMI, WordNet hypernym relations, the topic model word rank, and Pantel's distributional similarity score — as features for SVM ranking framework, and this produces a ranking of words for the most representative word or candidates that are selected by choosing the top-ranked words. This experiment demonstrates that selecting the topic word of the highest marginal probability gives including training the ranking model on labelled topics using other document collections produces better results. It predicts which words annotators tend to select as best words when presented with a list of ten words. Since annotators are not generally unanimous in their choice of best word, authors use ranking tasks, and treat the top-1, 2 and 3 words ranked by the system as the best words, and compare that to the top-1, 2 and 3 words chosen most frequently by annotators (Lau et al., 2010).

Mao et al. (2012) introduced a novel method that leverages the hierarchical relationships among topics for topic modeling. Unlike earlier approaches that treat each topic as an isolated entity, the proposed method aims to create a hierarchical structure of topics that is automatically labeled to aid in the interpretation of the topics. The method employs two algorithms that identify parent-child and sibling relations among topics to construct the hierarchical structure. The results showed that the accuracy of topic labeling increased significantly with an increase in the depth of the hierarchy. The proposed approach outperformed previous methods and demonstrated the usefulness of inter-topic relations in improving the accuracy of topic labeling. The hierarchical topic labeling method produced sensible outputs that facilitate effective interpretation of the topics.

Kou et al. (2016) proposed "Automatic Labelling of Topic Models Using Word Vectors and Letter Trigram Vectors" a new approach to automated topic labelling using both word vectors and letter trigram vectors. The proposed method aims to address the limitations of existing automated labelling techniques, such as the reliance on high-frequency words and the inability to capture the nuanced meaning of words. Their experimental results show that the proposed method outperforms existing automated labelling techniques in terms of accuracy and the ability to capture the nuanced meaning of words. Khan, Q and Chua, H N. (2021) demonstrated an automated topic labelling framework using zero-shot text classification. The proposed framework uses a pre-trained language model to classify topics based on their semantic similarity to a set of predefined topic labels. The authors compare their framework to several baseline methods and show that their approach outperforms the others in terms of accuracy and F1-score.

The framework is tested on a large dataset of news articles and shows promising results, indicating its potential usefulness in various applications, such as content analysis and information retrieval. The zero-shot classification utilizes transformer models. Transformer models are a type of neural network architecture that can be used for a variety of natural language processing tasks, including text classification. They are particularly effective at capturing contextual information and long-range dependencies in text data, which makes them well-suited for tasks like zero-shot text classification where the model needs to understand the context and meaning of words it has not seen before.

BERT, or Bidirectional Encoder Representations from Transformers, is a pre-trained deep learning model that can perform a variety of natural language processing tasks, including automated topic labeling. BERT uses a technique called masked language modeling to predict missing words in sentences, allowing it to understand the context and meaning of words within a sentence. There is a powerful pre-trained language model that is used for various natural language processing tasks, including topic modeling. One of the applications of BERT is automated topic labeling, which involves assigning appropriate labels to topics generated from a topic modeling algorithm (Tang et al., 2021). In automated topic labeling using BERT, the model learns to assign a label to a given topic by analyzing the text of the documents associated with that topic. BERT is trained on a massive amount of data and has a deep understanding of the relationships between words, making it highly accurate in identifying the topics discussed in a document.

To use BERT for automated topic labeling, first, the topic modeling algorithm generates a list of topics, usually in the form of a set of keywords or a list of related words. Then, BERT analyzes the text of the documents associated with each topic to determine the most appropriate label for that topic. Automated topic labeling using BERT has several benefits, including saving time and effort in manually labeling topics, increasing accuracy, and improving the interpretability of the results. By providing clear and informative labels for each topic, BERT allows users to understand the content of a large corpus of documents quickly and easily, enabling them to make informed decisions based on the insights gained from the analysis (Grootendorst, 2020).

Limitation of automated topic labeling:

One limitation of automated topic labeling is that it heavily relies on the quality of the input data. If the input data is noisy or contains irrelevant information, the automated topic labeling may produce inaccurate or misleading results. Automated topic modeling can be further affected by the quality and structure of the data. If the data is noisy or has a lot of irrelevant information, this can impact the accuracy of the generated topics.

Further, automated topic modeling can be sensitive to the parameters and settings used in the algorithm. If these are not set correctly, the resulting topics may not accurately represent the underlying data. Additionally, automated topic labeling may not be suitable for all types of

data, as some types of data may require more nuanced or subjective analysis (Lau et al., 2011).

Overall, while automated topic modeling can be a useful tool for analyzing large volumes of text data, it is important to be aware of its limitations and exercise caution when interpreting the results.

2.4 Stream 4: Text summarization techniques

Text summarization is the process of creating a shorter version of a text document while retaining its most important information. It involves analyzing a given text document and identifying the key points and ideas, and then presenting them in a condensed form. The goal of text summarization is to provide a quick and easy way to understand the main ideas of a lengthy text document without having to read the entire document. There are two main techniques of automated text summarization: extractive and abstractive. Extractive summarization involves selecting important sentences or phrases from the original text and using them to create a summary. This technique relies on natural language processing (NLP) algorithms to identify the most relevant and informative sentences in a document.

Abstractive summarization, on the other hand, involves generating new sentences that convey the same meaning as the original text. This technique uses more advanced NLP algorithms that are capable of understanding the context of the text and generating coherent summaries that may not be present in the original document.

Both techniques have their advantages and disadvantages. Extractive summarization is generally more accurate because it uses actual sentences from the original text, but it may not be able to capture the essence of the text as effectively as abstractive summarization. Abstractive summarization, on the other hand, can generate summaries that are more concise and readable, but it can also produce summaries that are less accurate or even misleading if the algorithms do not understand the text properly.

Twinandilla et al. (2018) proposed a multi-document text summarization method that utilizes both K-means clustering and Latent Dirichlet Allocation (LDA) to identify the most significant sentences from a set of documents. The approach involves first clustering similar sentences using K-means clustering, followed by topic modeling using LDA to identify the most relevant sentences from each cluster. The selected sentences are then combined to create

a summary of the key topics and themes present in the set of documents. The proposed method is evaluated using the ROUGE metric and compared to other multi-document summarization approaches, demonstrating its effectiveness in producing high-quality summaries.

Multi-document summarization is the process of creating a concise summary of multiple documents that captures the most important information from the original text. This is an important task in the field of natural language processing, as it can help individuals quickly understand the contents of large volumes of text.

Twinandilla et al. (2018) proposed method involves applying K-means clustering to group similar sentences together. K-means is a popular clustering algorithm that partitions data points into a specified number of clusters, with each cluster representing a group of data points that are similar to each other. In the context of multi-document summarization, K-means clustering is used to group similar sentences from different documents into clusters. After clustering, the authors apply LDA to each cluster to identify the most relevant sentences from each cluster. LDA is a topic modeling technique that can identify the underlying topics in a set of documents by identifying the words that are most likely to co-occur. In the context of multi-document summarization, LDA is used to identify the most relevant sentences from each cluster based on their topic. Finally, the selected sentences from each cluster are combined to create a summary of the key topics and themes present in the set of documents. The proposed method is evaluated using the ROUGE metric, which measures the overlap between the generated summary and the human-generated summary. The results demonstrate that the proposed method is effective in producing high-quality summaries.

Overall, the proposed method by Twinandilla et al. (2018) is effective in producing high-quality summaries and demonstrates the potential of combining different natural language processing techniques for text summarization.

Roul et al. (2018) also proposed a method combining Latent Dirichlet Allocation (LDA) with a sentence ranking algorithm to extract important sentences from multiple documents and generate a summary. The LDA model is used to identify the main topics in the documents, while the sentence ranking algorithm ranks the sentences based on their relevance to the identified topics. The proposed method is evaluated on a dataset of news articles, and the results show that it outperforms other state-of-the-art methods in terms of ROUGE scores,

which are commonly used to evaluate text summarization performance.

1.1 Stream 5: Chat GPT

Additionally, Chat GPT can assist in topic modelling by generating relevant keywords and topic suggestions based on a given dataset or text corpus. Chat GPT can also help in identifying patterns and relationships within the data that may be indicative of underlying topics or themes. Additionally, Chat GPT can be used to evaluate and refine the results of topic modelling algorithms by providing a human perspective on the relevance and accuracy of generated topics and their corresponding labels. Chat GPT can also suggest potential areas of improvement for the topic modelling process, such as refining data preprocessing methods or adjusting algorithm parameters.

In 2018, Anisin explored the use of GPT-3, a pre-trained deep learning model developed by OpenAI, for text summarization. GPT-3 is trained on a massive amount of data and can generate high-quality summaries that capture the essential information in a given text. However, to fine-tune GPT-3 for specific summarization tasks, a small dataset of examples is needed.

Fine-tuning GPT-3 for text summarization involves providing the model with a set of input-output pairs, where the input is a long text document and the output is a summary of the document. The model is then trained to predict the output given the input. During training, the model learns to identify the most important information in the document and generate a summary that captures that information.

One of the benefits of using GPT-3 for text summarization is its ability to generate summaries that are similar to those written by humans. The model is capable of producing human-like language and can generate summaries that are coherent, informative, and grammatically correct. Additionally, GPT-3 can handle different types of documents, including news articles, scientific papers, and social media posts.

Anisin (2018) also notes that GPT-3 can be used to generate summaries in different styles and tones. For example, the model can be fine-tuned to produce summaries that are formal or informal, objective or subjective, depending on the specific use case. This flexibility makes GPT-3 a powerful tool for a variety of applications, including news summarization, content

curation, and social media analysis.

However, there are some limitations to using GPT-3 for text summarization. One of the main challenges is the quality of the input data. Since GPT-3 is a deep learning model, it requires a large amount of high-quality data to perform well. If the input data is noisy or incomplete, the model may struggle to generate accurate summaries. Additionally, the model may produce biased or misleading summaries if the input data contains biased or misleading information. Another limitation is the complexity of the model. GPT-3 is a large and complex model with millions of parameters. As a result, it requires a significant number of computational resources to train and deploy. This can make it challenging to use GPT-3 in applications with limited resources, such as mobile devices or low-power computers.

Despite these limitations, the use of GPT-3 for text summarization has significant potential in various fields, including journalism, academia, and business. By providing high-quality summaries that capture the essential information in a given text, GPT-3 can save time and resources for organizations and individuals who need to quickly digest large amounts of information.

Chat GPT can potentially impact various aspects of academia and libraries, such as search and discovery, reference and information services, cataloguing and metadata generation, and content creation (Lund, B and Wang, T. 2023).

1.2 Lessons learned from the literature

From the literature, some of the key lessons learned includes:

- The importance of preprocessing and cleaning the data before running topic modelling algorithms, as this can significantly impact the accuracy and relevance of the generated topics.
- The need for careful selection and tuning of topic modelling algorithms and parameters, as different methods may be better suited for different types of data and contexts.

- The limitations of automated topic labeling may require manual review and refinement to ensure accuracy and relevance.
- The potential of combining multiple methods, such as clustering and topic modeling, is to improve the quality and interpretability of generated topics.
- The importance of evaluating and interpreting the generated topics in the context of the larger research question or application, rather than relying solely on automated outputs.
- In order for companies to realize their organizational visions, they need staff who are high-potential and looking toward the future
- The labor market is already reflecting a lack of high-quality human resources (HR), which inspires research into and planned management of high-potential, i.e., talented, employees
- It is important to understand the importance of talent management software and tools in emerging market economies
- Over the past decades, organizations have faced challenges in retaining good employees due to market competition and talent scarcity, thereby forcing leaders to improve their human resource strategies
- LDA is an example of a topic model and it builds a topic per document model and words per topic model, modelled as Dirichlet distributions
- The manual topic labelling needs considerable human labor to review the corpus's massive data
- Text-to-Text Transformer algorithm can be used for automated topic label generation
- The automated topic label uses text summarization techniques to identify the common theme between a list of documents
- Extractive summarization systems aim to extract salient snippets, sentences or

passages from documents

- Multi-document summarization is an automatic procedure that extracts important information from multiple documents
- LDA text summarization libraries can be used for extractive text summary from a corpus of documents
- Chat GPT can be a valuable tool in order comparing the obtained information on topic interpretation and can be useful in further extraction of meaningful topic label interpretation

1.3 Summary of this chapter

The literature review conducted for this project aimed to achieve the objective of finding a better way of interpreting topic labels. It was divided into five streams of study. The first stream involved acquiring domain knowledge and understanding the type of analysis that can be carried out within that domain. The second stream focused on the existing topic modeling algorithms and their application and limitations. The third and fourth streams aimed to understand additional algorithms such as automated topic labeling and text summarization, and how they can be combined with topic modeling to extract meaningful insights. Finally, in the last stream, the scope of Chat GPT was explored in terms of topic labeling interpretation and how it can contribute towards generating more meaningful insights from the topic model result.

In summary, the literature review covered a broad range of topics related to topic labeling and interpretation. It highlighted the importance of understanding domain-specific knowledge, the limitations of existing topic modeling algorithms, and the potential of incorporating additional algorithms such as automated topic labeling and text summarization to extract more meaningful insights. Additionally, it explored the potential of Chat GPT in facilitating topic labeling interpretation and providing a human perspective on the relevance and accuracy of generated topics and their corresponding labels. Overall, the literature

review provided valuable insights that can be used to improve the accuracy and effectiveness of topic labeling interpretation.

3 Research design and method

In this chapter, the research design and methodology used to achieve the goals set out in Chapter 1 are described. The specific activities involved in the research method are also explained, and the data that was collected and analyzed for the purpose of creating the deliverables are presented. The chapter concludes with a summary of its contents.

3.1 Research method

The purpose of this research project was to improve the interpretation of topic modelling results for Talent management companies. The project serves as an example of the methods used to interpret topic modelling results within the Talent Management domain. However, the approach used in this project is not limited to this domain and can be applied to any domain with good data. The project was carried out comprehensively, involving several steps to achieve the research objective.

Table 1: Research methodology.

Step	Activity was undertaken to produce deliverables	Outcome of the activity
1	Literature review on talent management software and companies	Understanding of the need for talent management tools
2	Build list of companies in Talent Management domain	Domain understanding and list of companies

3	Literature review on ways of improving the interpretation of topic modelling results	Understanding of the existing algorithms required to develop the process
4	Web scraping the data of 75 Talent management companies	A corpus of text docs for the topic model
5	Implement and apply topic model on the corpus of docs	List of Topics and associated ranked documents
6	Identify companies that are most representative of each given topic	Sorted company list based per topic
7	Implement a topic model interpretation based on the summarization of the most highly associated documents	Topic labels and brief descriptions uniquely identifying each topic
8	Incorporating insights from text documents of most highly associated companies to enhance the topic interpretation based on summarization	Summary of insights that enhance interpretation based on summarization
9	Provide a refined interpretation integrating both types of interpretation and follow-up Chat GPT analysis	Meaningful insights about a particular topic
10	Document the process and finalize project	A refined topic modelling

The initial stage involved conducting a thorough literature review on Talent Management software and companies, as well as exploring ways to improve the interpretation of topic modelling results. Next, web scraping was employed to collect data from 75 Talent Management companies. The topic model was implemented on the collected data, and the companies that were most representative of each given topic were identified. The resulting topic model interpretation was based on the summarization of the most highly associated documents. Further, insights from

the text documents of the most highly associated companies were incorporated to enhance the interpretation. The process culminated in a refined interpretation that integrated both types of interpretation, followed by a Chat GPT analysis to reinforce the findings. The entire process was comprehensively documented, and the project was finalized.

The individual activities are elaborated on in the next sections, and it is explained how the outcomes of these activities result in the project's deliverables.

3.2 Topic label interpretation framework

The project proposed a topic interpretation framework that utilizes the LDA topic model along with the LDA Genism summarization and Text-to-text transformer model for topic labelling.

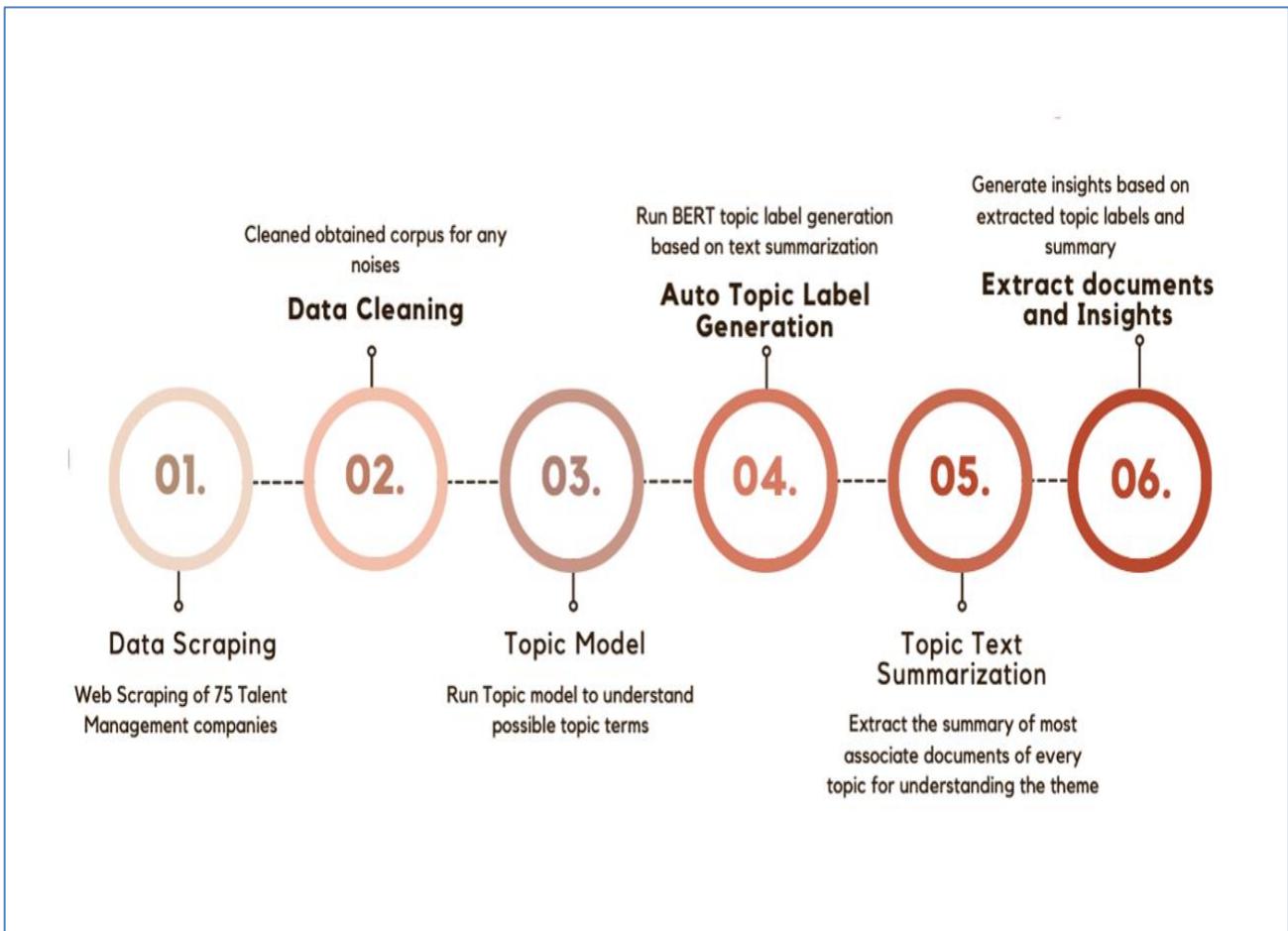


Figure 2: Topic label interpretation framework

The framework for the topic interpretation process begins with data scraping using an in-house web scraper. This process involves retrieving data from different sources such as online forums, blogs, social media, etc. After the data is scraped, it is sent for data cleaning to remove any noises including punctuation, invalid characters, and text lemmatizing. The cleaning process also includes removing stop words and processing multi-words to ensure that the data is accurate and clean.

The cleaned data will then be used as a corpus for the topic model. In the topic model, an optimum number of topics will be identified, and the model will produce the topic keywords with the stated number of topics. The topic modeling algorithm used in this framework is the Latent Dirichlet Allocation (LDA) algorithm, which is a generative statistical model. The LDA algorithm is commonly used in topic modeling tasks, and it assumes that each document is a mixture of a small number of topics, where each topic is represented by a distribution over words.

The LDA algorithm is used to extract meaningful insights from a corpus of documents by identifying the underlying topics in the text. The algorithm assigns each word in the document to a specific topic, and it does this by calculating the probability of a word being associated with a particular topic. The algorithm then groups together the words that have the highest probability of being associated with a specific topic to create a topic cluster. The LDA algorithm allows for a flexible approach to topic modeling, as it allows for a variable number of topics to be identified, and it can be used with different types of text data.

One of the limitations of the LDA algorithm is that it can be difficult to interpret the topics that are generated by the model. To overcome this limitation, the framework incorporates topic labelling and interpretation techniques to provide meaningful insights into the generated topics. The topic labelling process involves assigning a label or a name to each of the identified topics, which helps to provide a summary of the topic's main theme. The topic interpretation process involves analyzing the topic keywords and the associated documents to identify the key insights and trends in the data.

The framework also integrates text summarization techniques to extract meaningful data-

driven insights from the corpus. Text summarization involves creating a summary of the text by selecting the most important information from the document. The summarization process can be done using different techniques such as extractive summarization, where the summary is created by selecting the most important sentences from the document, or abstractive summarization, where the summary is created by generating new sentences that capture the essence of the document. The framework uses the extractive summarization technique to create summaries of the identified topics, which helps to provide a quick overview of the main insights.

The framework also incorporates the use of Chat GPT, a large language model, to analyze a large set of data and to model, label, and summarize the corpus to produce insights. Chat GPT is a state-of-the-art language model that is trained using deep learning techniques, and it is capable of generating human-like text. The framework uses Chat GPT to assist with the topic labelling and interpretation process by providing suggestions for topic labels based on the identified keywords. Chat GPT can also be used to summarize the identified topics, which helps to provide a quick overview of the main insights.

