Exploring ways for a better interpretation of Topic Models

By

Sourav Kumar 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

Exploring ways for a better interpretation of topic models

Sourav Singh
Masters in Applied Business Analytics
Carleton University
Ontario, ON K1S 5B7
souravsingh3@cmail.carleton.ca

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 not only identify the main themes or topics in the text data but also to understand the nuances and relationships between them.

This project will explore processes that combines the results of the LDA topic model with 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.

Contents

1 Introduction

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 deliverable,

- A corpus of text documents from 75 talent management companies

- Topic modelling results on the corpus of scraped web pages from Talent Management company websites
- A process to interpret each of the topic models based on the summarization of the associated documents and the application of ChatGPT
- 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 Related Work

1.4.1 What is Known

By leveraging the LDA topic model, word vectors, letter trigram vectors, seq2seq algorithms, and multiple text summarization algorithms such as SumBasic and MapReduce, we can create a comprehensive approach to topic modelling interpretation. The combination of the above techniques will enable us to identify and label topics automatically while extracting the most important information from relevant text documents. Doing so can provide a more accurate and meaningful interpretation of the topics identified by the model.

1.4.2 What is Unknown

A systematic approach to interpreting topics based on text documents extracted from company websites. Need for the process that could combine the above said algorithms along with topic modelling results for a better interpretation of the topics.

1.5 Contribution

The major contribution of this approach is the integration of multiple techniques and algorithms to improve the effectiveness and accuracy of topic modelling and interpretation. This approach has the potential to significantly improve the efficiency of topic modelling and text analytics, leading to better decision-making and insights for businesses and organizations across various industries. Additionally, the use of this integrated approach can help to promote further research and development in the field of text analytics and natural language processing.

1.6 Method

- Review the following literature streams:
  - Talent Management
  - Topic modelling approach
  - Topic labelling and interpretation
  - Text summarization techniques
  - ChatGPT
- Build a list of companies in Talent Management domain
- Web scrap the data of 75 Talent management companies
- Implement and apply topic model on the corpus of docs
- Identify companies that are most representative of each given topic
- 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 ChatGPT analysis
- Finally documenting the process and finalizing the result

1.7 Organization

This project consists of six chapters. The first chapter, "Introduction," provides a description of the project's goal, deliverables, relevance, what is known and unknown, contribution, and a summary of the method. In chapter 2, "Literature Review," the literature streams related to Talent Management, Topic modelling approach, Topic labelling and interpretation, Text summarization technique, and ChatGPT are discussed. Chapter 3, "Method," lists the steps taken to complete the project. The results are categorized into deliverables in Chapter 4, "Results." Chapter 5, "Discussion of Results," covers the problem addressed, new information discovered through research, and the relationship between the literature and the project's findings. Finally, in Chapter 6, "Conclusions and Recommendations," the report concludes with a summary of the findings and recommendations for further study.

2 Literature Review

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 (Boštjančič & Slana 2018).

Topic modeling 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. LDA 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.

Automated topic labelling 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 labellers 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. Kou et al. (2016) propose "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.

Twinandilla et al. (2018) suggested a multi-document text summarization method based on both K-means clustering and Latent Dirichlet Allocation (LDA) to identify the most significant sentences in a set of documents. The approach involves clustering similar sentences using K-means, 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 (Recall-Oriented Understudy for Gisting Evaluation) metric and compared to other multi-document summarization approaches, demonstrating its effectiveness in producing high-quality summaries. Roul et al. (2018) 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.

Additionally, ChatGPT can assist in topic modeling by generating relevant keywords and topic suggestions based on a given dataset or text corpus. ChatGPT can also help in identifying patterns and relationships within the data that may be indicative of underlying topics or themes. Additionally, ChatGPT can be used to evaluate and refine the results of topic modeling algorithms by providing a human perspective on the relevance and accuracy of generated topics and their corresponding labels. ChatGPT can also suggest potential areas of improvement for the topic modeling process, such as refining data preprocessing methods or adjusting algorithm parameters. Anisin (2018) discusses the use of GPT-3 (Generative Pre-trained Transformer 3) for text summarization and explains the process of how GPT-3 can be fine-tuned for summarization tasks using a small dataset of examples. It also covers the benefits of using GPT-3 for text summarization, including its ability to generate human-like summaries that capture the essential information in a given text. ChatGPT can have a potential impact on various aspects of academia and libraries, such as search and discovery, reference and information services, cataloging and metadata generation, and content creation (Lund, B & Wang, T. 2023).