In conclusion, the framework for the topic interpretation process involves data scraping, data cleaning, topic modeling using the LDA algorithm, topic labelling and interpretation, text summarization, and the use of Chat GPT. The framework provides a comprehensive approach to extracting meaningful insights from a corpus of documents, and it can be used in various applications such as market research, social media analysis, and sentiment analysis. The framework's flexibility and the use of different techniques make it a powerful tool for analyzing.

This TIM project utilizes three different approaches to generate insights and compare the best approach, these approaches are as below:

- **Approach 1:** Choosing the most associated document from each company in the topic to generate the insights
- **Approach 2:** Choosing the first 50 most associated documents in each topic to understand if the process can draw a meaningful interpretation of the topic
- **Approach 3:** Keeping entire associated documents within the topic after putting the threshold to understand the generated interpretation and comparison with the previous two approaches

The identified associated documents will further be experimented using Chat GPT to extract more insights in combination with the above-extracted topic labels and summary. This project will aim to produce meaningful insights and better topic interpretation.

3.3 Identified list of companies within the talent management domain

S.NO	NAME	URL
1	The Predictive Index	https://www.predictiveindex.com/software/hiring/
2	Test Gorilla	https://www.testgorilla.com/product-tour/
3	Mercer Mettl Assessments	https://mettl.com/
4	eSkill	https://eskill.com/
5	Canditech	https://www.canditech.io/
6	Testlify	https://testlify.com/
7	Maki	https://www.makipeople.com/
8	HireQuotient	https://hirequotient.com/
9	Alva Labs	https://www.alvalabs.io/
10	iMocha	https://www.imocha.io/
11	WeCP	https://www.wecreateproblems.com/
12	REALCODERZ	https://realcoderz.com/#
13	Evalart	https://evalart.com/en/
14	HighMatch	https://www.highmatch.com/
15	HR Avatar	https://www.hravatar.com/ta/index.xhtml
16	Test Dome	https://www.testdome.com/
17	Recruiter.hr	https://recruiter.hr/
18	Equalture	https://www.equalture.com/
19	Cangrade	https://www.cangrade.com/home-2/
20	AssessFirst	https://www.assessfirst.com/en/
21	WonScore from Wonderlic	https://www.wonderlictestprep.com/
22	Thomas International	https://www.thomas.co/
23	Glider AI	https://glider.ai/
24	Equip	https://equip.co/
25	Bryq	https://www.bryq.com/
26	Vervoe	https://vervoe.com/
27	Modern Hire	https://modernhire.com/

28	Crystal	https://www.crystalknows.com/
29	SuccessFinder	https://www.successfinder.com/
30	Plum	https://www.plum.io/
31	Criteria	https://www.criteriacorp.com/assessments
32	HireVue	https://www.hirevue.com/
33	Harver	https://harver.com/
34	Otomeyt	https://otomeyt.ai/
35	Indeed	https://ca.indeed.com/leadershiphub/introducing-indeed-hiring-platform
36	Paradox	https://www.paradox.ai/
37	Personas	https://persona.se/
38	Pixonality	https://pixonality.com/
39	Sprockets	https://sprockets.ai/
40	TestGrid	https://www.testgrid.com/
41	Test Partnership	https://www.testpartnership.com/
42	Byte Board	https://byteboard.dev/
43	Caliper Essentials	https://calipercorp.com/selection/
44	Outmatch	https://outmatch.com/
45	Journeyfront	https://www.journeyfront.com/
46	Kolbe	https://www.kolbe.com/
47	Kryterion	https://www.kryterion.com/test-candidate/
48	McQuaig	https://mcquaig.com/
49	Prevue Assessments	https://www.prevuehr.com/
50	PXT Select	https://www.pxtselect.com/
51	SHL Recruit	https://www.shl.com/solutions/talent-acquisition/
52	Symphony Talent	https://www.symphonytalent.com/
53	Searchlight.ai	https://searchlight.ai/
54	Toggl Hire	https://toggl.com/hire/

55	Talentoday Manager	https://www.talentoday.com/
56	Adaface	https://www.adaface.com/
57	Aon Assessment Solutions	https://assessment.aon.com/en-us/
58	ACQai	https://www.acqai.io/
59	Arctic Shores	https://www.arcticshores.com/
60	ASC Smarter Assessment	https://assess.com/
61	Talview	https://www.talview.com/
62	Pulsifi	https://pulsifi.me/
63	Xobin	https://xobin.com/
64	Talogy	https://www.talogy.com/en/
65	Drawmetrics	https://www.drawmetrics.com/
66	Ducknowl	https://ducknowl.com/
67	Weirdly	https://www.getweirdly.com/
68	XBInsight	https://www.xbinsight.com/
69	Priority Bridge	https://prioritybridge.com/
70	Brillium	https://www.brillium.com/
71	Ensize	https://ensize.global/
72	Interviewer.AI	https://interviewer.ai/
73	Metahire	https://metahire.ai/
74	Testinvite	https://www.testinvite.com/
75	Hire Success	https://www.hiresuccess.com/

Table 2: List of talent management companies

3.4 Data collection

The corpus of documents is generated by scraping the companies' web pages using the developed scraping tools. A total of 75 Talent Management Companies' data have been scrapped. The scraped dataset has the company's name, URL, and webpage contents as corpus documents for the topic model and further text summarization.

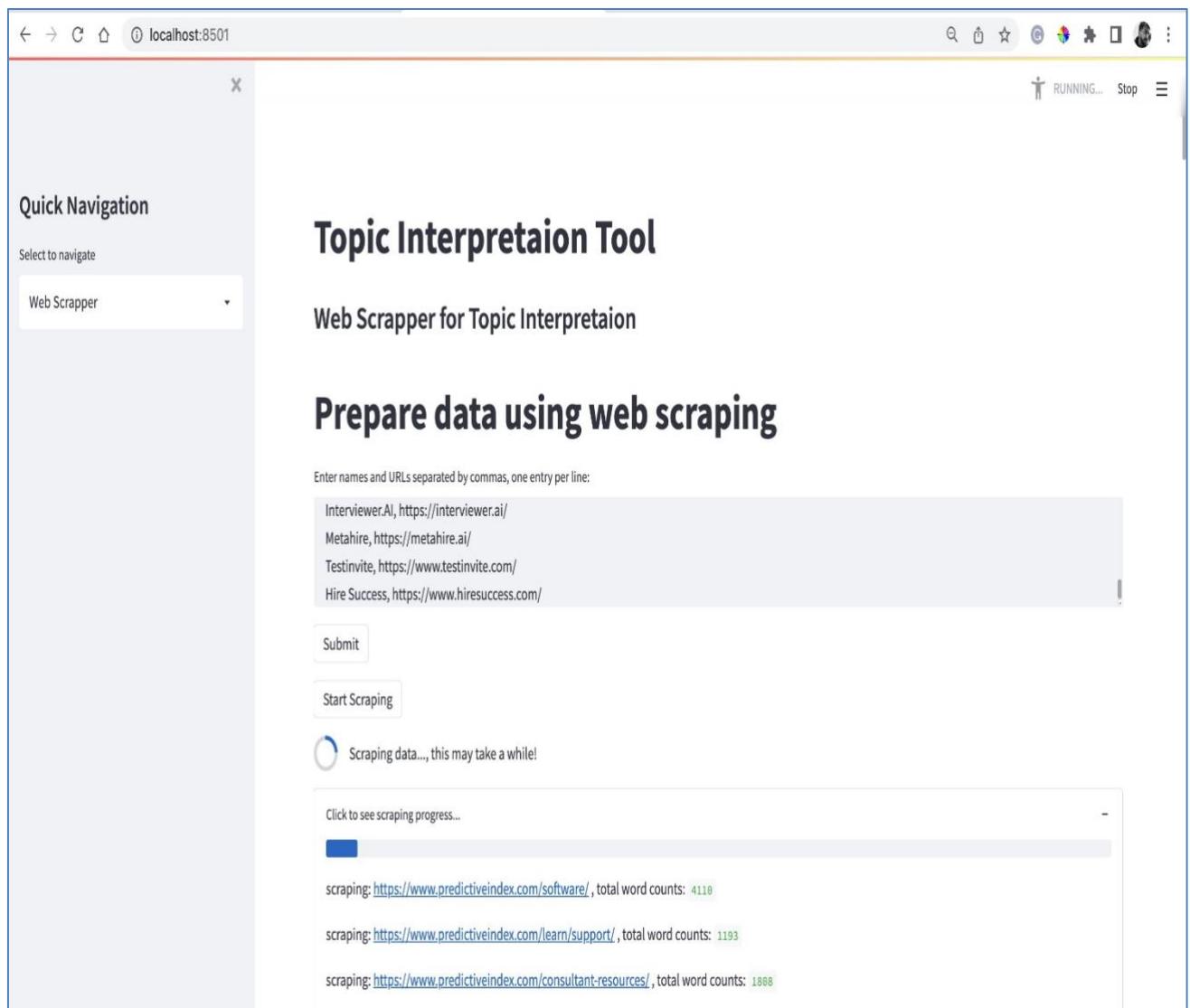


Figure 3: Data scraping tool

name	url	content
0 The Predictive Index	https://www.predictiveindex.com/software/	The best teams are magic. Roles and personalities complement each other. The whole becomes greater than the parts. Performance soars. Finally, there's a scientific way to design high-performing teams.
1 The Predictive Index	https://www.predictiveindex.com/learn/support/	~†(example: „Hire someone.“) Experience the power of the world. Às leading talent optimization platform Product Plans Call Support: 877-235-1541 Call Sales: 877-421-3717 Experience the power of the world. Às leading talent optimization platform Product Plans Call Support: 877-235-1541 Call Sales: 877-421-3717
2 The Predictive Index	https://www.predictiveindex.com/consultant-resource	Join 19K+ consultants and stay up to date with our top resources for consultants. FEATURED NEW FEATURED NEW Meritas uses people data to find the right people for the right roles for their clients.
3 The Predictive Index	https://www.predictiveindex.com/consultants/partner	How Six20 Partners provided urgent solutions that helped clients get things done Talent optimization consultant David Lunken is using PI to transform stress into value Great timing: IN2GREAT leverages talent optimization.
4 The Predictive Index	https://www.predictiveindex.com/learn/	Learn to make better hires, build high-performing teams, increase employee engagement, and improve relationships. Àthrough talent optimization. Sign up for free to get unlimited access to all of our free online training.
5 The Predictive Index	https://es.predictiveindex.com/software/hiring/	El marco de optimizaciòn del talento Marco probado para alinear la estrategia empresarial y la estrategia de personal para obtener resultados empresariales. Make the best hiring decision for a company.
6 The Predictive Index	https://www.predictiveindex.com/assessments/asses	ASSESSMENTS & VALIDITY Our assessments have been validated through rigorous scientific testing and research over decades. They're tried and true: We've assessed more than 25 million people.
7 The Predictive Index	https://www.predictiveindex.com/	Hire the right people, inspire their best work, design dream teams, and sustain engagement for the long haul. Meet the talent optimization platform Hire with certainty. Lead with purpose. Build a high-performing team.
8 The Predictive Index	https://www.predictiveindex.com/team-types/	Your team has an identity and personality. Team Types help you understand how that identity manifests as strengths and blind spots. Àand how to use each to the group's advantage. Book a seat.
9 The Predictive Index	https://www.predictiveindex.com/assessments/cognit	Did you know that cognitive ability is the number one predictor of job performance? The timed PI Cognitive AssessmentTM is a cognitive ability test for employment. This assessment measures a person's cognitive abilities, such as memory, reasoning, and problem-solving.
10 The Predictive Index	https://fr.predictiveindex.com/software/hiring/	Le cadre d'optimisation des talents Cadre prouvé pour aligner les stratégies commerciale et des ressources humaines afin d'obtenir des résultats professionnels. Make the best hiring decision for a company.
11 The Predictive Index	https://www.predictiveindex.com/software/diagnose	Isolate the cause of low employee engagement, Àand take action to fix it. If your people are disengaged, this is what's happening in your workplace every day. And it's preventing you from reaching your goals.
12 The Predictive Index	https://www.predictiveindex.com/learn/talent-optimi	300+ executives reveal the cost of the Great Quit, Àand the top drivers of employee retention. In November 2021, The Predictive Index surveyed 325 executives about The Great Resignation. No one is immune.
13 The Predictive Index	https://www.predictiveindex.com/learn/view-all/?_re	Free online training to get to the next level as a leader. Experience the power of the world. Às leading talent optimization platform Product Plans Call Support: 877-235-1541 Call Sales: 877-421-3717
14 The Predictive Index	https://www.predictiveindex.com/trust-center/	The Predictive Index is committed to high standards of security to deliver trust and assurance to our customers. Read more about PI's privacy practices and how we protect your data. Website.
15 The Predictive Index	https://www.predictiveindex.com/security/responsible	Predictive Index is committed to the security of customer information. We value the assistance of security researchers in keeping our systems secure. If you are a security researcher, we appreciate your work.
16 The Predictive Index	https://www.predictiveindex.com/learn/view-all/?_is	Free online training to get to the next level as a leader. Experience the power of the world. Às leading talent optimization platform Product Plans Call Support: 877-235-1541 Call Sales: 877-421-3717
17 The Predictive Index	https://www.predictiveindex.com/blog/	Join 150K+ leaders in our weekly email. Join 150K+ leaders in our weekly email. Andrew Barker. Let's go over what psychological safety means in the workplace, why it is important & what you can do to promote it.
18 The Predictive Index	https://www.predictiveindex.com/terms/website-term	Introduction The provisions of these Terms of Use form an agreement between Predictive Index, LLC („PI“), ÀWe, ÀOur, ÀU, ÀOur, ÀU or ÀUs, ÀU) and you. The following terms apply to your use of this website.
19 The Predictive Index	https://www.predictiveindex.com/learn/view-all/?_re	Free online training to get to the next level as a leader. Experience the power of the world. Às leading talent optimization platform Product Plans Call Support: 877-235-1541 Call Sales: 877-421-3717
20 The Predictive Index	https://www.predictiveindex.com/assessments/empli	Disengagement is tricky to spot, and even trickier to solve. PI lets you know why morale is low. Àfrom poor communication to culture issues. Àand gives you immediate next steps to mobilize, motivate, and engage.
21 The Predictive Index	https://www.predictiveindex.com/company/contact-u	Connect with us below to learn how to achieve your business goals through a deeper understanding of your people. M-F 8:30am, À 5:30pm EST Phone (781) 493-7535 Toll Free (877) 235-1541 p
22 The Predictive Index	https://www.predictiveindex.com/terms/	Who We Are The Predictive Index, LLC („PI“), ÀWe, ÀOur, ÀU, ÀOur, ÀU or ÀUs, ÀU) is a cloud-based SaaS platform that provides our Clients with Talent Optimization software and other products.
23 The Predictive Index	https://www.predictiveindex.com/news-press/	Experience the power of the world. Às leading talent optimization platform Product Plans Call Support: 877-235-1541 Call Sales: 877-421-3717 Experience the power of people data. Àfree Try the latest news.
24 The Predictive Index	https://www.predictiveindex.com/plans/	Priced by Activated Candidates Everything in Free, plus, À Priced by Activated Candidates Priced by Job & Activated Candidates Priced by Eligible Employee Count Priced by Teams Everything in Free.
25 The Predictive Index	https://www.predictiveindex.com/privacy/	This informs visitors to our public website of the procedures we follow and the technology we use to protect your information from unauthorized access, disclosure, or release. This policy outlines the types of personal information we collect, how we collect it, and how we use it.
26 The Predictive Index	https://de.predictiveindex.com/software/hiring/	Der Talent-Optimierungsrahmen BevvyShires Rahmenwerk für die Vereinbarung von Geschäftsstrategie und Personalstrategie für Geschäftsergebnisse. Make the best hiring decision for any company.
27 The Predictive Index	https://www.predictiveindex.com/assessments/job-a	Too often, you discover that the people you hired don't seem to be the same people you interviewed. Hiring is a tricky game indeed. But there's a better way to play that game. Job descriptions.
28 The Predictive Index	https://www.predictiveindex.com/software/hiring/	Make the best hiring decision for any role using millions of data points with PI Hire. Àleading pre-employment testing tool.* Try PI Hire completely free. Àno credit card required. TRUSTED BY: Fortune 500 companies.
29 The Predictive Index	https://www.predictiveindex.com/software/talent-str	Join thousands of managers using PI Design to build cohesive, collaborative, successful teams. Invite unlimited team members, and get all-access features for your first team. No credit card required.
30 The Predictive Index	https://www.predictiveindex.com/learn/talent-optimi	The PI Talent Optimization Certification will forever change the way you lead. After completing the certification, you'll be able to speak about talent strategy with authority and influence positively.
31 The Predictive Index	https://go1.predictiveindex.com/free	With this free 6-minute sample assessment, you can experience what our flagship Behavioral Assessment can uncover about you, your team, and how.

Figure 4: Sample of scraped data

3.5 Data analysis

Cleaning of data is required in topic modeling to improve the accuracy and effectiveness of the topic model result. Cleaning involves removing any irrelevant or noisy data that may hinder the interpretation of the text. This includes removing special characters, stop words, multi-words, and any other irrelevant data that does not contribute to the meaning of the text. The presence of such irrelevant data can lead to biased results or inaccurate summaries.

Additionally, data cleaning helps to ensure that the text data is standardized and consistent, making it easier to analyze and interpret. By removing inconsistencies and standardizing the text, the text summarization process can be made more efficient and accurate.

Data cleaning is required to improve the accuracy and efficiency of automated text summarization by removing any irrelevant or noisy data that can hinder the interpretation of

the text, and by standardizing the text to ensure consistency and ease of analysis (Fakhitah et al., 2019).

Hence, prior to analyzing the corpus, the date was cleaned. To accomplish this, Python libraries are used to identify and remove noise such as special characters, stop words, and multi-words. In addition, extraneous data such as emojis, punctuation, and non-English words are also eliminated.

The cleaned corpus is then used for the topic model and then the model is evaluated based on its interpretability. The final output of the process is evaluated once before finalizing the process.

The developed tool requires to:

- ❖ Upload the scraped corpus
- ❖ Upload the stop words txt file
- ❖ Upload the multiword txt file
- ❖ And click on the Clean Corpus button to get the cleaned data
- ❖ The corpus is cleaned for:
 - Punctuation
 - Non-English contents
 - All contents to lowercase
 - Removed stop-words and processed multi-words
 - Performed Text lemmatization

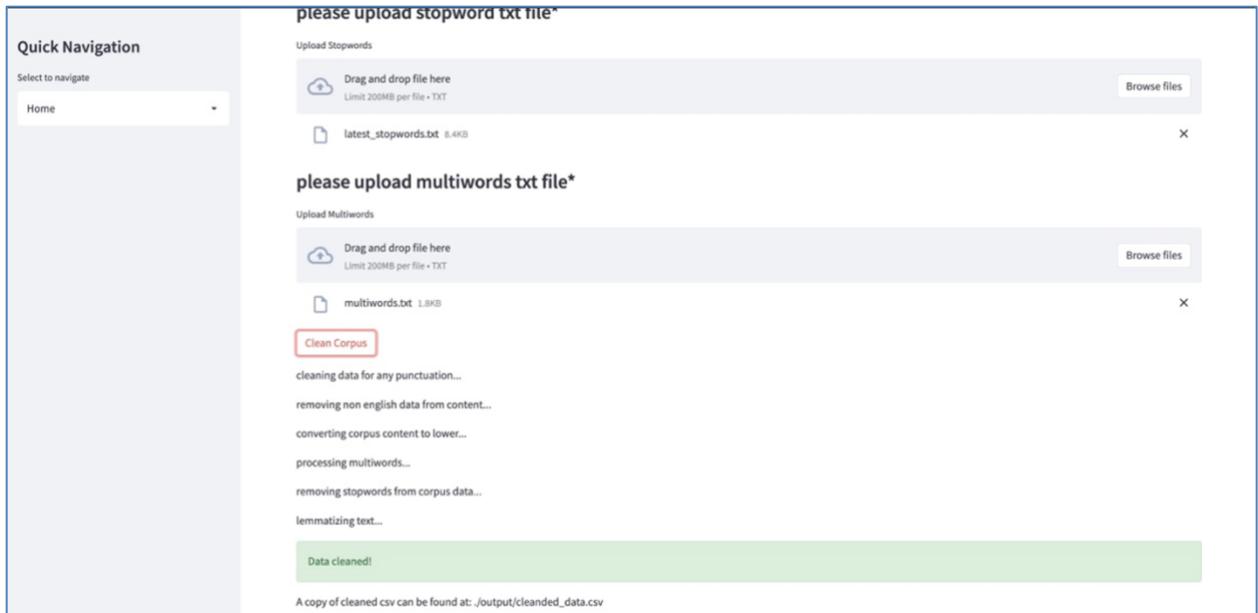


Figure 5: Data cleaning process

After cleaning the data, the next step in topic modeling is to use the cleaned data as input for the tool. The tool will then determine the ideal number of topics and generate the corresponding topic keywords.

3.6 Implement and apply the topic model

Topic modeling is a type of statistical modeling used for uncovering hidden patterns or themes within a large corpus of text data. It involves identifying a set of topics and their corresponding keywords that best represent the underlying themes within the corpus. The goal of topic modeling is to help in organizing, summarizing, and exploring large volumes of textual data by providing insights into the key themes and topics being discussed.

3.6.1 Identify the optimal number of topics using coherence score

LDA topic coherence score is a metric used to evaluate the quality of topics generated by Latent Dirichlet Allocation (LDA), a popular topic modeling algorithm. Topic coherence measures how interpretable and coherent the generated topics are by evaluating the degree of

semantic similarity between the most important words within each topic.

The coherence score is calculated based on the frequency and distribution of words in the corpus and is used to assess the performance of different topic models and the optimal number of topics to extract. A higher coherence score indicates more coherent and meaningful topics.

After the data is cleaned, the LDA topic coherence score is used to determine the optimal number of topics for the corpus. The coherence score should be greater than 0.5, which is considered good for most cases. The optimal number of topics is determined to be around 10, based on the highest coherence score achieved. This methodology is based on the work of Rosner et al. (2014).

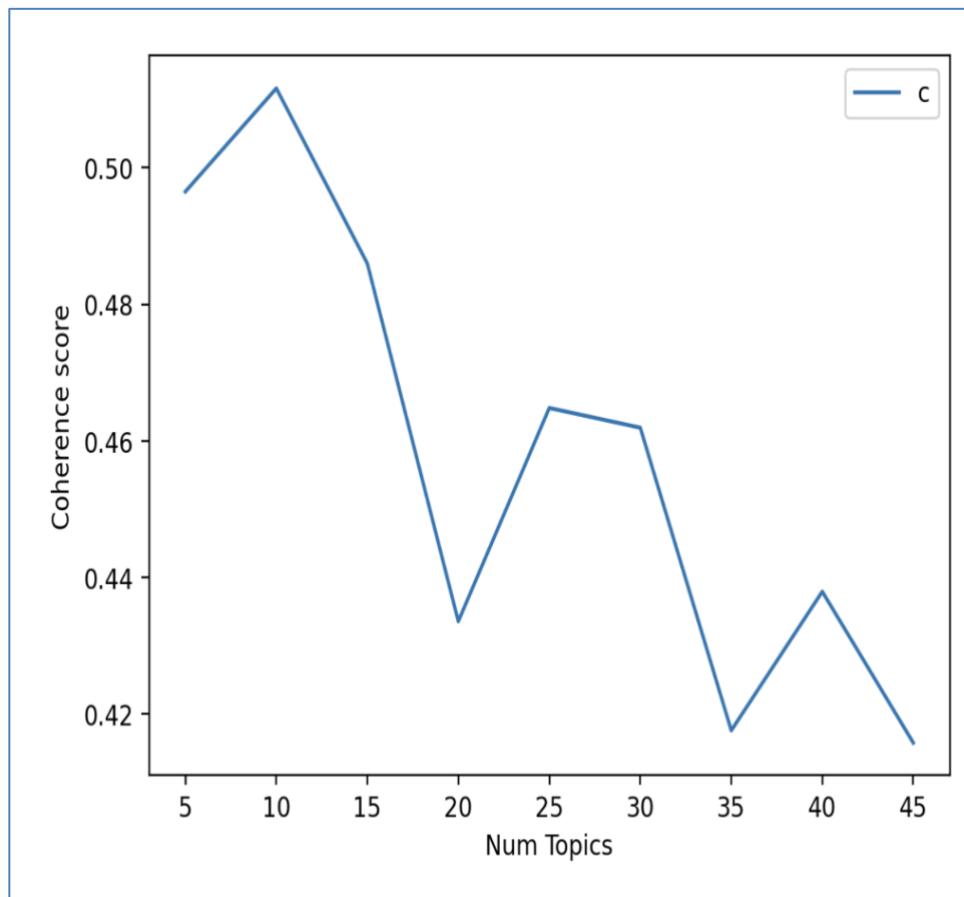


Figure 6: Topic model coherence graph

3.6.2 Topic model

The optimal number of topics for the topic model was determined by running the model with different numbers of topics until distinguishable keywords for each topic were obtained. The LDA topic coherence score was used to determine the best number of topics for the data, with a score greater than 0.5 being considered good for most cases. The coherence score was highest when the number of topics was around 10, but after experimentation, it was found that the optimal number of topics for the data was 9. The resulting topic models generated with 9 topics had distinguishable keywords for each topic.

3.6.3 Data visualization

Figure 6 shows a **word cloud**: A word cloud is a graphical representation of words that frequently appear in a text corpus. It is a visual representation of text data that shows the most frequently used words in larger fonts and the less frequently used words in smaller fonts. Word clouds are often used to provide a quick overview of the most important themes in a text corpus. They are popular in data analysis, data mining, and natural language processing (Carpineti, 2020).

In the context of this project, a word cloud was generated for the cleaned corpus of talent management company web pages to provide a visual representation of the most frequently used words in the corpus. The word cloud helped to identify the key themes that are common in the talent management industry. As seen in Figure 7, the word cloud shows that some of the most commonly used words in the corpus are "talent," "management," "employee," "development," "performance," "training," "career," and "leadership."

Word clouds are useful for data visualization and can be used in a variety of applications such as marketing, social media analysis, and trend analysis. They provide an easy-to-understand representation of text data, especially when the corpus is large and complex. However, it is important to note that word clouds have limitations. They do not provide detailed information about the relationship between words, and they do not take into account the context in which the words are used. Therefore, they should be used in conjunction with other data analysis

tools and techniques to gain a comprehensive understanding of the text corpus.

In conclusion, word clouds are a useful tool for visualizing the most frequently used words in a text corpus. They provide an easy-to-understand overview of the key themes in the corpus and can be used to identify trends and patterns. However, they should be used in conjunction with other data analysis techniques to gain a comprehensive understanding of the corpus.



Figure 7: word cloud

Figure 7: shows the **inter-topic distance map** visualization, it is a tool used to depict the relationships between the topics in a corpus generated from a topic model. The visualization helps to explore how the clusters of related topics are connected. The map consists of nodes, which represent the individual topics, and edges, which represent the distances between the topics. The edges are colored based on the strength of the relationship between the topics. The visualization is often used to identify groups of similar topics that are closely related to each other, as well as to identify topics that are outliers and not strongly related to other topics in the corpus.

The inter-topic distance map visualization can provide valuable insights into the structure of a corpus generated from a topic model, as well as help to identify areas where additional analysis or refinement may be needed. By exploring the relationships between topics, researchers can gain a deeper understanding of the underlying themes and concepts within the corpus, as well as identify patterns and trends that may not be immediately apparent from a simple analysis of the individual topics. Overall, the inter-topic distance map visualization is a useful tool for researchers looking to gain a deeper understanding of the relationships between topics in a corpus generated from a topic model (Chuang and Heer, 2012).

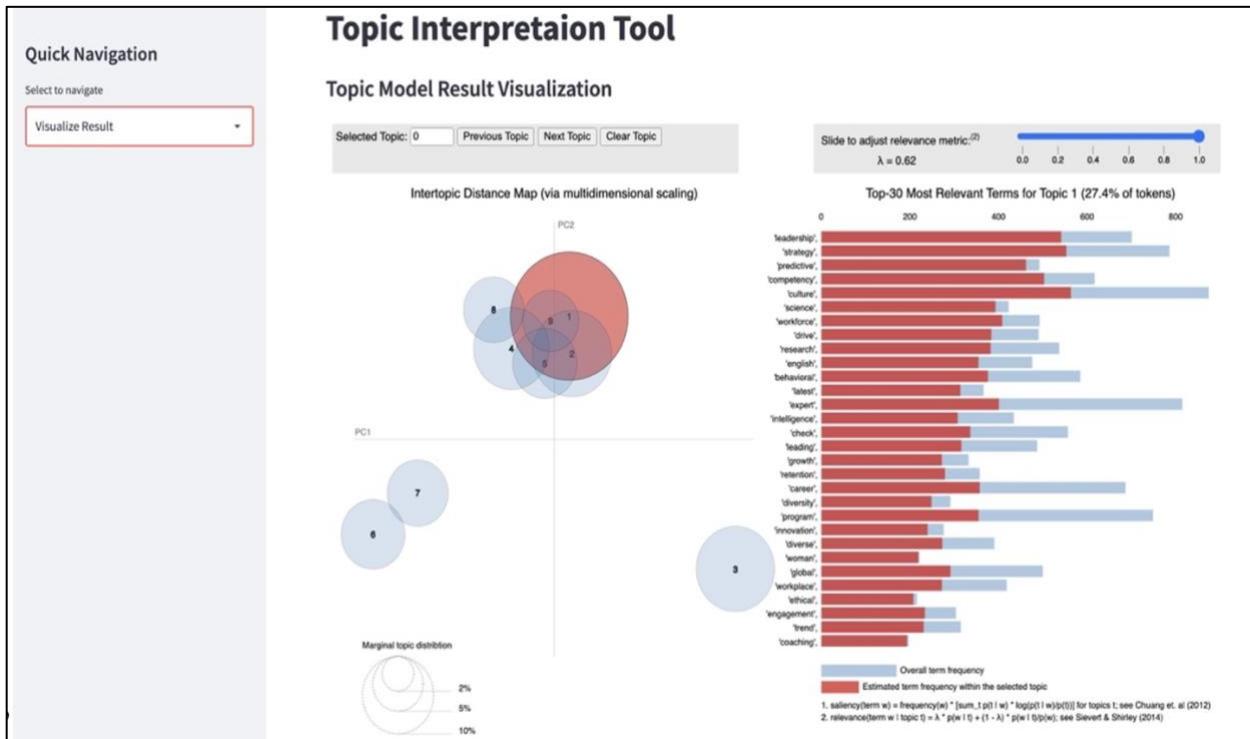


Figure 8: inter-topic visualization map

3.7 Companies representing each topic

After running a topic model, it is possible to generate the documents that are associated with each topic and their corresponding contribution percentage. This means that the algorithm has assigned a certain topic to each document in the dataset, and that topic is assigned a certain weight based on the words and phrases found in that document.

Once the topics have been identified and associated with their respective documents, text summarization can be performed. This involves generating a concise summary of the main points discussed in each document. This can be done using various algorithms and techniques such as sentence extraction, keyword extraction, and summarization using neural networks.

Another technique that can be used with the topic model is threshold cut-off. This involves setting a threshold for the topic contribution percentage so that only documents with a contribution percentage above that threshold are considered relevant for that particular topic. This can help to refine the topic model and make it more accurate, as well as reduce noise in the data.

Overall, the process of running a topic model involves identifying the topics that are present in a given dataset, assigning those topics to the documents in the dataset, and generating associated documents and summaries. The LDA topic coherence score can be used to evaluate the quality and relevance of the topics generated by the model. The resulting topic model can be used for a variety of purposes, such as information retrieval, content analysis, and knowledge discovery.

Topic documents						
GET DOCUMENTS						
	Topic_Num	Topic_Perc_Contrib	Keywords	Text	Name	
0	0.0000	0.9991999864578247	'subscription', 'applicabl...	1.1 Acceptance of these T...	Pulsifi	
1	0.0000	0.9991999864578247	'subscription', 'applicabl...	Keep your hiring team effi...	Interviewer.AI	
2	0.0000	0.9987000226974487	'subscription', 'applicabl...	English English Search US...	Talogy	
3	0.0000	0.9986000061035156	'subscription', 'applicabl...	Thank you for taking the t...	Adaface	
4	0.0000	0.9986000061035156	'subscription', 'applicabl...	Keep your hiring team effi...	Interviewer.AI	
5	0.0000	0.9977999925613403	'subscription', 'applicabl...	Last updated June 16, 20...	Test Gorilla	
6	0.0000	0.9930999875068665	'subscription', 'applicabl...	Send us a message below ...	Caliper Essentials	
7	0.0000	0.9930999875068665	'subscription', 'applicabl...	Send us a message below ...	Caliper Essentials	
8	0.0000	0.9832000136375427	'subscription', 'applicabl...	Claritas est etiam process...	Recruiter.hr	
9	0.0000	0.9815000295639038	'subscription', 'applicabl...	Claritas est etiam process...	Recruiter.hr	

Figure 9: sample topic documents

3.7.1 Cutting off document based on the topic percentage contribution

Topic percentage contribution refers to the proportion of a particular topic in a given document or text corpus. In topic modeling, each document is typically assigned a set of topics and their corresponding probability distributions. The topic percentage contribution refers to the proportion of a particular topic in a given document, represented as a percentage of the total number of topics in that document. For example, if a document is assigned three topics with probabilities of 0.4, 0.3, and 0.3, respectively, the topic percentage contribution for the first topic would be 40, then 30% and 30% respectively.

Now the cutting off of documents based on the topic percentage contribution involves setting a threshold for the percentage of a document that needs to be related to a particular topic in order for it to be included in the analysis. For example, if the threshold is set at 70%, then only documents that are at least 70% related to a given topic will be included in the analysis for that topic.

The purpose of this approach is to ensure that only documents that are highly relevant to a particular topic are included in the analysis. This can help to reduce noise and improve the accuracy of the topic modeling results.

Understanding the topic percentage contribution is important because it allows us to identify the most relevant topics in a given document or corpus, and to track changes in topic prevalence over time. Topic percentage contribution can also be used to group documents by their most dominant topics, allowing for more efficient analysis and interpretation of large datasets. Overall, topic percentage contribution is a key metric in topic modeling that helps us to understand the underlying themes and patterns in text data.

For this project, we have experimented with the documents in multiple settings such as 70%, 60% and 50% depending upon the approach and document counts under each topic:

please provide topic percentage cut off for your document!

Enter cut off, example 0.85, please save a copy of previous document, it will overwrite the existing doc

selected cut off: 0.7

PROCEED DOC CUT OFF

Figure 10: topic documents threshold

3.8 Summary of this chapter

This section outlines the research method used in the project and the steps taken to produce the intended deliverables. The first step involved conducting a literature review on Talent management domain, topic modeling, labeling and text summarization techniques. This review helped in identifying the limitations of the current practices and the challenges faced by users, which served as a motivation for the project. The second step involved choosing an appropriate topic model, labelling process and MDS algorithm that could help interpret the topic model. This step also involved studying the features and implementation methods offered by other researchers to refine the process of interpretation.

The results of the project were presented using the Streamlit application, which allowed for a user-friendly interface and easy access to the project's outcomes. Overall, this research method section highlights the systematic and comprehensive approach used in the project to address the challenges faced interpreting the results of topic modeling.

4 Summary of results

The results of the research project, which consists of four deliverables, are presented in this section.

- The first deliverable is the cleaned corpus of data from 75 talent management companies
- The second deliverable is the topic modelling results on the corpus of scraped web pages from Talent management company websites
- The third deliverable is a process to interpret each of the topic models based on the summarization of the associated documents and the application of Chat GPT
- The fourth is a refined process that could generate additional topic interpretation insights based on the companies associated with each topic

The results of each deliverable are discussed in detail in the following sections:

4.1 Building the corpus of text documents

The corpus of documents is generated by scraping the companies' web pages using the developed scraping tools. A total of 75 Talent Management Companies' data have been scrapped. The scraped dataset has the company's name, URL and webpage contents as corpus documents. The corpus also tokenized, a tokenized corpus refers to a collection of text documents that have been processed and converted into individual words or phrases, known as tokens. Tokenization is a natural language processing technique that involves breaking down a document into smaller units for analysis. These tokens can then be further processed to identify patterns and relationships within the text, such as frequently occurring words or phrases. A tokenized corpus is often used as input for various natural languages processing tasks, such as topic modeling, sentiment analysis, and text classification. By tokenizing a corpus, it becomes easier to perform these tasks as the text is converted into a format that can be easily analyzed by computers.

Figure 11: cleaned tokenized corpus

The cleaned corpus was then analyzed using TF-IDF to extract initial insights from the data. TF-IDF (Term Frequency-Inverse Document Frequency) is a statistical method that is commonly used for analyzing textual data. It is used to determine the importance of a word in a document by calculating the frequency of the word in the document (term frequency) and dividing it by the frequency of the word in the corpus (inverse document frequency).

The TF-IDF algorithm is often used in natural language processing and information retrieval to extract the most significant words or phrases from a collection of texts. To identify the most discussed topic using TF-IDF, we need to calculate the TF-IDF score for each word in the corpus and then identify the words with the highest scores. These words are likely to be the most significant and therefore provide insights into the most discussed topic.

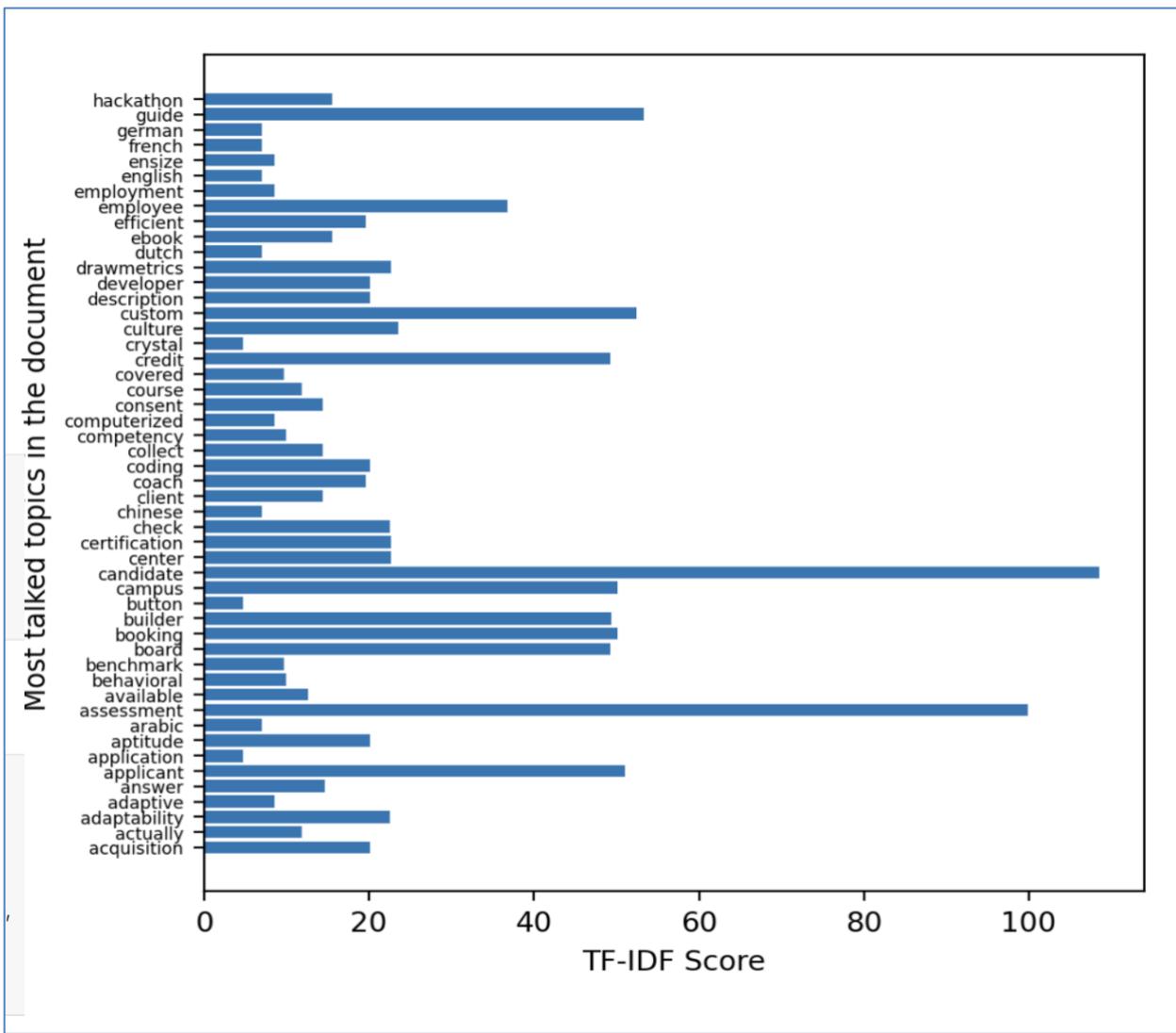


Figure 12: Most frequent words in the corpus

4.2 The topic model

The corpus was cleaned by removing abnormal letters, punctuation marks, converting all uppercase letters to lowercase, and removing duplicate statements before performing topic modeling.

Topic Model

Enter number of topics values 1-50

You have selected 9 topics, please click below to proceed

PROCEED

procedding...

 Topic modelling

Topic Number	Topic words
1	subscription, applicable, liability, material, condition, provision, responsible, damage, dispute, limitation,
2	culture, strategy, leadership, competency, predictive, workforce, expert, science, drive, research,
3	course, practice, minute, phone, title, deimeasure, choose, comment, actually, employer,
4	builder, coffee, trait, serving, wayne, distributor, vendor, award, hundred, daunting,
5	purpose, reference, protection, device, crystal, thirdparty, legal, transfer, necessary, protect,
6	questionnaire, complete, director, standard, simulation, field, adaptive, staff, necessary, helped,
7	integrate, trial, workflow, credit, board, update, refer, anywhere, smarter, navigate,
8	proctoring, remote, examination, secure, certification, feature, automated, recruiting, proctor, screen,
9	campus, booking, efficient, expert, description, succeed, productive, manage, domain, coach,

Figure 13: topic model result

The obtained topic model can be considered optimal for the given corpus as it includes a diverse range of topics and is both concise and interpretable. It strikes a balance between capturing the main topics and subtopics within the data. Topics are the main themes or subjects that are discussed in the data. For example, in a corpus about talent management, some of the main topics might be employee recruitment, performance management, and leadership development.

And the subtopics are more specific or detailed aspects of a topic.

For example, under the topic of employee recruitment, there might be subtopics such as candidate sourcing, interview techniques, and job offers.

By identifying both the main topics and subtopics within the data, we can get a more comprehensive understanding of the different areas related to talent management. The model's keywords offer valuable insights into the underlying themes and patterns present in a corpus of text data.

Topic model keywords at a glance:

Topic 1: 'subscription', 'applicable', 'liability', 'material', 'condition', 'provision', 'responsible', 'damage', 'dispute', 'limitation'

Based on the given words, this topic seems to be related to legal terms and agreements. It may include clauses or provisions related to subscription services, such as terms and conditions, liability, material, and limitations. It may also include words related to resolving disputes, including responsibility and damages. Overall, this topic may be related to the legal language and regulations surrounding the use of subscription services or other similar products.