2.1 Purpose

The purpose of the literature review is obtain the following information:

- Need for Talent Management tools and companies
- Understanding the existing topic modelling alogorithms and their limitations
- Understaing the existing alogorithms for tex summarization and how they can be integrated with topic model to extract meaning data driven insights
- Understanding the impact and use of ChatGPT on anlyzing the large set of data and how it can be helpful in modeling, labelling and summarizing the corpus to produce insights

2.2 Framework

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

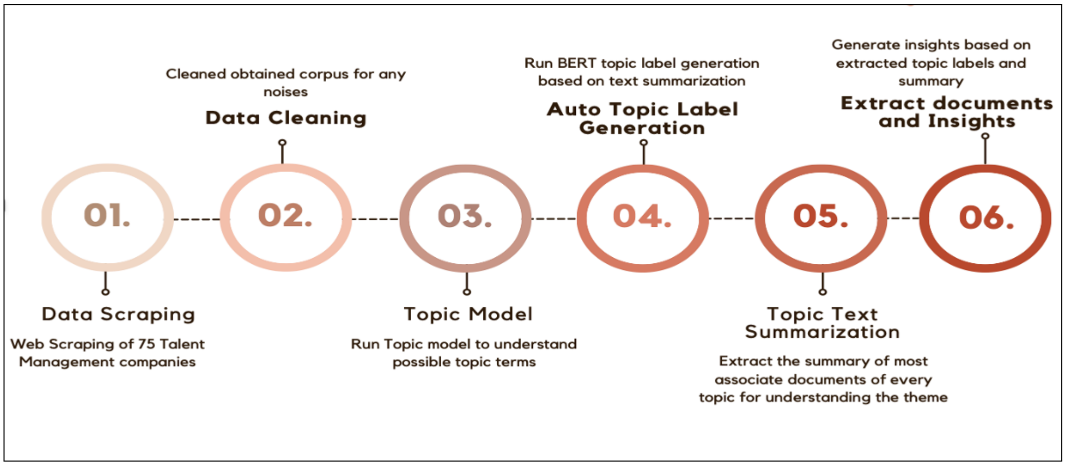

Cleaned obtained corpus for any noises
Data Cleaning
Run BERT topic label generation based on text summarization
Auto Topic Label Generation
Generate insights based on extracted topic labels and summary
Extract documents and Insights
01.
02.
03.
04.
05.
06.
Data Scraping
Web Scraping of 75 Talent Management companies
Topic Model
Run Topic model to understand possible topic terms
Topic Text Summarization
Extract the summary of most associate documents of every topic for understanding the theme

Figure 1: Topic interpretation framework

The framework states the overall process of topic interpretation starts with data scraping using an in-house web scrapper, then the scraped is sent for data cleaning where it will be cleaned for any noises including punctuation, invalid characters, text lemmatizing, stopwords and processing the multi-words. This cleaned data will then be used as a corpus for our topic model, where an optimum number of topics will be identified and the model produce the topic keywords with the stated number of topics.

The project utilizes four 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

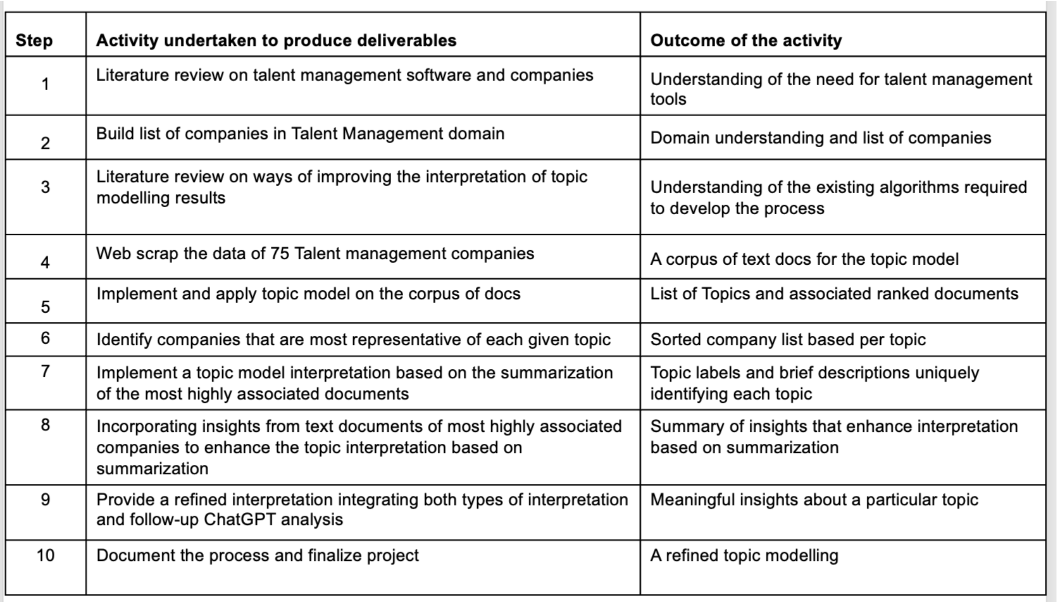
- Approach 2: Choosing the first 50 most associated documents in each topic