Topic 2: 'culture', 'strategy', 'leadership', 'competency', 'predictive', 'workforce', 'expert', 'science', 'drive', 'research'

This topic appears to be concerned with managing human resources and organizational behavior. The included keywords indicating a focus on developing strong organizational culture and leadership to drive success. The inclusion of 'competency' suggests a focus on developing and managing workforce skills and knowledge. The terms 'predictive' and 'expert' may suggest the use of data science and research to inform decision-making and develop strategies. Overall, this topic seems to be centered around how organizations can effectively manage and develop their employees to achieve success.

Topic 3: 'course', 'practice', 'minute', 'phone', 'title', 'deimeasure', 'choose', 'comment', 'actually',, 'employer'

The relationship between these keywords and talent management doesn't look very clear just by looking at these topic keywords. Some of the keywords, such as 'phone' and 'title', do not seem to have a direct connection. However, the presence of keywords like 'course' and 'practice' could indicate that the topic is related to employee training and development, while 'employer' suggests a potential focus on acquiring and retaining talent. It is challenging to establish a specific connection to talent management without additional information or examination.

Topic 4: 'builder', 'coffee', 'trait', 'serving', 'wayne', 'distributor', 'vendor', 'award', 'hundred', 'daunting'

The keyword 'trait' could possibly relate to identifying and developing desirable skills and characteristics in employees. The mention of an 'award' suggests a possible focus on recognition and rewards for employees. However, without additional context or analysis, it is unclear how these keywords relate to talent management in a meaningful way.

Topic 5: 'purpose', 'reference', 'protection', 'device', 'crystal', 'thirdparty', 'legal', 'transfer', 'necessary', 'protect'

In this topic, the presence of keywords such as 'legal' and 'third-party' indicates a potential emphasis on safeguarding sensitive information and proprietary knowledge. The term 'transfer' could be linked to the internal or external mobility of employees.

Topic 6: 'questionnaire', 'complete', 'director', 'standard', 'simulation', 'field', 'adaptive', 'staff', 'necessary', 'helped'

The presence of the keyword 'questionnaire' indicates that there might be a focus on gathering feedback or evaluating performance of employees. The keywords 'director' and 'staff' may suggest a possible focus on developing leadership and training employees. The term 'adaptive' may indicate a focus on creating a flexible and responsive workforce.

Topic 7: 'integrate', 'trial', 'workflow', 'credit', 'board', 'update', 'refer', 'anywhere', 'smarter', 'navigate'

The keywords 'workflow' and 'update' suggest a possible focus on improving HR processes and systems. The term 'credit' could refer to employee recognition programs or performance evaluation metrics. The inclusion of 'smarter' and 'navigate' may indicate a focus on improving the user experience of HR software or tools.

Topic 8: 'proctoring', 'remote', 'examination', 'secure', 'certification', 'feature', 'automated', 'recruiting', 'proctor', 'screen'

The topic appears to be centered around remote examination proctoring and certification, with a focus on creating a secure and automated system for testing. The inclusion of 'recruiting' suggests that certification may be used as a means of attracting talent. The keyword 'screen' may refer to identity verification of exam takers. Overall, this topic seems to be about developing a secure and efficient method for remote testing and certification.

Topic 9: 'campus', 'booking', 'efficient', 'expert', 'description', 'succeed', 'productive', 'manage', 'domain', 'coach'

The keywords in this topic suggest that the focus is on managing a facility, possibly in a specific industry or field. The keywords 'expert' and 'coach' indicate that there may be a focus on staff development and training to improve productivity and success. The keyword 'domain' may suggest a specific area of expertise or knowledge required for managing the facility. Additionally, the keyword 'booking' suggests that the management of reservations and schedules may be important in this context. Overall, it seems that this topic is related to the effective management of a facility, with a focus on staff development and productivity.

Summarizing this, upon analyzing the topic keywords, it appears that there are some topics that lack a clear and distinct theme. This makes interpreting the meaning of these topics challenging.

Although we can infer that the keywords are related to a specific theme, it is difficult to fully understand the topic without using additional tools.

While the topic keywords provide some indication of a potential theme, they are not sufficient on their own to fully interpret the meaning of a topic as can be seen with above Topic 3 and Topic 4.

In order to gain a deeper understanding, we will be using additional tools such as examining the associated documents, automated topic labeling and summaries of the associated documents within the topic. This will help us identify common patterns or topics that are consistently mentioned. Moreover, we will also utilize Chat GPT to extract more insights and provide further analysis of the topic model. By combining these tools, we can enhance our understanding of the topics and their underlying themes.

Let us see how the combination of these techniques under different settings can help interpret the topic model result in further sections.

4.2.1 Approach 1: Choosing the most associated documents from each company related to a topic

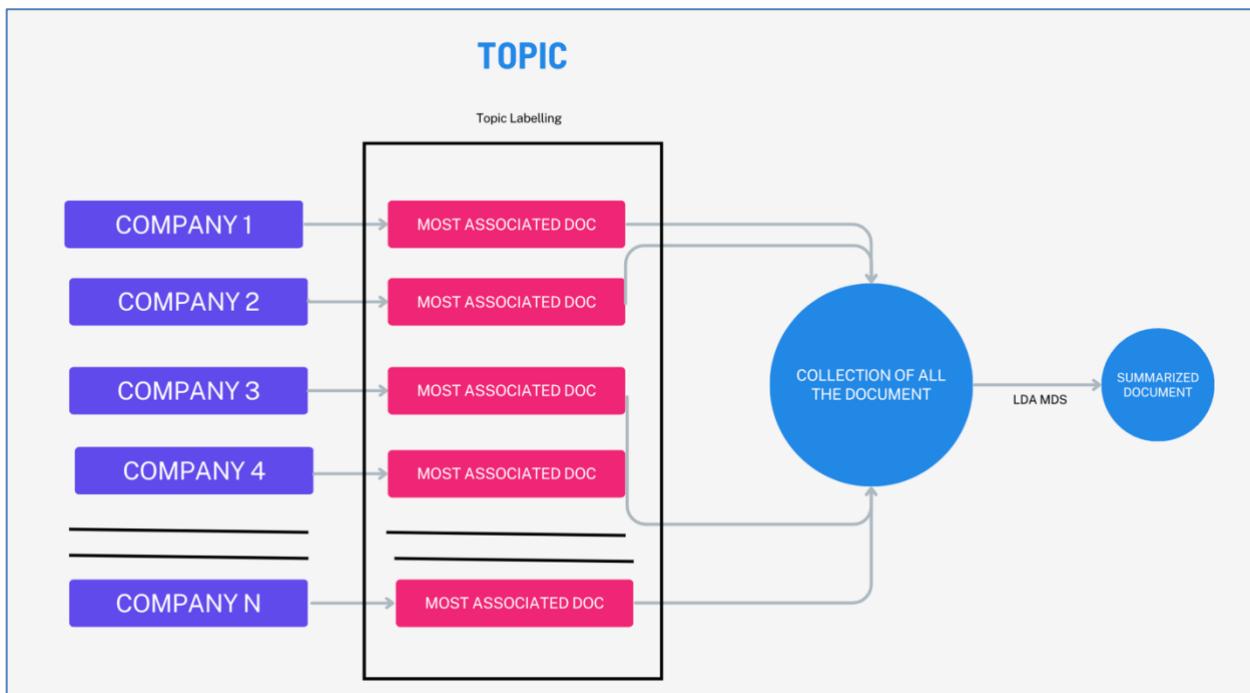


Figure 14: topic interpretation: Approach 1

In this approach, for a given topic the most highly associated document from each company for the given topic was picked from the collection of these highly associated documents, the text summarization and topic label interpretation were performed.

Generating topic labels!

Topic 0

Topic Keywrods: 'subscription', 'applicable', 'liability', 'material', 'condition', 'provision', 'responsible', 'damage', 'dispute', 'limitation',

Topic label: Recruiting - The Best Way to Find the Best Talents

Topic 1

Topic Keywrods: 'culture', 'strategy', 'leadership', 'competency', 'predictive', 'workforce', 'expert', 'science', 'drive', 'research',

Topic label: Ingredients to Build a Powerful Talent Management Platform

Topic 2

Topic Keywrods: 'course', 'practice', 'minute', 'phone', 'title', 'deimeasure', 'choose', 'comment', 'actually', 'employer',

Topic label: How to Improve Your Talent Management

Topic 3

Topic Keywrods: 'builder', 'coffee', 'trait', 'serving', 'wayne', 'distributor', 'vendor', 'award', 'hundred', 'daunting',

Topic label: Ways to Find the Best Candidates for Your Next Job Search

Topic 4

Topic Keywrods: 'purpose', 'reference', 'protection', 'device', 'crystal', 'thirdparty', 'legal', 'transfer', 'necessary', 'protect',

Topic label: How to Protect Your Personal Data

Figure 15: Automatic topic label generation

Topic 5 Topic Keywrods: 'questionnaire', 'complete', 'director', 'standard', 'simulation', 'field', 'adaptive', 'staff', 'necessary', 'helped', Topic label: eSkill - A Great Pre-Selection Tool for Recruiting
Topic 6 Topic Keywrods: 'integrate', 'trial', 'workflow', 'credit', 'board', 'update', 'refer', 'anywhere', 'smarter', 'navigate', Topic label: Testlify - Discover the Best Hire With Insights
Topic 7 Topic Keywrods: 'proctoring', 'remote', 'examination', 'secure', 'certification', 'feature', 'automated', 'recruiting', 'proctor', 'screen', Topic label: Achieve a Better Job Search With a Mobile App

Figure 16: Automatic topic label generation (contd...)

4.2.1.1 Grouping companies based on the generated topic label

Grouping of companies based on topic labels involves clustering companies together based on their topic contributions in a topic model. Companies that have similar topic distributions or have a higher contribution to specific topics can be grouped together to identify patterns or trends within the dataset.

This can help in gaining a better understanding of the topics being discussed within a particular industry or domain and identifying key players or competitors in that space. Grouping can be done using various clustering techniques, such as k-means clustering, hierarchical clustering, or latent Dirichlet allocation (LDA) clustering.

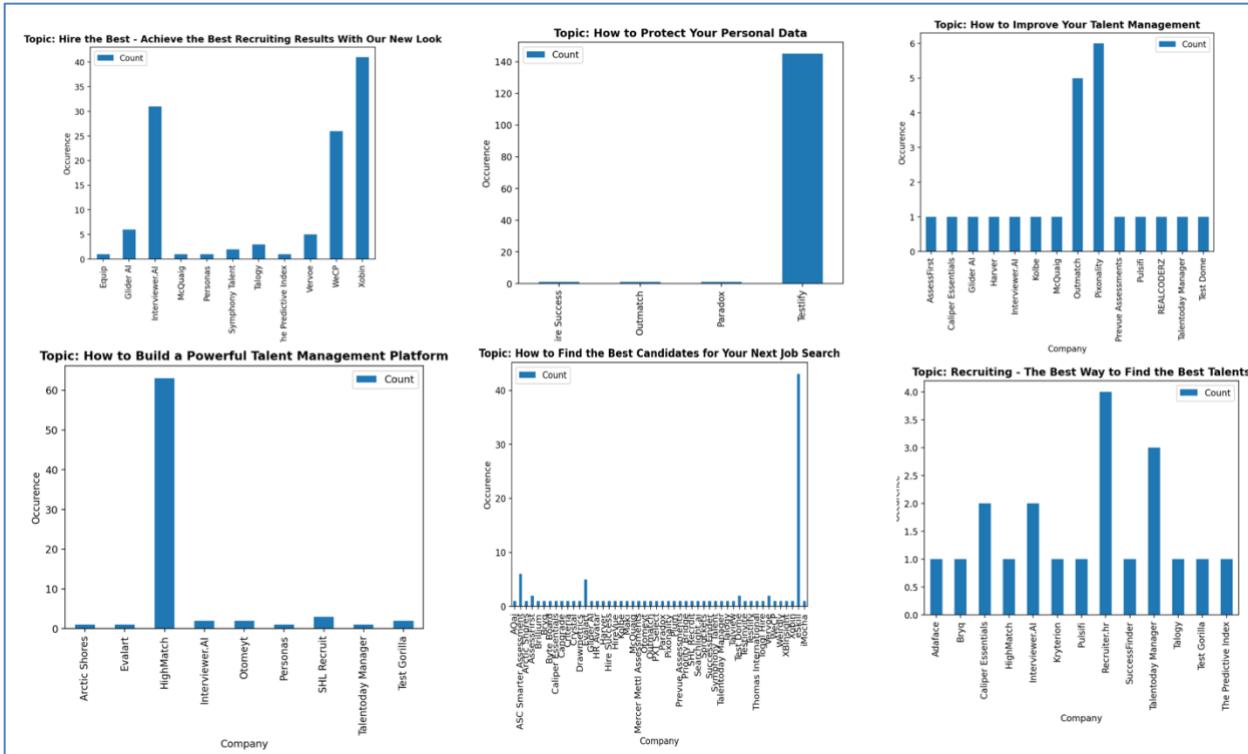


Figure 17: Companies grouped per topic label

In the context of talent management tools and software, protecting personal data is crucial as it involves the storage and processing of sensitive information such as resumes, contact information, and employment history. Online recruitment software providers must implement proper security measures to ensure that the data is protected from unauthorized access, disclosure, and alteration.

Topic: “How to protect your personal data”



Figure 18: Companies grouped on topic

There are multiple ways to do this, one way to protect personal data is through encryption, which involves converting the data into a coded language that can only be deciphered by authorized users with the proper decryption key. Another important aspect is implementing access controls to limit the people who have access to the data. This involves setting up roles and permissions for users, such as recruiters or HR personnel, to ensure that they only have access to the data that is necessary for their job function.

Data backups and disaster recovery plans are also important in case of data loss or system failure. Online recruitment software providers must have a robust backup system in place to ensure that the data can be restored in case of a disaster. Additionally, regular testing of the disaster recovery plan is essential to ensure that it works as expected.

Overall, protecting personal data in online recruitment software requires a multi-faceted approach that involves encryption, access controls, data backups, and disaster recovery plans. It is important for companies to prioritize data protection and ensure that their software providers are implementing the necessary security measures to protect sensitive information.

Now, if we see the above topic distribution graph **Testlify** has almost 140 numbers of documents talking about security and personal data protection, which seems to be a good option when an organization is looking to have secure tool requirements.

The results of the analysis show that the topic "How to find the best candidates for your next job search" is widely discussed by all the companies in the dataset, indicating that they all focus on assisting organizations with their new hire process, which is a common topic in the Talent management domain.

Topic: “Recruiting – The best way to find the best talent”

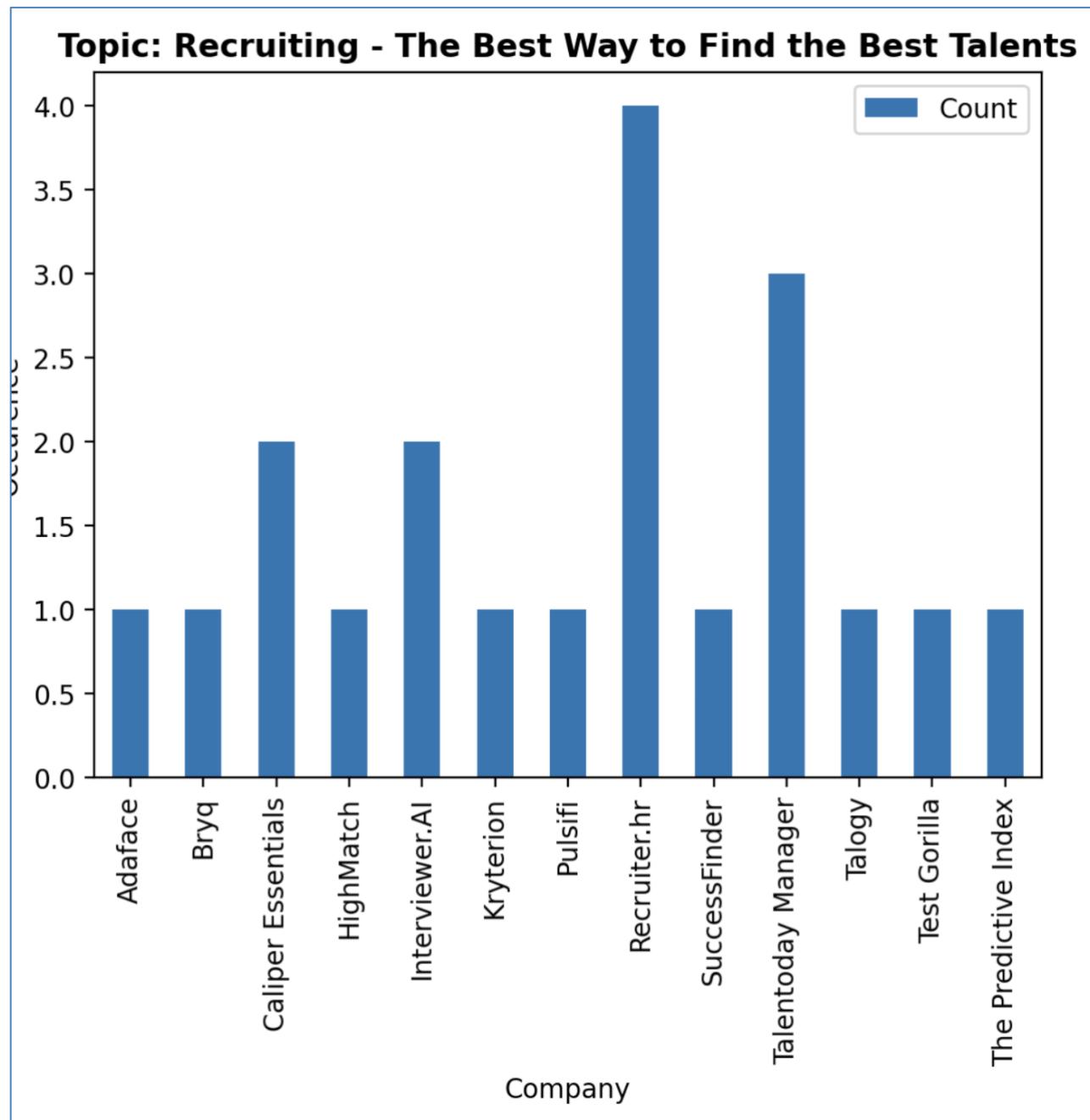


Figure 19: Companies grouped on topic

The data collected and analyzed in this project is related to companies and tools in the talent management domain. Upon analyzing the data, it became apparent that the majority of these companies were focusing on promoting themselves by highlighting their ability to find the best

talents using their available tools. This was seen in the common topic of "How to find the best candidates for your next job search" which was present across all the companies in the dataset.

Furthermore, the promotion of tools and strategies to find the best talent was also observed in the topics related to interviewing processes, candidate evaluation, and recruiting techniques. These topics suggest that companies are prioritizing their ability to effectively source and select top-quality candidates for their clients.

Overall, the data analysis shows that companies in the talent management domain are emphasizing their ability to source and select top-quality candidates using their available tools, while also prioritizing data privacy and protection.

Topic: "How to improve your talent management"

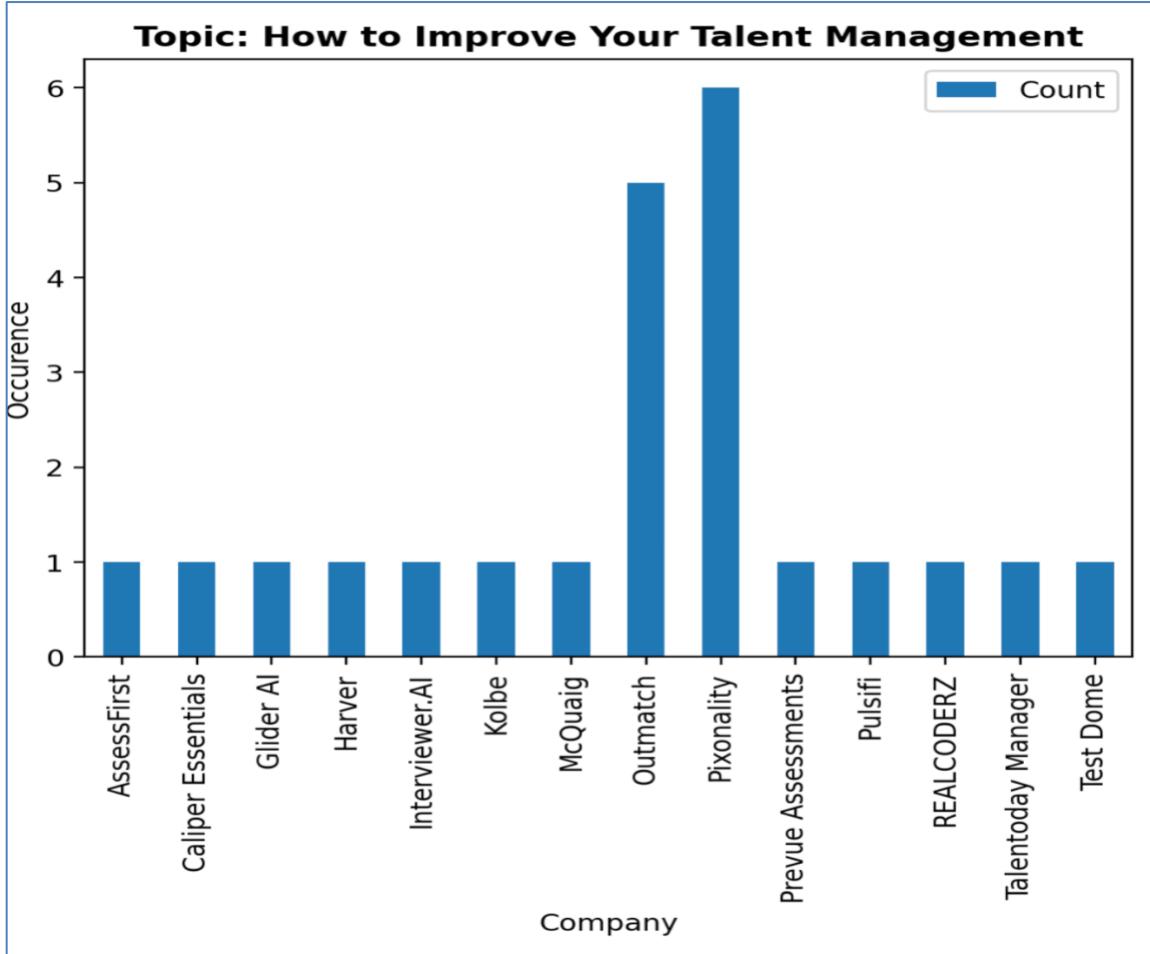


Figure 20: Companies grouped on topic

Based on the result, it can be inferred that a significant number of companies (around 14 out of 75) in the dataset are emphasizing their efforts in improving the overall talent management process. This suggests that the companies are aware of the importance of efficient talent management and are actively working towards enhancing their processes. The data also indicates that these companies are using various tools and techniques to improve the talent management process, as seen from the topics such as "How to find the best candidates for your next job search" and "How to improve candidate experience through recruitment software". Additionally, it is evident that companies are promoting their tools and services by highlighting their ability to

help organizations find the best talents using their available tools, which suggests a competitive market for talent management solutions.

Approach 1 has been successful till now in interpreting topic labels by grouping companies together and identifying common themes through an automated topic labeling process. This approach extracts the most associated documents from each topic and provides insights into which companies are focused on which topic. By extracting the most associated documents from each topic, this approach provides valuable insights into which companies are focused on which topic, allowing for a better understanding of the talent management landscape. Till now it seems to be a more effective way of analyzing a large corpus of talent management data and identifying patterns and trends.

4.2.1.2 Automated multi-document text summarization for Approach 1

The project employed the LDA text summarization algorithm to generate summaries for the documents collected under each company. This was done to gain a clearer understanding of the topics interpreted by the topic model in previous steps.

Topic 1:

- **Generated topic label:** Ingredients to Build a Powerful Talent Management Platform
- **Topic keywords:** 'culture', 'strategy', 'leadership', 'competency', 'predictive', 'workforce', 'expert', 'science', 'drive', 'research'

Interpretation Process:

1. Get the most associated topic documents from each company

A	B	C	D	E	F
Topic_Num	Topic_Perc_(Keywords)		Text	Name	Topic Label
1	0.9976 'culture', 'strategy', 'leadership', 'compel Perfect for those who have completed their AQai Level 1 Certificati AQai'		Ingredients to Build a Powerful Talent Management Platform		
1	0.9967 'culture', 'strategy', 'leadership', 'compel Chief Learning Officer - CLO Medi Chief Learning Officer is a multi Searchlight.ai		Ingredients to Build a Powerful Talent Management Platform		
1	0.9963 'culture', 'strategy', 'leadership', 'compel Growing up on a farm in Canada, the Aberhart brothers always knei Kolbe		Ingredients to Build a Powerful Talent Management Platform		
1	0.996 'culture', 'strategy', 'leadership', 'compel Recruitment and HR leaders are hiring faster, increasing employee AssessFirst		Ingredients to Build a Powerful Talent Management Platform		
1	0.996 'culture', 'strategy', 'leadership', 'compel Slash Your Sourcing Bills with Sprockets Pools. Learn More. You spe Sprockets		Ingredients to Build a Powerful Talent Management Platform		
1	0.996 'culture', 'strategy', 'leadership', 'compel We cultivate a work environment that is open and collaborative, tei TestGrid		Ingredients to Build a Powerful Talent Management Platform		
1	0.9955 'culture', 'strategy', 'leadership', 'compel English English Search US English German French Arabic Danish Du Talogy		Ingredients to Build a Powerful Talent Management Platform		
1	0.995 'culture', 'strategy', 'leadership', 'compel Our mission is to make hiring the right candidate easier Our missio iMocha		Ingredients to Build a Powerful Talent Management Platform		
1	0.9949 'culture', 'strategy', 'leadership', 'compel Award-Winning Talent Assessment Provider Matches People and Pt Plum		Ingredients to Build a Powerful Talent Management Platform		

Figure 21: sample of most associated company document

Generated Summary:

From the summarized output, we can understand that the text mentions various topics related to talent management, including data insights, interpreting reports, and enhancing team performance. It also includes examples of companies and organizations that have used AQai assessments for employee development and digital readiness. Additionally, there are mentions of AssessFirst's personality, motivation, and reasoning assessments, as well as their contributor program and services for talent acquisition and development. The text concludes with a commitment to achieving tangible ROI and success through talent strategies.

Easy access to all session recordings Deep Dive - AQ Data, Research & Science- Understand correlations, dimension distributions, and advanced data insights- Explain key points regarding the science behind AQme- Answer key questions regarding the science behind AQme- Identify the opportunities from data to achieve business outcomes- Identify the opportunities from data to achieve progress toward goals- Map insights using report data Interpreting AQ Reports With Confidence- Access the additional report functionality data that is available to Level 2 professionals- Interpret the insights gained from the new team dimensions- Connect the insights to how it is applied to various business operating methods- Demonstrate feedback on the team report Communicating AQ Results For Maximum Value & Impact- Connect organizational goals to specific dimensions- Communicate succinctly additional insights unknown to the user- Deliver team feedback with confidence 10X Your Value & Impact- Identify the clients you can position AQ data intelligence and reports with- Map AQ opportunities to your own business goals- Build out your value-proposition in relation to adaptability- Identify opportunities in the AQ trainer journey, AQ marketplace, and Level 2 Roundtables Setup for back into focus, with the key concepts from the AQai Foundations course.

Working closely with our AQ consultants, IMEG took different teams, including recent members of their graduate program, through their assessments to gain a better understanding of where to focus employee development programs to ensure an efficient and productive workforce A specialized agency in industrial development within the United Nations (UN), after investing \$1bn into traditional financial management, used AQai assessment and coaching to reform their back office processes, enhancing their digital readiness as a team.

100% of students tested 86% School's recommendation rate Hiring Candidate Experience Talent development Internal mobility Identifying potential Team performance AssessFirst X Science Shape (Personality) Drive (Motivations) Brain (Reasoning) See our case studies Subscription plans Book a product tour ATS, HRIS and Third-Party Apps Distribute AssessFirst Business Contributor Program HRMission Blog Whitepapers About us Leadership & Management Team Brand Assets Jobs F.A.Q Contact Legal Notice Manage Cookies Terms of Use Terms of Service Privacy PolicySlash Your Sourcing Bills with Sprockets Pools.

Will is committed to his two key goals of ensuring all of his clients: To do so, Will adheres to his 3 principles: pride himself on his abilities in designing talent acquisition and development strategies that provide tangible Return-On-Investment (ROI) to their existing and new clients, in order to enable them to make faster and smarter decisions and empower them to achieve success through their Stuart is a highly experienced professional skilled in Business Planning, Business Development, Operations Management, Learning Management and Customer Relationship Management (CRM).

Figure 22: short summary of topic 1: most associated doc from each company

Overall, we can extract that the topic 1:

- Discusses about to AQme, a tool for measuring and analyzing organizational performance
- Includes insights on the science behind AQme and identifying opportunities to achieve business outcomes using AQ (Adaptability Quotient) data intelligence and reports
- Covers interpreting AQ reports and communicating AQ results for maximum impact
- Examples are given of how organizations have used AQai assessments and coaching to improve their workforce and enhance digital readiness
- Other tools and programs related to talent acquisition and development are mentioned, such as AssessFirst X Science and Sprockets Pools
- Brief descriptions of goals, principles, and professional skills are also provided

4.2.1.3 Incorporating Chat GPT with Approach: 1

Copied the extracted documents for topic 1 most associated document from each company and created a google doc link to use as a reference while using Chat GPT.

- Using Chat GPT, we can further analyze that:
- Chat GPT summarizes the collection of documents with additional hidden information such as the importance of the topic keywords in a Talent management company
- It also emphasizes how an organization can be helped by a Talent management tool
- By incorporating the findings generated in the project and collating Chat GPT findings, topic 1 can be easily interpreted
- The approach produces a very clear interpretation of the topic backed by analytical data



After reviewing the provided Google doc, it appears to be a research paper or article discussing various topics related to organizational culture, leadership, and workforce competency. The paper covers the following topics:

- The importance of organizational culture in shaping employee behavior and performance
- The role of strategy in shaping organizational culture and achieving business goals
- The impact of leadership on organizational culture and employee engagement
- The use of competency frameworks to assess and develop workforce skills and capabilities
- The value of predictive science and expert research in making informed decisions about workforce management
- The use of data analytics and machine learning in workforce planning and management

Overall, the paper provides insights into how organizations can develop and maintain a positive culture, effective leadership, and a competent workforce through the use of various tools and techniques, including data analysis and expert research.



topic label this article



Topic: Strategies for Developing and Managing Organizational Culture, Leadership, and Workforce Competency.



Figure 23: insights from chat GPT for topic 1

4.2.2 Approach 2: Choosing the 50 most associated documents from each company within a topic

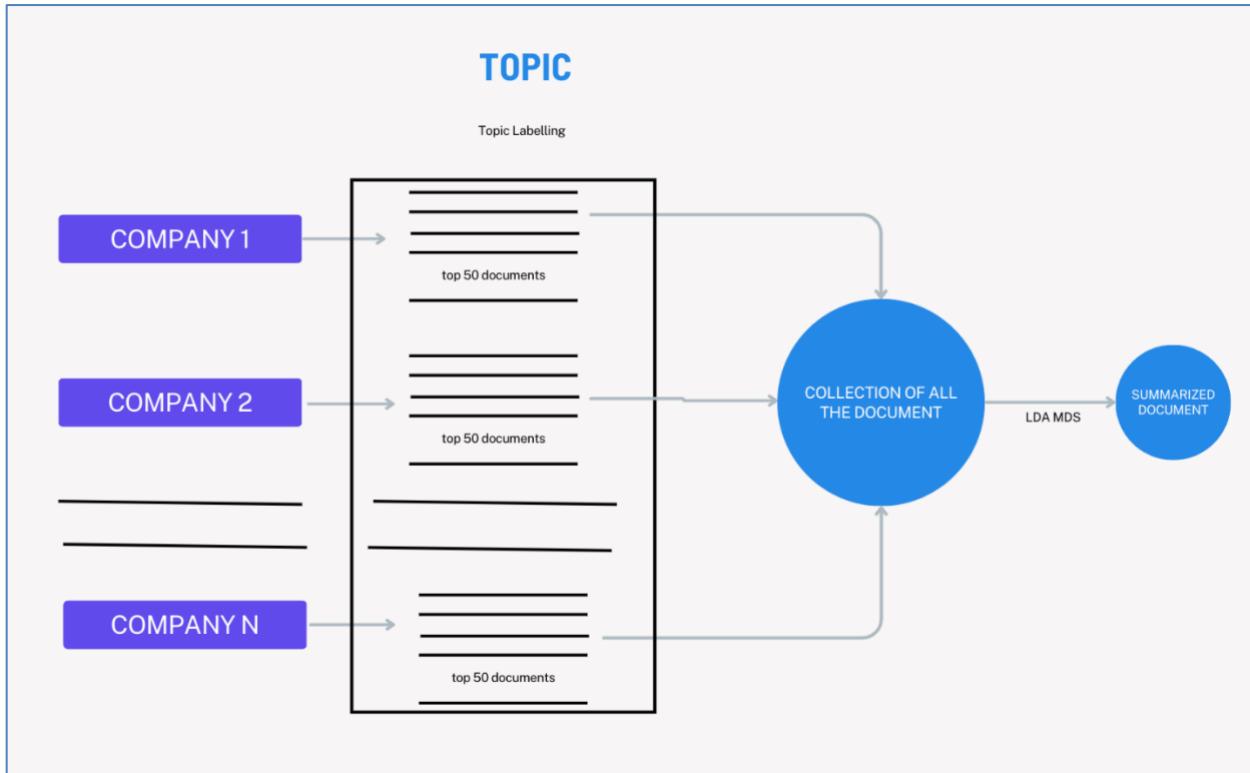


Figure 24: topic interpretation: Approach 2

In this approach, for a given topic the top 50 associated document from each company for the given topic was picked and from the collection of these highly associated documents, the text summarization and topic label interpretation were performed.

Generating topic labels!

Topic 0

Topic Keywrods: 'subscription', 'applicable', 'liability', 'material', 'condition', 'provision', 'responsible', 'damage', 'dispute', 'limitation',

Topic label: Using the Services to Assess the Potential of Candidates

Topic 1

Topic Keywrods: 'culture', 'strategy', 'leadership', 'competency', 'predictive', 'workforce', 'expert', 'science', 'drive', 'research',

Topic label: Adaptability Assessments - The Powerful Kolbe Index

Topic 2

Topic Keywrods: 'course', 'practice', 'minute', 'phone', 'title', 'deimeasure', 'choose', 'comment', 'actually', 'employer',

Topic label: How to Pass the Wonderlic SLE Exam

Topic 3

Topic Keywrods: 'builder', 'coffee', 'trait', 'serving', 'wayne', 'distributor', 'vendor', 'award', 'hundred', 'daunting',

Topic label: How to Improve Your Hiring Process With HighMatch

Topic 4

Topic Keywrods: 'purpose', 'reference', 'protection', 'device', 'crystal', 'thirdparty', 'legal', 'transfer', 'necessary', 'protect',

Topic label: The WAREwithal Platform - A Comprehensive Multi-Rate Feedback Tool

Figure 25: generated topic label: Approach 2

Topic 5

Topic Keywrods: 'questionnaire', 'complete', 'director', 'standard', 'simulation', 'field', 'adaptive', 'staff', 'necessary', 'helped',

Topic label: eSkill - A Great Pre-employment Assessment Tool

Topic 6

Topic Keywrods: 'integrate', 'trial', 'workflow', 'credit', 'board', 'update', 'refer', 'anywhere', 'smarter', 'navigate',

Topic label: Hire With Aims - Discover the Best Hire With Aims A Step A

Topic 7

Topic Keywrods: 'proctoring', 'remote', 'examination', 'secure', 'certification', 'feature', 'automated', 'recruiting', 'proctor', 'screen',

Topic label: Mercer | Mettl - The Best Platform for Online Exams

Topic 8

Topic Keywrods: 'campus', 'booking', 'efficient', 'expert', 'description', 'succeed', 'productive', 'manage', 'domain', 'coach',

Topic label: How to Boost Your Recruiting Efficiency With Online Coding Tests

Figure 26: generated topic label: Approach 2 (contd...)

4.2.2.1 Grouping companies based on the generated label

Grouping of companies based on topic labels involves clustering companies together based on their topic contributions in a topic model. Companies that have similar topic distributions or have a higher contribution to specific topics can be grouped together to identify patterns or trends within the dataset.

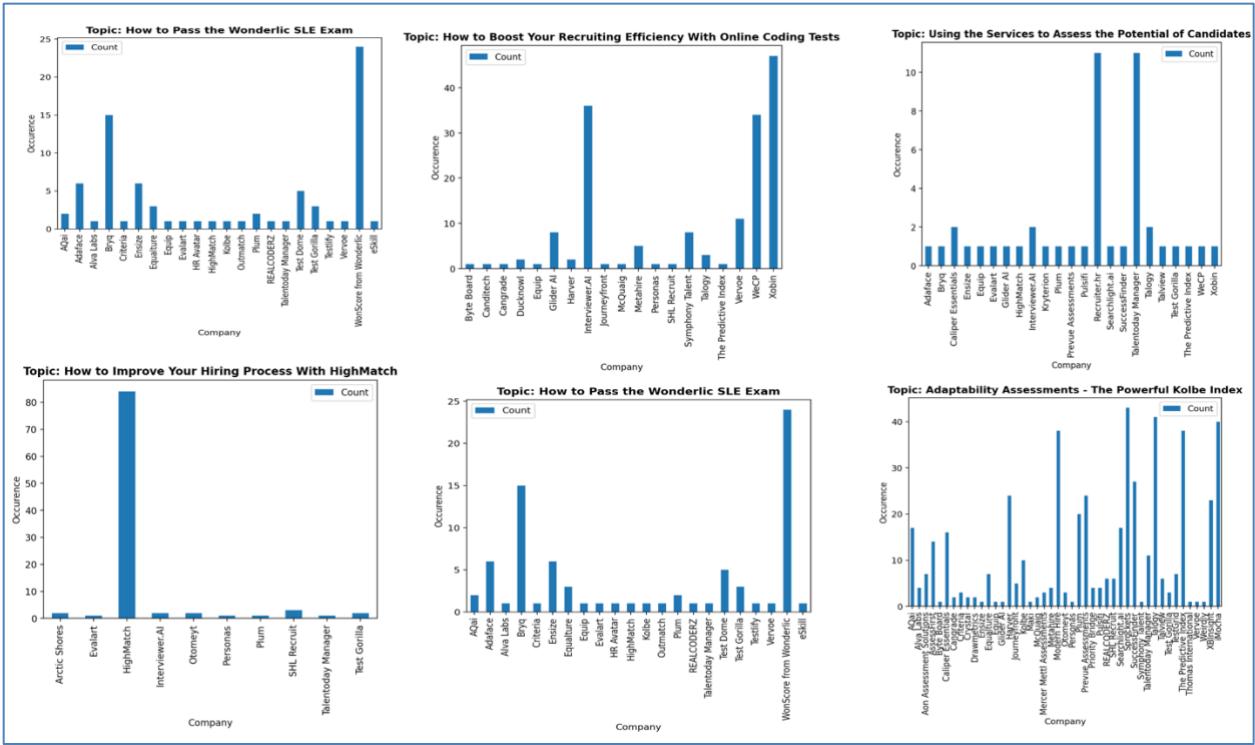


Figure 27: companies grouped based on the topic label: Approach 2

This approach has resulted in more detailed labels due to the increased amount of data supporting each topic label.

For example:

Topic: How to boost your recruiting efficiency with an online coding test

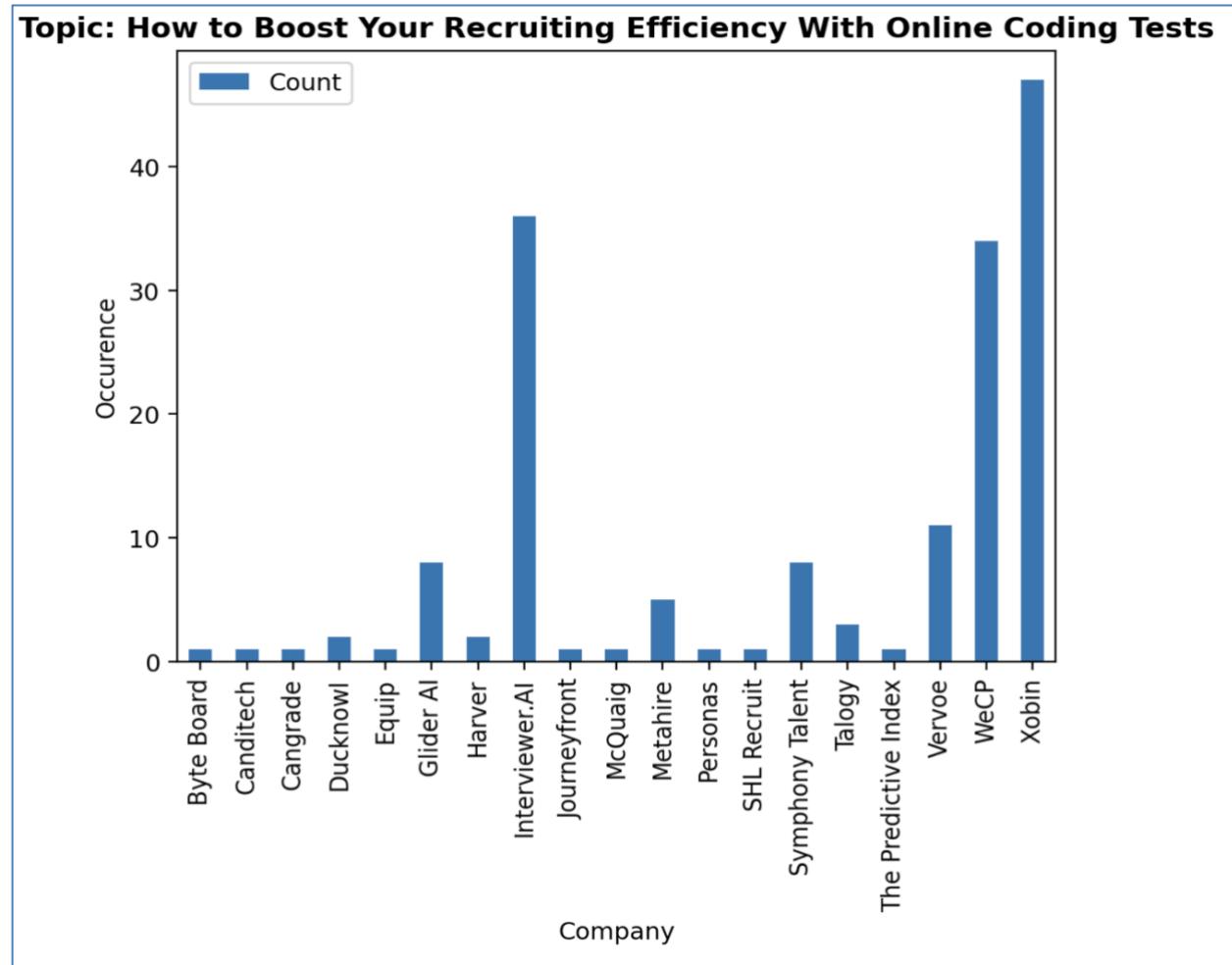


Figure 28: companies representing topic: online coding test

An online coding test can be a valuable tool to boost recruiting efficiency by providing a streamlined process for evaluating candidates' coding skills. The test can be customized to match specific job requirements and can help to filter out unqualified candidates early in the hiring process. This can save recruiters time and effort, as they can focus on the most promising candidates. Additionally, an online coding test can help to identify top candidates who may have been overlooked in traditional resume screening methods. It also allows for remote testing, which can be especially beneficial in today's virtual work environment. Overall, using an online

coding test can be an effective way to increase recruiting efficiency and find the best candidates for a position.