- Approach 3: Keeping entire associated documents within the topic after putting the threshold

- Approach 4: using Cosine Similarity to pick the associated documents within the topic

The identified associated documents will further be analyzed by ChatGPT 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.

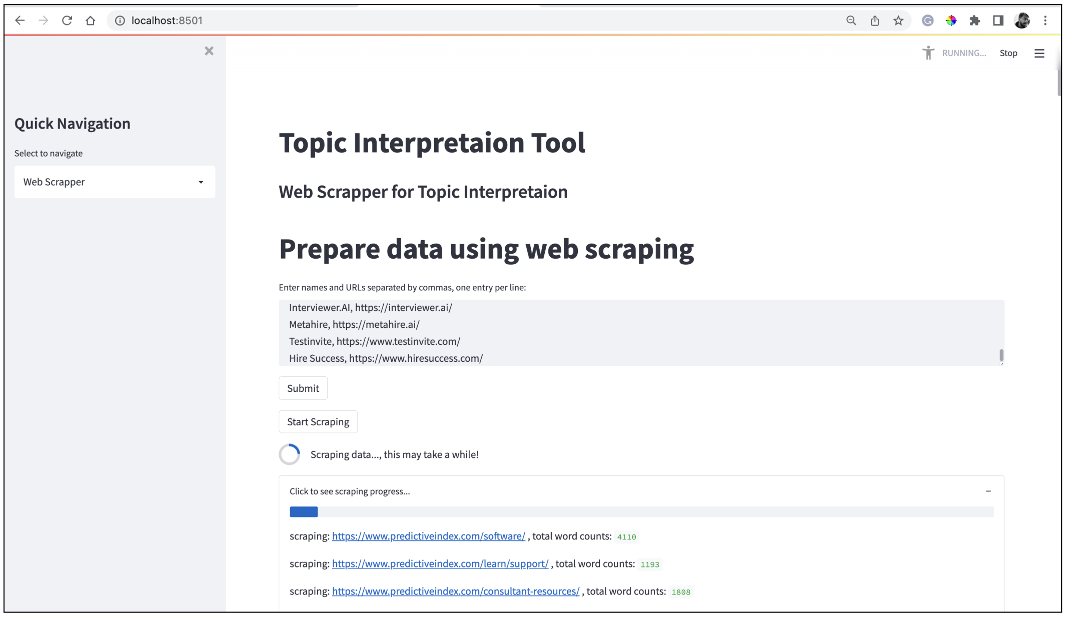
3 Research design and method



Step	Activity 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 scrap 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 ChatGPT analysis	Meaningful insights about a particular topic
10	Document the process and finalize project	A refined topic modelling

3.1 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 scraped. The scraped dataset has the company's name, URL and webpage contents as corpus documents for the topic model and further text summarization.



Topic Interpretaion Tool

Web Scrapper for Topic Interpretaion

Prepare data using web scraping

Enter names and URLs separated by commas, one entry per line:

Interviewer.AI, https://interviewer.ai/
Metahire, https://metahire.ai/
Testinvite, https://www.testinvite.com/
Hire Success, https://www.hiresuccess.com/

Submit

Start Scraping

Scraping data..., this may take a while!