Overall, looking at the above graph three companies namely:

- Interviewer.ai
- WeCP
- Xobin

Are focusing more on online coding test for hiring the best talent.

The previous approach only selected one document from each company, which may have missed out on important details and nuances related to each company's focus on talent management. However, by using a larger sample size of documents from each company in the second approach, more detailed and nuanced topic labels could be generated, providing a better understanding of each company's focus within the talent management domain.

4.2.2.2 Automated multi-document text summarization for Approach 2

Topic 1:

- **Generated Label:** Adaptability assessment: A powerful Kolbe index

Topic Keywords: 'culture', 'strategy', 'leadership', 'competency', 'predictive', 'workforce', 'expert', 'science', 'drive', 'research'

1	0.9976 'culture', 'strategy', 'leadership', 'competency', 'predi Perfect for those who have completed their AQai Level 1 Certification and want to become a fully accredited AQai pro AQai
1	0.9972 'culture', 'strategy', 'leadership', 'competency', 'predi A training designed for corporations looking to improve their people's relationship with change Take your career to the AQai
1	0.9967 'culture', 'strategy', 'leadership', 'competency', 'predi Chief Learning Officer - CLO Media Chief Learning Officer is a multimedia publication focused on the importance, ben Searchlight.ai
1	0.9963 'culture', 'strategy', 'leadership', 'competency', 'predi Growing up on a farm in Canada, the Aberhart brothers always knew they were different than other members of their Kolbe
1	0.996 'culture', 'strategy', 'leadership', 'competency', 'predi Recruitment and HR leaders are hiring faster, increasing employee performance, driving diversity and reducing staff c AssessFirst
1	0.996 'culture', 'strategy', 'leadership', 'competency', 'predi Slash Your Sourcing Bills with Sprockets Pools. Learn More. You spend hours fine-tuning a job description, review resu Sprockets
1	0.996 'culture', 'strategy', 'leadership', 'competency', 'predi We cultivate a work environment that is open and collaborative, team oriented and supportive, most importantly, infc TestGrid
1	0.9955 'culture', 'strategy', 'leadership', 'competency', 'predi English English Search US English German French Arabic Danish Dutch Norwegian Swedish Chinese UK English Arabic Tatalogy
1	0.995 'culture', 'strategy', 'leadership', 'competency', 'predi Our mission is to make hiring the right candidate easier Our mission is to make hiring the right candidate easier Our iMocha
1	0.995 'culture', 'strategy', 'leadership', 'competency', 'predi Our mission is to make hiring the right candidate easier Our mission is to make hiring the right candidate easier Our iMocha
1	0.9949 'culture', 'strategy', 'leadership', 'competency', 'predi Award-Winning Talent Assessment Provider Matches People and Positions, Building Productive and Fulfilled Workforce Plum
1	0.9949 'culture', 'strategy', 'leadership', 'competency', 'predi Our mission is to make hiring the right candidate easier Our mission is to make hiring the right candidate easier Our iMocha
1	0.9949 'culture', 'strategy', 'leadership', 'competency', 'predi Our mission is to make hiring the right candidate easier Our mission is to make hiring the right candidate easier Our iMocha
1	0.9948 'culture', 'strategy', 'leadership', 'competency', 'predi Our mission is to make hiring the right candidate easier Our mission is to make hiring the right candidate easier Our iMocha
1	0.9946 'culture', 'strategy', 'leadership', 'competency', 'predi Slash Your Sourcing Bills with Sprockets Pools. Learn More. Sprockets is the missing ingredient in your HR tech recipe, Sprockets
1	0.9946 'culture', 'strategy', 'leadership', 'competency', 'predi English English Search US English German French Arabic Danish Dutch Norwegian Swedish Chinese UK English Arabic Tatalogy
1	0.9946 'culture', 'strategy', 'leadership', 'competency', 'predi Our mission is to make hiring the right candidate easier Our mission is to make hiring the right candidate easier Our iMocha
1	0.9946 'culture', 'strategy', 'leadership', 'competency', 'predi Our mission is to make hiring the right candidate easier Our mission is to make hiring the right candidate easier Our iMocha
1	0.9942 'culture', 'strategy', 'leadership', 'competency', 'predi Slash Your Sourcing Bills with Sprockets Pools. Learn More. Insights and Inspiration Explore our collection of resources Sprockets
1	0.9942 'culture', 'strategy', 'leadership', 'competency', 'predi Slash Your Sourcing Bills with Sprockets Pools. Learn More. Insights and Inspiration Explore our collection of resources Sprockets
1	0.9942 'culture', 'strategy', 'leadership', 'competency', 'predi Slash Your Sourcing Bills with Sprockets Pools. Learn More. Insights and Inspiration Explore our collection of resources Sprockets
1	0.9942 'culture', 'strategy', 'leadership', 'competency', 'predi Slash Your Sourcing Bills with Sprockets Pools. Learn More. Insights and Inspiration Explore our collection of resources Sprockets
1	0.9942 'culture', 'strategy', 'leadership', 'competency', 'predi Modern Hire combines trusted science and proven technology to solve today's hiring challenges. Our job-specific Solut Modern Hire
1	0.9941 'culture', 'strategy', 'leadership', 'competency', 'predi English English Search US English German French Arabic Danish Dutch Norwegian Swedish Chinese UK English Arabic Tatalogy
1	0.9939 'culture', 'strategy', 'leadership', 'competency', 'predi Slash Your Sourcing Bills with Sprockets Pools. Learn More. solution creates a unique success profile based on your cu Sprockets
1	0.9939 'culture', 'strategy', 'leadership', 'competency', 'predi made it! Welcome to A World. For over 15 years our team has brought a spark to brands around the world. We are ex McQuaig

Figure 29: sample of the top 50 associated company document

Generated summary:

Easy access to all session recordings Deep Dive - AQ Data, Research & Science- Understand correlations, dimension distributions, and advanced data insights- Explain key points regarding the science behind AQme- Answer key questions regarding the science behind AQme- Identify the opportunities from data to achieve business outcomes- Identify the opportunities from data to achieve progress toward goals- Map insights using report data Interpreting AQ Reports With Confidence- Access the additional report functionality data that is available to Level 2 professionals- Interpret the insights gained from the new team dimensions- Connect the insights to how it is applied to various business operating methods- Demonstrate feedback on the team report Communicating AQ Results For Maximum Value & Impact- Connect organizational goals to specific dimensions- Communicate succinctly additional insights unknown to the user- Deliver team feedback with confidence 10X Your Value & Impact- Identify the clients you can position AQ data intelligence and reports with- Map AQ opportunities to your own business goals- Build out your value-proposition in relation to adaptability- Identify opportunities in the AQ trainer journey, AQ marketplace, and Level 2 Roundtables Setup for back into focus, with the key concepts from the AQ Foundations course. Working closely with our AQ consultants, IMEG took different teams, including recent members of their graduate program, through their assessments to gain a better understanding of where to focus employee development programs to ensure an efficient and productive workforce A specialized agency in industrial development within the United Nations (UN), after investing 1bn into traditional financial management, used AQai assessment and coaching to reform their back-office processes, enhancing their digital readiness as a team. Working closely with our AQ consultants, IMEG took different teams, including recent members of their graduate program, through their assessments to gain a better understanding of where to focus employee development programs to ensure an efficient and productive workforce. A specialized agency in industrial development with in the United Nations (UN), after investing 1bn into traditional financial management, used AQai assessment and coaching to reform their back office processes, enhancing their digital readiness as a team. 100% of students tested 86% School's recommendation rate Hiring Candidate Experience Talent development Internal mobility Identifying potential Team performance AssessFirst X Science Shape (Personality) Drive (Motivations) Brain (Reasoning) See our case studies Subscription plans Book a product tour ATS, HRIS and Third-Party Apps Distribute AssessFirst Business Contributor Program HRMission Blog Whitepapers About us Leadership & Management Team Brand Assets Jobs F.A.Q Contact Legal Notice Manage Cookies Terms of Use Terms of Service Privacy PolicySlash Your Sourcing Bills with Sprockets Pools.

Figure 30: short summary: top 50 associated docs from each company

From the summary, it can be easily extracted that:

The text includes information about the AQme platform, including sessions on data insights, interpreting AQ reports, and communicating AQ results. The AQai assessment and coaching tool has been used by different teams, including a specialized agency in the UN, to improve employee development and digital readiness. The text also mentions the AssessFirst X Science

tool for evaluating personality, motivations, and reasoning in candidates, as well as subscription plans and integration with HR software. The text ends with a brief mention of Sprockets Pools, a tool for reducing sourcing costs.

The summary discusses the benefits and opportunities of using AQme, a data intelligence platform.

- The platform offers access to session recordings, data insights, and report functionality for advanced data interpretation and analysis.
- The platform also provides team dimensions, which can be used to improve employee development programs and team performance.
- AQai assessment and coaching can be used to enhance digital readiness in organizations.
- The platform can be used for talent development, identifying potential candidates, and improving the hiring candidate experience.
- The article also mentions AssessFirst X Science, which focuses on personality, motivations, and reasoning to aid in talent development.
- The article includes case studies, subscription plans, and information on integrating with third-party apps.
- Sprockets Pools is mentioned as a way to save on sourcing bills.

4.2.2.3 Incorporating Chat GPT with Approach: 2

Copied the extracted documents from Step 1 and created a google doc link to use as a reference while using Chat GPT.

By incorporating the findings generated in the project and collating Chat GPT findings, topic 1 can be easily interpreted:

The approach produces a very clear interpretation of the topic backed by analytical data. Compared to Approach 1, this approach can be more effective in interpreting the topic label if the first 50 documents are randomly distributed across the companies.

However, this approach requires more time and processing power compared to Approach 1 due to the increased number of documents.



Based on the document provided, some possible discussion points or topics could include:

1. The importance of aligning organizational culture with business goals and values, and strategies for creating a positive workplace culture.
2. The role of leadership in driving workforce development initiatives and creating a culture of continuous learning and improvement.
3. Identifying and developing key competencies among employees to improve performance and increase engagement, including strategies for assessing and measuring competencies.
4. The use of data-driven approaches and predictive analytics to inform talent management decisions and identify future organizational needs, including strategies for collecting, analyzing, and utilizing workforce data.
5. Attracting and retaining top talent, including strategies for recruiting diverse candidates and building a strong employer brand.
6. The importance of ongoing professional development and training for employees, and strategies for designing and implementing effective training programs.
7. The role of technology in workforce development and talent management, including the use of learning management systems, performance management software, and other tools.



[https://docs.google.com/document/d/1uSqvFqnf2pj_oIYPKXaRQJlgSOuUZSf_7c0mEMFujSo/edit?
usp=sharing](https://docs.google.com/document/d/1uSqvFqnf2pj_oIYPKXaRQJlgSOuUZSf_7c0mEMFujSo/edit?usp=sharing)



Based on the content in the Google Doc provided, a possible topic label for the article could be "Strategies and Best Practices for Workforce Development and Talent Management." Upvote Downvote

Figure 31: Chat GPT insights for the topic

4.2.3 Approach 3: Choosing all the documents from each company within a topic

This approach involves selecting all the documents related to a particular company within a given topic, rather than just selecting one document per company. This allows for more detailed topic labeling due to the larger amount of data supporting each company. This approach may result in a larger amount of data to analyze, but it can provide more comprehensive insights into the performance and practices of each company within a particular topic.

Topic: Using REST API to Manage a Questionnaire

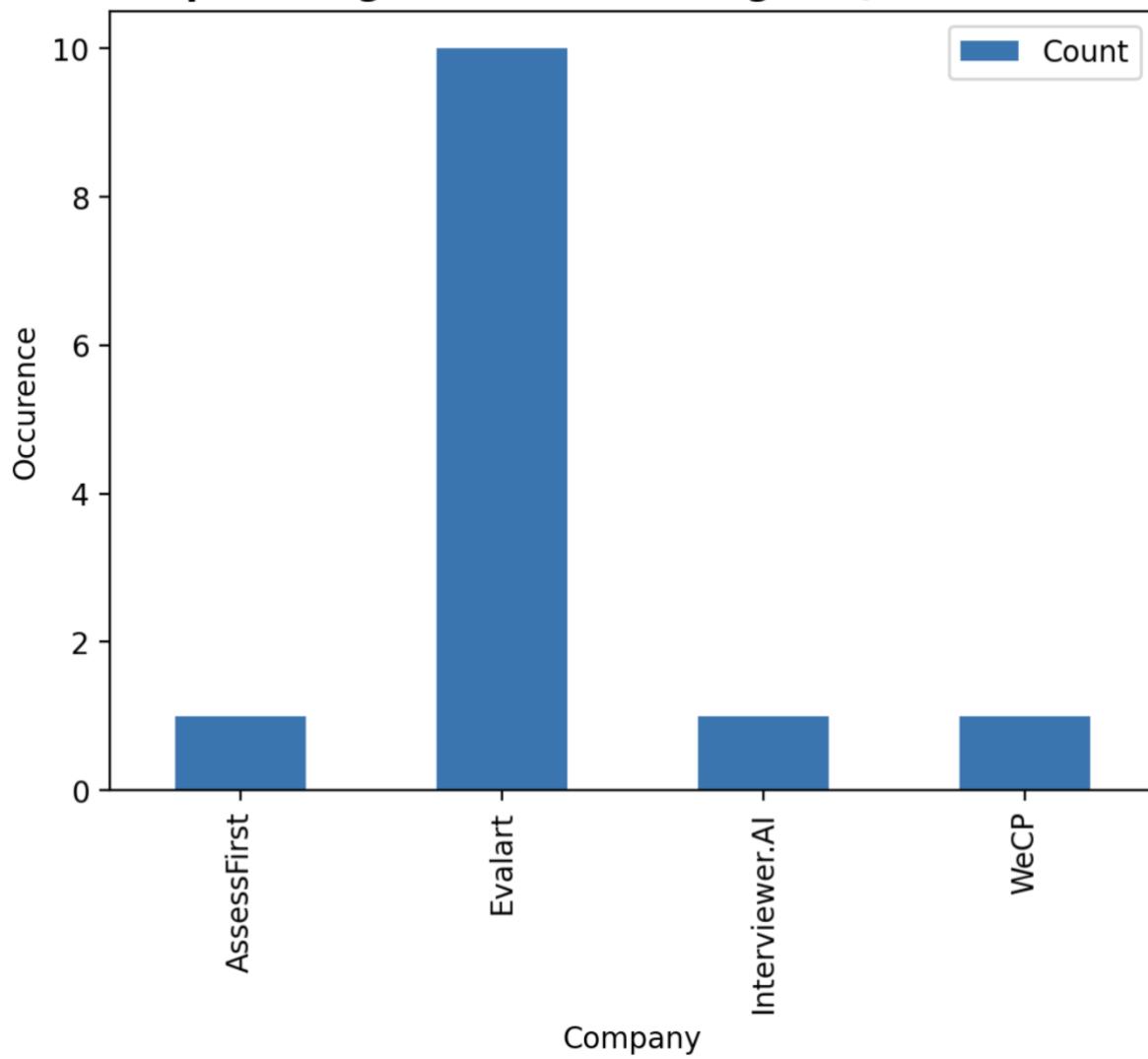


Figure 32: companies grouped on a topic

API can be used to automate the administration and scoring of the questionnaires, allowing for a more efficient recruitment process. Responses to the questionnaires can be collected automatically and scored according to pre-defined criteria, allowing for a faster and more accurate evaluation of candidates. API can be further used to integrate the questionnaire into existing recruitment software or workflows. This can include automatically triggering the questionnaire at specific points in the recruitment process, such as after a candidate submits an application, and integrating the questionnaire results into candidate profiles and evaluations.

Overall, using Rest API to manage questionnaires for talent recruitment can help companies to streamline and optimize their recruitment processes, allowing them to find and hire the best candidates more efficiently and effectively.

Overall, using a larger number of documents can provide more insights from the topic model result, which can be missed if only a small number of documents are used. More data helps in capturing a broader range of perspectives, ideas, and language patterns, which can lead to more accurate and comprehensive topic modeling results. Additionally, a larger dataset also helps in reducing the impact of noise and outliers and provides a more robust and stable topic model.

4.3 Summary of this chapter

This section presents the deliverables of the research project, which collectively provides the outcome and the process of achieving our deliverables. We have produced the summaries which will help TIM students to perform an analysis on the essence of the topic words that have been produced by the topic modelling, text summarization and automated topic labeling to help interpret the topic better.

5 Analysis and discussion

In this chapter, the problem addressed by this research project is introduced, along with the solutions it proposes. The discussion focuses on the findings presented in Chapter 4, which demonstrate how the existing technique of topic modeling, combined with natural language processing, can be leveraged to fulfil the client's requirements, specifically the TIM program. The section explains how the results obtained from topic modeling can be interpreted with the help of natural language processing techniques.

5.1 Problem and gap addressed

Topic modeling is a technique used to identify patterns and themes in a collection of documents. However, users and modellers may have different interpretations of the results generated by the topic model, which can make it difficult to understand the true topics and their relevance to the documents. This is because the output typically consists of a list of words that serve as "themes," and it may be challenging to determine which theme is most relevant and what is being discussed. To address this issue, an appropriate method has been selected to help users obtain more accurate results and better interpret the topic model output, ultimately helping them achieve their objectives.

5.2 Interpretation of the results

This project aimed to identify the different approaches used for topic labeling/interpretation methods. A comprehensive literature review was conducted to accomplish this objective, utilizing scholarly articles and journals from various libraries and sources such as Google Scholar. The review uncovered a significant gap in the existing methods of interpreting topic models, which was the motivation for this project. Researchers and TIM students faced challenges when attempting to interpret topic models, which is why the first deliverable aimed to address these limitations by outlining the method and process behind topic models and proposing solutions to these limitations.

The research project involved scraping data from 75 talent management companies' URLs, which was used for topic modeling and text summarization to extract meaningful insights. The project aimed to address the challenge of interpreting topic models and identify the best text analytics or interpretation method for the context of the experiment.

5.3 Comparison between the different approaches used for the topic interpretation

To compare the interpretation from three different approaches, the result was analyzed, and it was evident that more documents within a topic helped uncover hidden details and better topic interpretation. However, doing so required a significant amount of computation and effort. The study also highlighted that Chat GPT could help in various ways, such as summarizing text and gathering insights about a topic using the document at a large extent. The results showed that Chat GPT could assist in understanding topics and providing insights quickly and efficiently.

5.3.1 Approach 1

The approach of grouping companies and identifying common themes through an automated topic labeling process has been successful in interpreting topic labels. This approach has the advantage of being more efficient in analyzing a large corpus of talent management data and identifying patterns and trends.

By extracting the most associated documents from each topic, approach 1 provides valuable insights into which companies are focused on which topic, allowing for a better understanding of the talent management landscape. It is worth noting that since only one document per company per topic was selected, it requires fewer computational resources and power, and has proven to be quite effective in producing the topic interpretation in large amounts of data.

However, the downside of using this approach is that it may miss out on key documents that have an equal or slightly less percentage of topic contribution but contain more important information about the topic, leading to a potentially misleading interpretation of the topic.

Additionally, since only one document is selected per company per topic, there is a possibility of losing important context and variability within the company's talent management practices.

This approach is suitable when there is a large amount of data, and a broad overview of the topics is required. It provides an efficient way to analyze a large corpus of talent management data and identify patterns and trends. However, it may not be suitable when detailed information on specific documents or subtopics is required. For a more detailed overview, where the computational time and resources are not limited, it may be beneficial to consider other approaches as well.

5.3.2 Approach 2

This approach focuses on identifying the most relevant documents within each topic by selecting the top 50 documents per company per topic. This provides a more accurate and detailed view of the topic, allowing for a deeper understanding of the talent management landscape. The results obtained using this approach are more promising and effective, as they capture a larger amount of hidden information within each topic.

However, this approach requires more computational resources and time to produce the results as it involves analyzing a larger amount of data. It is more suitable for larger datasets, where the amount of information is substantial and there are many relevant documents for each topic.

However, the downside of this approach is that it may miss out on some key documents that have an equal or slightly less percentage contribution to the topic but may contain more important information about the topic. This can lead to a misleading interpretation of the topic, especially if these key documents are not included in the top 50 associated documents.

Furthermore, this approach may not be suitable for smaller datasets, as the top 50 documents may contain repetitive information and may not provide enough variety to accurately capture the main themes and subtopics within the data. Overall, while approach 2 is a useful method for analyzing large datasets, it is important to exercise caution and consider the limitations of the approach when interpreting the results.

5.3.3 Approach 3

The third approach used in this project involved selecting documents dynamically based on a selected threshold, which was set at 70% for the topic percentage contribution of each document. This approach is more practical than the previous two, as it allows for the selection of documents that are highly associated with the topic being analyzed.

This approach is capable of producing interpretations that may not be extracted using the previous two approaches since it considers the topic percentage contribution of each document. The data selected in this approach is more relevant and highly associated with the topic, which makes it more likely to produce accurate interpretations.

However, the downside of this approach is that it requires a lot of computational resources and time, depending on the size of the dataset. Despite this drawback, the results obtained from this approach are more promising and accurate than the other two approaches. This approach is especially effective when dealing with a respectable amount of data.

When using the approach of selecting the top associated documents based on a cut-off threshold, it is important to carefully consider the chosen threshold value. If the threshold is set too low, there may be a risk of including repetitive or irrelevant documents that can skew the topic interpretation. On the other hand, if the threshold is set too high, some topics may not have a sufficient number of associated documents, which can lead to an incomplete or inaccurate understanding of the topic.

Therefore, finding the optimal cut-off threshold value that balances the number of associated documents and the relevance of the topic is important for producing reliable and meaningful topic interpretations. This requires careful experimentation and tuning, as different datasets and research questions may require different threshold values.

In conclusion, the third approach used in this project is the most practical and effective in extracting relevant and highly associated data. However, its drawback is that it requires a lot of computational resources and time, which is why it is recommended to use this approach only when dealing with a large dataset.

5.4 Summary of the topic model interpretation approaches

The three approaches used in this project for topic modeling in talent management data are:

- **Selecting the most associated document per company per topic**

This approach is efficient in analyzing large corpus of data and provides insights into which companies are focused on which topic. It is effective in identifying patterns and trends and requires minimal computational resources. However, it only selects one document per company per topic, which can miss out on important information.

- **Selecting the top document per company per topic.**

This approach is effective in producing topic interpretation, and requires optimal computational resources and time. It extracts the most associated documents from each topic, providing valuable insights into which companies are focused on which topic. However, if the dataset is too small, the top 50 associated documents may seem redundant and can mislead the interpretation result.

- **Using a topic percentage contribution threshold.**

This approach is practical and dynamic, as the documents are selected based on the topic percentage contribution cutoff threshold. It is capable of producing interpretations that may not be extracted using the previous two approaches, as the data is more relevant and highly associated. However, it requires a lot of computational resources and time, depending on the dataset size.

The choice of which approach is better depends on the specific goals and resources of a project. If the dataset is large and computational resources are limited, approach 1 or approach 2 may be more suitable since they require less computation power and can still provide valuable insights. However, if computational resources are not a concern and a more accurate and comprehensive interpretation of the data is desired, approach 3 may be the best option. Approach 3 has the advantage of dynamically selecting associated documents based on the topic percentage

contribution, resulting in a more relevant and highly associated dataset for analysis. However, it requires more computational resources and may not be practical for smaller datasets.

Overall, it is important to consider the trade-offs between accuracy and efficiency when choosing an approach for topic labeling and interpretation, and to carefully evaluate the results to ensure they align with the goals of the project.

5.5 Link to the literature

The project has utilized the insights gathered from the literature review to address the challenges faced in topic modelling, which is the difficulty in interpreting the results. The project has identified the gap and proposed a solution by selecting an appropriate interpretation method. The project's findings have contributed to the existing literature by refining the process and bridging the gaps identified in the literature review. The literature review has been an integral part of the project, providing a foundation for understanding the present techniques of topic modeling and topic labelling techniques and how they have been implemented using various methods. Overall, the project's results and work are closely related to the literature review for this project, which has helped in identifying the challenges faced in topic modeling and proposing a solution by selecting an appropriate interpretation method. The project has contributed to the existing literature by refining the process and bridging the gaps identified in the literature review.

5.6 Summary of this chapter

The project's main goal is to improve the topic modelling interpretation process for large datasets of text documents. To achieve this, an improved process has been identified and a new method has been developed and implemented. The primary objective is to reduce the time and effort

required by researchers and users to obtain meaningful insights from large amounts of text data without sacrificing the quality of information. The aim is to provide them with valuable insights that are good enough to meet the purpose of their study. By achieving this goal, the project aims to bridge the gap between the limitations of current topic modelling techniques and the needs of researchers and users in various fields, such as talent management. The project also involves a literature review to understand the current state of topic modelling techniques and to identify gaps in the field that need to be addressed. The results of the project demonstrate that more documents within a topic lead to better topic interpretation, but this requires significant computation and effort. Additionally, the project shows that Chat GPT can help with various aspects of text analysis, such as summarizing text and extracting insights about a topic from large amounts of documents. Overall, the project contributes to improving the process of topic modelling and interpretation, making it easier and more efficient for users to obtain valuable insights from large amounts of text data.

6 Conclusions, Limitations, and Future Research

This section is divided into three sections. Section 6.1 provides the conclusion of this project, section 6.2 discusses the limitations of this project and last section 6.3 provides the suggestions for future research.

6.1 Conclusions

In summary, this chapter highlights that the project's main objective was to improve the interpretation process of topic models to provide more meaningful insights to users. The results of the experiment showed that using a topic label generator alongside multi document text summarizer with topic modeling can help achieve this objective. The manual approach currently used by researchers is time-consuming and may not provide enough depth of information for the user's needs. However, the experiment conducted in this project has demonstrated that the use of a summarizer can improve the interpretation of topic models and provide more actionable insights for users. The insights obtained from the experiment on the given dataset revealed new industry trends and recent developments in the field of talent management domain. The chapter concludes that this project has successfully addressed the problem of interpreting topic models and has provided a solution that can be applied in various fields of study.

6.2 Limitations

The limitations of the project include:

- **Dataset size:** The project was limited by the size of the dataset used. The insights generated may not be representative of the entire talent management industry
- **Language:** The project was conducted only in the English language, and may not be applicable to other languages

- **Topic model quality:** The quality of the topic models generated depends on the parameters used, and the number of topics chosen. The accuracy and usefulness of the results may be impacted if these parameters are not chosen correctly
- **Interpretation:** The interpretation of the results is subjective and depends on the user's understanding of the domain and the context of the data. Different users may interpret the same results differently
- An additional limitation of this project could be the requirement for **computational power** to train and run the text summarizer model. As the dataset size and complexity increase, the computational requirements for training and using the summarizer model will also increase. This may limit the scalability of the approach for larger datasets or for users with limited access to powerful computing resources
- Furthermore, the **performance of the summarizer model** may also be limited by the quality and diversity of the training data used to develop the model. If the training data is not representative of the target dataset or does not contain enough diversity in terms of language, domain, and topics, the summarizer model may not perform well on new and unseen data. Thus, further improvements and training of the text summarizer model may be necessary to address these limitations

6.3 Future Research

There are several avenues for future research related to this project:

Firstly, the summarization model used in this project can be further improved by fine-tuning it on domain-specific data, which can result in better-quality summaries. Secondly, the topic modeling process can be optimized by exploring different algorithms and techniques, such as Latent Dirichlet Allocation (LDA) or Non-negative Matrix Factorization (NMF) and comparing their

effectiveness in identifying meaningful topics in large datasets.

Another potential area for future research is exploring the use of different machine learning algorithms for the task of topic labelling and interpretation, such as deep learning models like Transformers. These models have shown promising results in natural language processing tasks and may be able to provide more accurate and nuanced topic interpretations.

Additionally, this project has been focused on the talent management industry, but the methodology can be applied to other industries or domains to extract meaningful insights from large datasets. Finally, future research can also explore the use of interactive visualizations and dashboards to enable users to explore and analyze topic models in a more intuitive and user-friendly manner.

Reference

- Alkhodair, S. A., Fung, B. C. M., Rahman, O., & Hung, P. C. K. (2018) "Improving Interpretations of Topic Modeling in Microblogs." *Journal of the Association for Information Science and Technology* 69.4. 528–540.
- Alokaili, A., Aletras, N., & Stevenson, M. (2020). Automatic Generation of Topic Labels. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1965–1968.
<https://doi.org/10.1145/3397271.3401185>
- Anisin, A. (2021). How to use GPT-3 for text summarization.
<https://towardsdatascience.com/how-to-use-gpt-3-for-text-summarization-903dd6a056a6>
- Anon (2018) Study Findings on Psychology Are Outlined in Reports from University of Ljubljana (The Role of Talent Management Comparing Medium-Sized and Large Companies - Major Challenges in Attracting and Retaining Talented Employees).
NewsRX LLC
- Aweisi, A., Arora, D., Emby, R., Rehman, M., Tanev, G., & Tanev, S. (2021). Using Web Text Analytics to Categorize the Business Focus of Innovative Digital Health Companies. *Technology Innovation Management Review*, 11(7/8), 65–78.
- Boštjančič, E., & Slana, Z. (2018). The Role of Talent Management Comparing Medium-Sized and Large Companies - Major Challenges in Attracting and Retaining Talented Employees. *Frontiers in Psychology*, 9, 1750–1750.
<https://doi.org/10.3389/fpsyg.2018.01750>
- Cano Basave, Amparo & Xu, Ruifeng. (2014). Automatic labelling of topic models learned from Twitter by summarisation. In: The 52nd Annual Meeting of the Association for

Computational Linguistics: *Proceedings of the Conference: Volume 2: Short Papers, Association for Computational Linguistics (ACL)*, pp. 618–624

Carpineti, M. (2020). Word clouds: A quick overview. <https://towardsdatascience.com/word-clouds-a-quick-overview-3c28e84a4077>

Chuang, J., Manning, C. D., & Heer, J. (2012). Termite: Visualization techniques for assessing textual topic models. *Proceedings of the International Working Conference on Advanced Visual Interfaces*, 74-77. DOI: 10.1145/2254556.2254572

Grootendorst, M. (2020) Topic modeling with bert, Medium. Towards Data Science. Available at: <https://towardsdatascience.com/topic-modeling-with-bert-779f7db187e6> (Accessed: April 4, 2023).

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, 78(11), 15169–15211.

Kaliannan, M., Darmalinggam, D., Dorasamy, M., & Abraham, M. (2023). Inclusive talent development as a key talent management approach: A systematic literature review. *Human Resource Management Review*, 33(1), 100926–.

Khan, Qaisar & Chua, Hui Na. (2021). An Automated Topics Labelling Framework Using Zero-Shot Text Classification. *Journal of Engineering Science and Technology*. 2021. 46 - 59.

Kou, W., Li, F., & Baldwin, T. (2016). Automatic Labelling of Topic Models Using Word Vectors and Letter Trigram Vectors. *Information Retrieval Technology*, 253–264

Lau, J & Grieser, K & Newman, D & Baldwin, T. (2011). Automatic Labelling of Topic Models.. 1536– 1545.

Lund, B & Wang, T. (2023). Chatting about ChatGPT: How may AI and GPT impact academia and libraries?. 40. 10.1108/LHTN-01-2023-0009.

Mao, Y., Lu, Z., & Yang, Y. (2012). Automatic labelling of hierarchical topics. *Journal of the American Society for Information Science and Technology*, 63(12), 2423-2437.

Pereira, V., Collings, D. G., Wood, G., & Mellahi, K. (2022). Evaluating talent management in emerging market economies: societal, firm and individual perspectives. International *Journal of Human Resource Management*, 33(11), 2171–2191.

<https://doi.org/10.1080/09585192.2022.2067941>

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv.org*.

Rosner, F., Hinneburg, A., Röder, M., Nettling, M., & Both, A. (2014). Evaluating topic coherence measures. *arXiv.org*.

Roul, R. K., Mehrotra, S., Pungaliya, Y., & Sahoo, J. K. (2018). A New Automatic Multi-document Text Summarization using Topic Modeling. *Distributed Computing and Internet Technology*, 212–221. https://doi.org/10.1007/978-3-030-05366-6_17.

Sinha, R. (2020) Latent dirichlet allocation and topic modelling, Medium. Analytics Vidhya.
Available at: <https://medium.com/analytics-vidhya/latent-dirichlet-allocation-and-topic-modelling- eea49dc3eea6> (Accessed: April 4, 2023).

Song, Y., Pan, S., Liu, S., Zhou, M., & Qian, W. (2009). Topic and keyword re-ranking for LDA-based topic modeling. *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, 1757–1760

Tang, X., Mou, H., Liu, J. (2021). Research on automatic labeling of imbalanced texts of customer complaints based on text enhancement and layer-by-layer semantic matching. *Sci Rep* 11, 11849 <https://doi.org/10.1038/s41598-021-91189-0>

Truica, C.-O., & Apostol, E.-S. 2021. TLATR: Automatic Topic Labeling Using Automatic (Domain-Specific) Term Recognition. *IEEE Access*, 9, 76624–76641

Twinandilla, S., Adhy, S., Surarso, B., & Kusumaningrum, R. (2018). Multi-Document Summarization Using K-Means and Latent Dirichlet Allocation (LDA) – Significance Sentences. *Procedia Computer Science*, 135, 663–670

Appendix

Python libraries used in this project:

- Natural language toolkit
- Pandas
- Matplot
- Seaborn
- Word cloud
- Regular Expression
- Gensim
- PPrint
- pyLDAvis
- PIL
- Torch
- Transformers
- Random
- CSV
- Numpy
- LDA Summarizer
- BERT Topic labeler
- Streamlit

python code for web scrapper

```
1 # import library
2 from bs4 import BeautifulSoup
3 import requests
4 import re
5 from urllib.parse import urlparse
6 from StoreToCSV import saveToCSV
7 from django.core.validators import URLValidator
8 from django.core.exceptions import ValidationError
9 import csv
10 import streamlit as st
11
12 from ReadCSV import *
13
14 HTTPS= "https://"
15 COUNT = 0
16 header = ["name", "url", "content"]
17
18 company_list=[]
19
20 # scraped_data_dict = {
21 #     "company": "",
22 #     "url": "",
23 #     "data": ""
24 # }
25 def increment():
26     global COUNT
27     COUNT = COUNT+1
28     print(COUNT)
29
30 scraped_data_list=[]
31
32 def write_datas():
33     with open('./input/scraped_data.csv', 'w', newline='') as csv_file:
34         writer = csv.writer(csv_file)
35         writer.writerow(header)
36         #writer.writerow(row)
37
38 def write_data(data):
39     with open('./input/scraped_data.csv', 'a', newline='') as csv_file:
40         writer = csv.writer(csv_file, quoting=csv.QUOTE_MINIMAL)
41         for row in data:
42             writer.writerow(row)
43
44 def rem_if_not_in_list():
45     df = pd.read_csv('./input/scraped_data.csv')
46     list(set(company_list))
47     # Drop the rows where col2 is not in the accepted_values list
48     df = df[df['name'].isin(company_list)]
49     df.to_csv('./input/scraped_data.csv')
50
51
```

python code for Streamlit home page design and text cleaning

```

1 |_author_ = "Sourav Singh"
2 |_version_ = "1.0.1"
3 |_project_ = "this a home page for Topic Interpretation tool"
4
5 import streamlit as st
6 import pandas as pd
7 import re
8 import nltk
9 from nltk.corpus import stopwords
10 import matplotlib.pyplot as plt
11 from wordcloud import WordCloud
12 #importing wordnet lemmatizer for linking words with similar meanings to one word
13 from nltk.stem.wordnet import WordNetLemmatizer
14 w_tokenizer = nltk.tokenize.WhitespaceTokenizer()
15 import nltk_stopword
16 import string
17 from PIL import Image
18 from remove_stopwords import run_preprocess
19 from insights_vis import insights_visualize
20
21 from rem_non_english import clean_column
22 from inputUrlName import takeUserData
23 from scraper import do_scrap
24 import word_dist
25
26
27 # multiwords as list
28 multiwords_list[]
29 not_cleaned = True
30
31 #lemmatization function
32 def lemmatize_data(csv_data):
33     st.write("Lemmatizing the corpus data...")
34     lemmatizer = nltk.stem.WordNetLemmatizer()
35     lem_word = [lemma.lemmatize(word) for word in lem_word]
36     return lem_word
37
38 #remove punctuations from data
39 #creating a new column in csv as content_processed for storing removed punctuation, lowered case and stopwords free data
40 def punc_remove(csv_data):
41     st.write("removing punctuation for any punctuation...")
42     csv_data['content'] = csv_data['content'].fillna(value='missing_data')
43     punc = '''([!()-{};,:<>@#$%^&*~]'''
44     st.write('removing non english data from content...')
45     csv_data['content'] = clean_column(csv_data['content'])
46     #st.write(csv_data.head())
47     csv_data['content_processed'] = \
48         csv_data['content'].map(lambda x: str(x).translate(str.maketrans('', '', string.punctuation)))
49     #remove any numbers
50     csv_data['content_processed'] = csv_data['content_processed'].str.replace('\d+', '')
51     csv_data['content'].map(lambda x: re.sub('[,.!?]', '', x))
52     csv_data = csv_data.replace(' ', '')
53     csv_data = csv_data.replace('_', '')
54     csv_data = csv_data.replace('-', '')
55     csv_data = csv_data.replace('`', '')
56

```

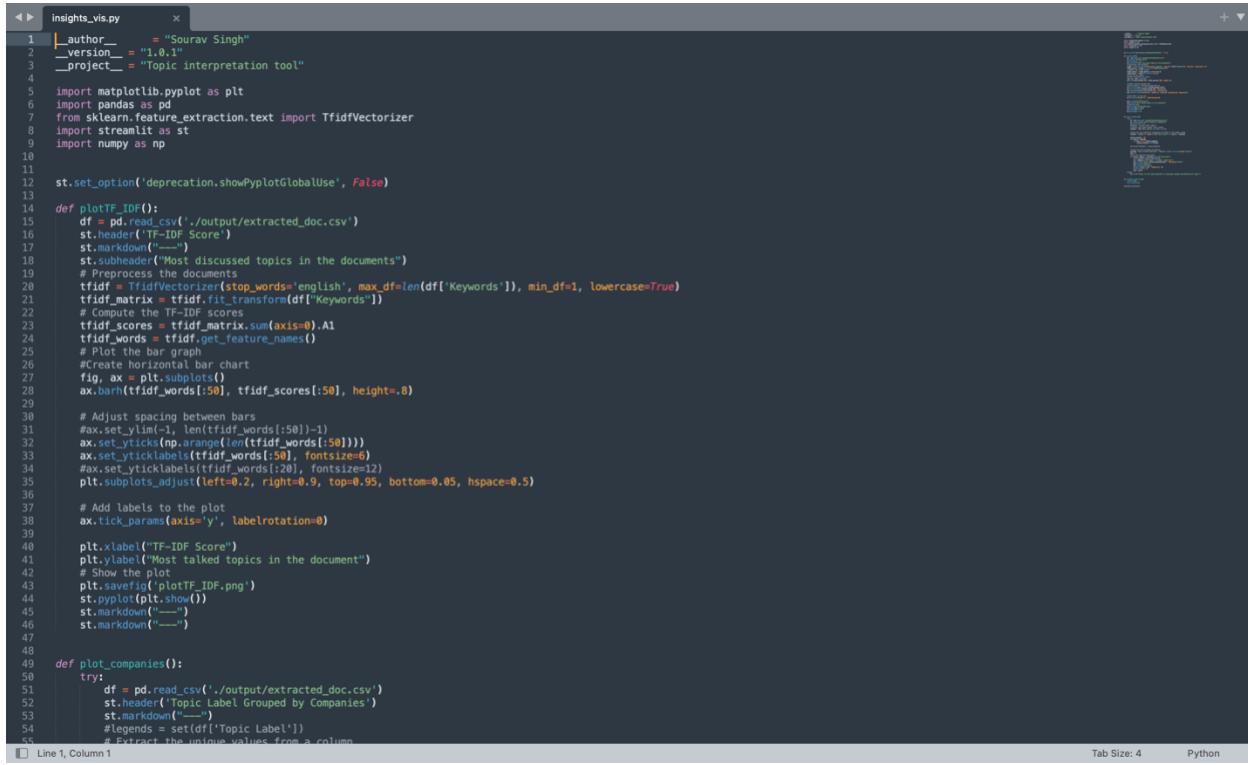
python code for topic model

```

1 |_author_ = "Sourav Singh"
2 |_version_ = "1.0.1"
3
4 import gensim.corpora as corpora
5 from gensim.models.ldamodel import LdaModel
6 from gensim.models import RegexpTokenizer
7 from parquet import parquet
8 import pandas as pd
9 from itertools import chain
10 import numpy as np
11 from nltk.stem.wordnet import WordNetLemmatizer
12 from nltk.tokenize import WhitespaceTokenizer
13 from nltk.tokenize import RegexpTokenizer
14 from transformers import BertCoherenceModel
15 import matplotlib.pyplot as plt
16 import pyDavis
17 import pyDavis.davis
18 import torch
19 from PIL import Image
20 import streamlit.components.v1 as components
21 import sum
22 from summarize import summarize
23 from gensim.summarization import summarize
24 #bert summarize
25 from summarizer import Summarizer
26 import torch
27 from transformers import T5ForConditionalGeneration, T5Tokenizer, pipeline
28 from random import sample
29 from typing import List
30 import os
31 import color_coding as cc
32 from bertLabel import generate_label
33 from gensim import corpora, similarities
34 import numpy as np
35
36 max_articles_for_label = 5
37
38 w_tokenizer = WhitespaceTokenizer()
39 #lemmatization
40 lemm = WordNetLemmatizer()
41
42 def write_data(data):
43     header = ['topic_num', 'keywords', 'summary']
44     with open('./output/summary_data.csv', 'w', newline='') as csv_file:
45         writer = csv.writer(csv_file)
46         writer.writerow(header)
47         for row in data:
48             writer.writerow(row)
49
50 #count number of data in a topic
51 def count_till_value_change(column, start_index):
52     start_index = start_index
53     current_value = start_index
54     while current_value == start_index:
55         start_index += 1
56         current_value = start_index
57
58

```

python code for data visualization

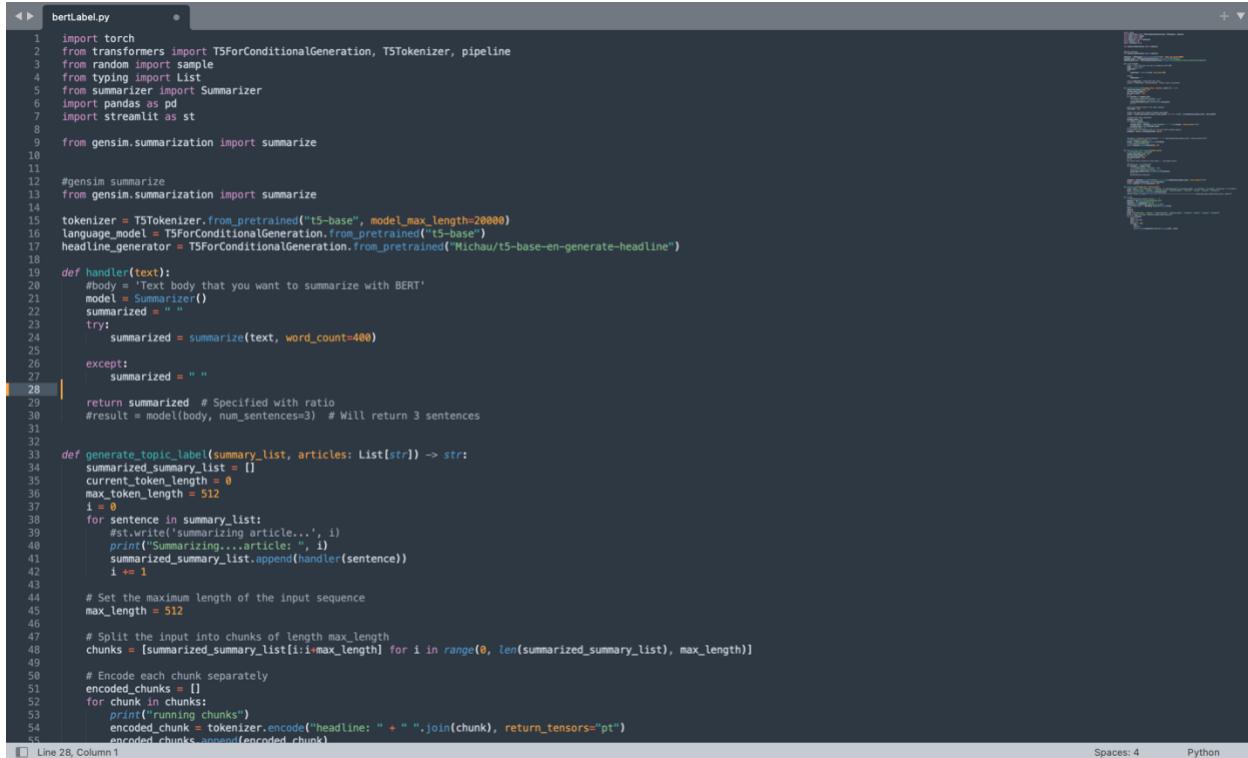


```

1 | author_ = "Sourav Singh"
2 | version_ = "1.0.1"
3 | project_ = "Topic interpretation tool"
4 |
5 | import matplotlib.pyplot as plt
6 | import pandas as pd
7 | from sklearn.feature_extraction.text import TfidfVectorizer
8 | import streamlit as st
9 | import numpy as np
10|
11| st.set_option('deprecation.showPyplotGlobalUse', False)
12|
13| def plotTF_IDF():
14|     df = pd.read_csv('./output/extracted_doc.csv')
15|     st.header("TF-IDF Score")
16|     st.markdown("—")
17|     st.subheader("Most discussed topics in the documents")
18|     # Preprocess the documents
19|     tfidf = TfidfVectorizer(stop_words='english', max_df=len(df['Keywords']), min_df=1, lowercase=True)
20|     tfidf_matrix = tfidf.fit_transform(df['Keywords'])
21|     # Compute the TF-IDF scores
22|     tfidf_scores = tfidf_matrix.sum(axis=0).A1
23|     tfidf_words = tfidf.get_feature_names()
24|     # Plot the bar graph
25|     # Create horizontal bar chart
26|     fig, ax = plt.subplots()
27|     ax.bar(tfidf_words[:50], tfidf_scores[:50], height=.8)
28|
29|     # Adjust spacing between bars
30|     #ax.set_ylim(1, len(tfidf_words)-1)
31|     ax.set_yticks(np.arange(len(tfidf_words[:50])))
32|     ax.set_yticklabels(tfidf_words[:50], fontsize=6)
33|     #ax.set_yticklabels(tfidf.words[:20], fontsize=12)
34|     #ax.set_yticks(np.arange(len(tfidf_words[:20])), fontsize=12)
35|     plt.subplots_adjust(left=0.2, right=0.9, top=0.95, bottom=0.05, hspace=0.5)
36|
37|     # Add labels to the plot
38|     ax.tick_params(axis='y', labelrotation=0)
39|
40|     plt.xlabel("TF-IDF Score")
41|     plt.ylabel("Most talked topics in the document")
42|     # Show the plot
43|     plt.savefig('plotTF_IDF.png')
44|     st.pyplot(plt.show())
45|     st.markdown("—")
46|     st.markdown("—")
47|
48| def plot_companies():
49|     try:
50|         df = pd.read_csv('./output/extracted_doc.csv')
51|         st.header("Topic Label Grouped by Companies")
52|         st.markdown("—")
53|         #legends = set(df['Topic Label'])
54|         # Extract the unique values from a column
55|
56|         # Line 1, Column 1
57|         Tab Size: 4 Python

```

python code for generating topic label



```

1 | import torch
2 | from transformers import T5ForConditionalGeneration, T5Tokenizer, pipeline
3 | from random import sample
4 | from typing import List
5 | from summarizer import Summarizer
6 | import pandas as pd
7 | import streamlit as st
8 |
9 | from gensim.summarization import summarize
10|
11| #gensim summarize
12| from gensim.summarization import summarize
13|
14| tokenizer = T5Tokenizer.from_pretrained("t5-base", model_max_length=20000)
15| language_model = T5ForConditionalGeneration.from_pretrained("t5-base")
16| headline_generator = T5ForConditionalGeneration.from_pretrained("Michau/t5-base-en-generate-headline")
17|
18| def handler(text):
19|     #body = 'Text body that you want to summarize with BERT'
20|     model = Summarizer()
21|     summarized = ""
22|     try:
23|         summarized = summarize(text, word_count=400)
24|     except:
25|         summarized = ""
26|
27|     return summarized # Specified with ratio
28|     #result = model(body, num_sentences=3) # Will return 3 sentences
29|
30|
31| def generate_topic_label(summary_list, articles: List[str]) -> str:
32|     summarized_summary_list = []
33|     current_token_length = 0
34|     max_token_length = 512
35|     i = 0
36|     for sentence in summary_list:
37|         #st.write('summarizing article...', i)
38|         print("Summarizing... article: ", i)
39|         summarized_summary_list.append(handler(sentence))
40|         i += 1
41|
42|     # Set the maximum length of the input sequence
43|     max_length = 512
44|
45|     # Split the input into chunks of length max_length
46|     chunks = [summarized_summary_list[i:i+max_length] for i in range(0, len(summarized_summary_list), max_length)]
47|
48|     # Encode each chunk separately
49|     encoded_chunks = []
50|     for chunk in chunks:
51|         print("running chunks")
52|         encoded_chunk = tokenizer.encode("headline: " + " ".join(chunk), return_tensors="pt")
53|         encoded_chunks.append(encoded_chunk)
54|
55|
56|     # Line 28, Column 1
57|     Spaces: 4 Python

```

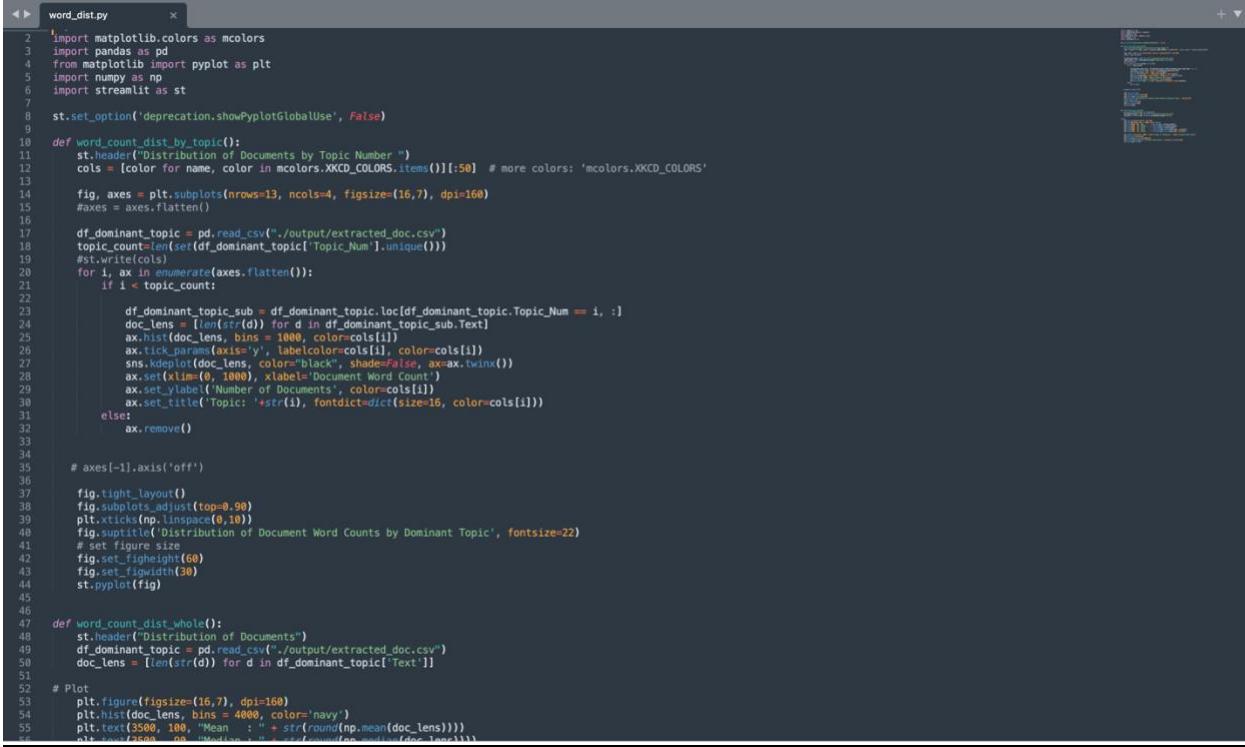
python code for text summarization

```
270 #todo add only text from each company
271 def summary_extractor():
272     #model = Summarizer()
273     model = Summarizer(model='distilbert-base-uncased')
274     print("extracting summary of documents")
275     datafile = pd.read_csv("./output/extracted_doc.csv")
276     filename = list(datafile['Text'])
277     topic_number = datafile['Topic_Num']
278     summary_list=[]
279     text=""
280     count=0
281     slist=[]
282     topic_num = 0
283     start_row = 0
284     end_row = 0
285     #end_row = doc.cut_off_len()
286     while end_row < len(topic_number) and topic_num<int(9):
287         end_row = end_row + count_till_value_changes(topic_number, start_index=start_row)
288         st.write(f"start row before: {start_row}")
289         st.write(f"end row before: {end_row}")
290         for i in filename:
291             count+= 1
292             summary_list.append(i)
293             text += i
294             slist.append(i)
295             if(count == end_row):
296                 print("documents for the topic")
297                 topicList= datafile[Keywords].unique()[topic_num]
298                 st.write(f"Topic {topic_num} : {topicList}")
299                 st.write(f"Topic Keywords: {topicList}")
300                 st.write(f"Number of documents summarized: {len(slist)}")
301                 st.write(f"Summary: {text}")
302                 text=""
303                 count=0
304                 topic_num=topic_num+1
305                 slist=[]
306                 start_row = end_row
307                 end_row = end_row + count_till_value_changes(topic_number, start_index=start_row)
308
309
310     #updated summary methods dynamic
311     def _get_trace_updated():
312         #st.session_state['value'] = '8'
313         print("extracting summary of documents")
314         df = pd.read_csv("./output/extracted_doc.csv")
315         st.write(df.head(5))
316         data = []
317         for i in range(int(st.session_state['value'])):
318             #st.write(f'i is : {i}')
319             df_selected = df[df['Topic_Num'] == i]
320             #df_selected = df_selected.drop_duplicates(subset=['Name']).head(30)
321             df_selected = df_selected.head(50);
322             st.write(f'{len(df_selected)}')
323             #st.write(df_selected.head(5))
```

python code for plotting TF-IDF graph

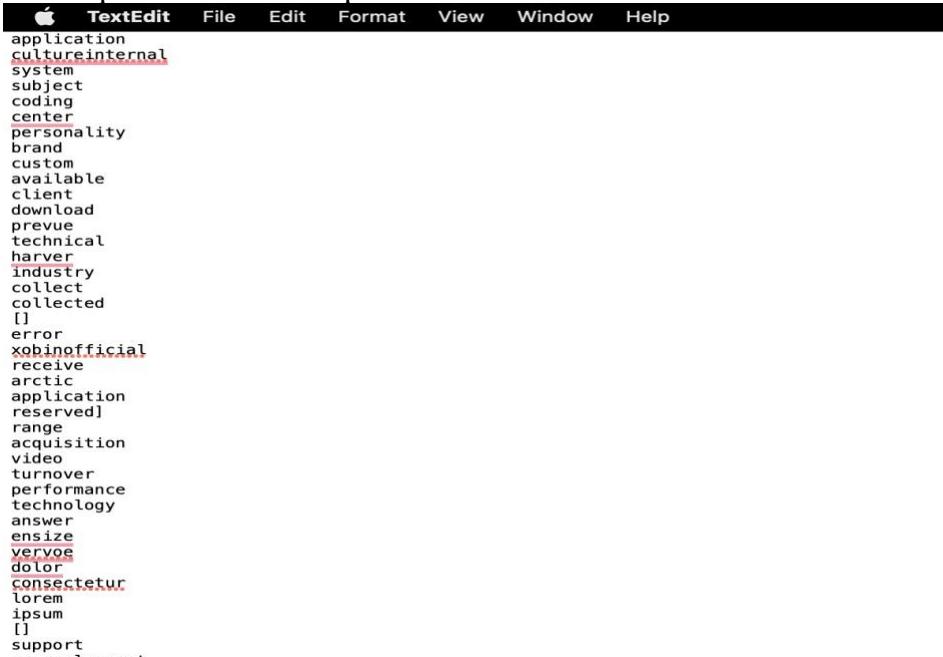
```
1 import matplotlib.pyplot as plt
2 import pandas as pd
3 from sklearn.feature_extraction.text import TfidfVectorizer
4
5 def plotTF_IDF():
6     # Load the CSV file into a DataFrame
7     df = pd.read_csv("extracted_doc.csv")
8
9     # Preprocess the documents
10    tfidf = TfidfVectorizer(stop_words='english', max_df=0.5, min_df=2, lowercase=True)
11    tfidf_matrix = tfidf.fit_transform(df["Keywords"])
12
13    # Compute the TF-IDF scores
14    tfidf_scores = tfidf_matrix.sum(axis=0).A1
15    tfidf_words = tfidf.get_feature_names()
16
17    # Plot the bar graph
18    plt.barh(tfidf_words[:50], tfidf_scores[:50])
19
20    # Add labels to the plot
21    plt.xlabel("TF-IDF Score")
22    plt.ylabel("Most talked topics in the document")
23
24    # Show the plot
25    plt.show()
26
27 plotTF_IDF()
```

python code for plotting word distribution



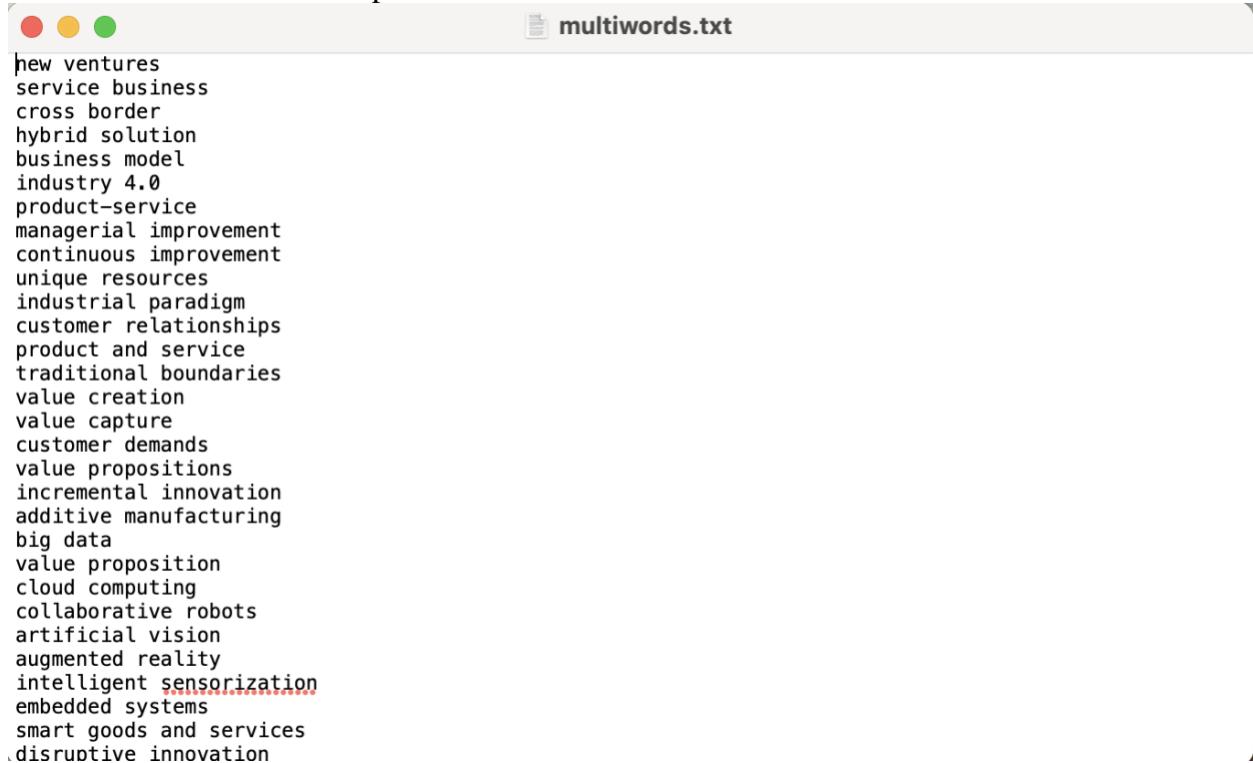
```
1 word_dist.py
2
3 import matplotlib.colors as mcolors
4 import pandas as pd
5 from matplotlib import pyplot as plt
6 import numpy as np
7 import streamlit as st
8
9 st.set_option('deprecation.showPyplotGlobalUse', False)
10
11 def word_count_dist_by_topic():
12     st.header("Distribution of Documents by Topic Number")
13     cols = [color for name, color in mcolors.XKCD_COLORS.items()][:50] # more colors: 'mcolors.XKCD_COLORS'
14
15     fig, axes = plt.subplots(nrows=13, ncols=4, figsize=(16,7), dpi=160)
16     #axes = axes.flatten()
17
18     df_dominant_topic = pd.read_csv("./output/extracted_doc.csv")
19     topic_count = len(df_dominant_topic['Topic_Num'].unique())
20     #st.write(cols)
21     for i, ax in enumerate(axes.flatten()):
22         if i < topic_count:
23             df_dominant_topic_sub = df_dominant_topic.loc[df_dominant_topic.Topic_Num == i, :]
24             doc_lens = [len(str(d)) for d in df_dominant_topic_sub.Text]
25             ax.hist(doc_lens, bins = 1000, color=cols[i])
26             ax.tick_params(axis='y', labelcolor=cols[i], color=cols[i])
27             sns.kdeplot(doc_lens, color="black", shade=False, ax=ax.twinx())
28             ax.set(xlim(0, 1000), xlabel='Document Word Count')
29             ax.set_ylabel('Number of Documents', color=cols[i])
30             ax.set_title(f'Topic: {str(i)}', fontdict=dict(size=16, color=cols[i]))
31         else:
32             ax.remove()
33
34
35 # axes[-1].axis('off')
36
37 fig.tight_layout()
38 fig.subplots_adjust(top=0.90)
39 plt.xticks(np.linspace(0,10))
40 fig.suptitle('Distribution of Document Word Counts by Dominant Topic', fontsize=22)
41 # set figure size
42 fig.set_figheight(60)
43 fig.set_figwidth(30)
44 st.pyplot(fig)
45
46
47 def word_count_dist_whole():
48     st.header("Distribution of Documents")
49     df_dominant_topic = pd.read_csv("./output/extracted_doc.csv")
50     doc_lens = [len(str(d)) for d in df_dominant_topic['Text']]
51
52 # Plot
53 plt.figure(figsize=(16,7), dpi=160)
54 plt.hist(doc_lens, bins = 4000, color='navy')
55 plt.text(3500, 100, "Mean : " + str(round(np.mean(doc_lens))))
56 plt.text(3500, 100, "Min : " + str(min(doc_lens)))
57 plt.text(3500, 100, "Max : " + str(max(doc_lens)))
58
```

stopwords txt file sample



application
cultureinternal
system
subject
coding
center
personality
brand
custom
available
client
download
prevue
technical
harver
industry
collect
collected
[]
error
xobinofficial
receive
arctic
application
reserved]
range
acquisition
video
turnover
performance
technology
answer
ensize
vervoe
dolor
consectetur
lorem
ipsum
[]
support

multi words txt file sample



The screenshot shows a text editor window with a light gray background. At the top left, there are three colored circular icons: red, yellow, and green. To the right of these icons, the file name "multiwords.txt" is displayed in a small, dark font. The main content area contains a list of approximately 30 business-related terms, each on a new line. The terms include: new ventures, service business, cross border, hybrid solution, business model, industry 4.0, product-service, managerial improvement, continuous improvement, unique resources, industrial paradigm, customer relationships, product and service, traditional boundaries, value creation, value capture, customer demands, value propositions, incremental innovation, additive manufacturing, big data, value proposition, cloud computing, collaborative robots, artificial vision, augmented reality, intelligent sensorization, embedded systems, smart goods and services, and disruptive innovation.

```
new ventures
service business
cross border
hybrid solution
business model
industry 4.0
product-service
managerial improvement
continuous improvement
unique resources
industrial paradigm
customer relationships
product and service
traditional boundaries
value creation
value capture
customer demands
value propositions
incremental innovation
additive manufacturing
big data
value proposition
cloud computing
collaborative robots
artificial vision
augmented reality
intelligent sensorization
embedded systems
smart goods and services
disruptive innovation
```