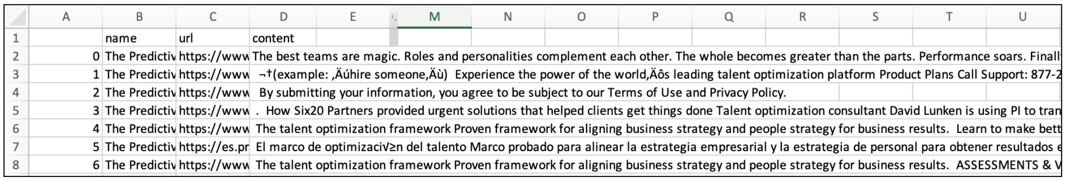
Click to see scraping progress...

scraping: https://www.predictiveindex.com/software/ , total word counts: 4110

scraping: https://www.predictiveindex.com/learn/support/ , total word counts: 1193

scraping: https://www.predictiveindex.com/consultant-resources/ , total word counts: 1808

Figure 2: Data scraping tool



	A	B	C	D
1		name	url	content
2	0	The Predictiv	https://www	The best teams are magic. Roles and personalities complement each other. The whole becomes greater than the parts. Performance soars. Finall
3	1	The Predictiv	https://www	¬†(example: ‚Äúhire someone‚Äù) Experience the power of the world‚Äôs leading talent optimization platform Product Plans Call Support: 877-2
4	2	The Predictiv	https://www	By submitting your information, you agree to be subject to our Terms of Use and Privacy Policy.
5	3	The Predictiv	https://www	. How Six20 Partners provided urgent solutions that helped clients get things done Talent optimization consultant David Lunken is using PI to tran
6	4	The Predictiv	https://www	The talent optimization framework Proven framework for aligning business strategy and people strategy for business results. Learn to make bett
7	5	The Predictiv	https://es.pr	El marco de optimizaci√≥n del talento Marco probado para alinear la estrategia empresarial y la estrategia de personal para obtener resultados e
8	6	The Predictiv	https://www	The talent optimization framework Proven framework for aligning business strategy and people strategy for business results. ASSESSMENTS & V

Figure 3: Sample of scraped data

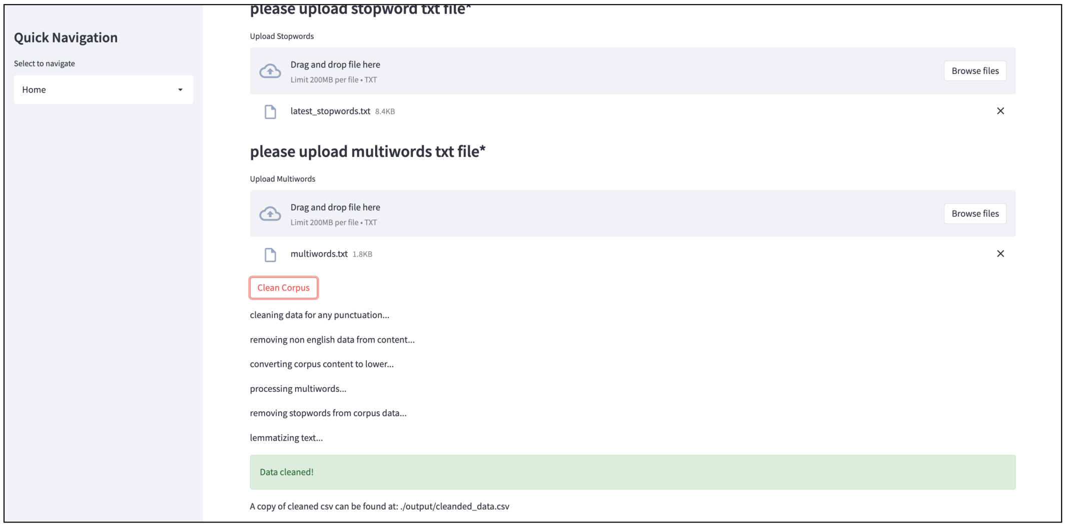
3.2 Data analysis

The corpus is checked and cleaned for any noise such as special characters, stop words or multi-words using python libraries and any irrelevant data such as emojis, punctuation, and non-English words are also cleaned. 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 tool requires to:

- Upload the scraped corpus
- Upload the stopwords txt file
- Upload the multiwords txt file
- And click on Clean Corpus button to get the cleaned data
- The corpus is cleaned for:

- Punctuation
- Non-English contents
- All contents to lower case
- Removed stop-words and processed multi-words
- Performed Text lemmatization



cleaning data for any punctuation...

removing non english data from content...

converting corpus content to lower...

processing multiwords...

removing stopwords from corpus data...

lemmatizing text...

Data cleaned!

A copy of cleaned csv can be found at: ./output/cleanded_data.csv

Figure 4: Data cleaning process

3.3 Topic Modeling

3.3.1 Identifying the optimal number of topics

The LDA topic coherence score was utilized to determine the best number of topics for the data. An optimal number of topics should have a coherence score greater than 0.5, which is considered good for most cases according to Rosner et al. (2014). The coherence score was highest when the number of topics was around 10.

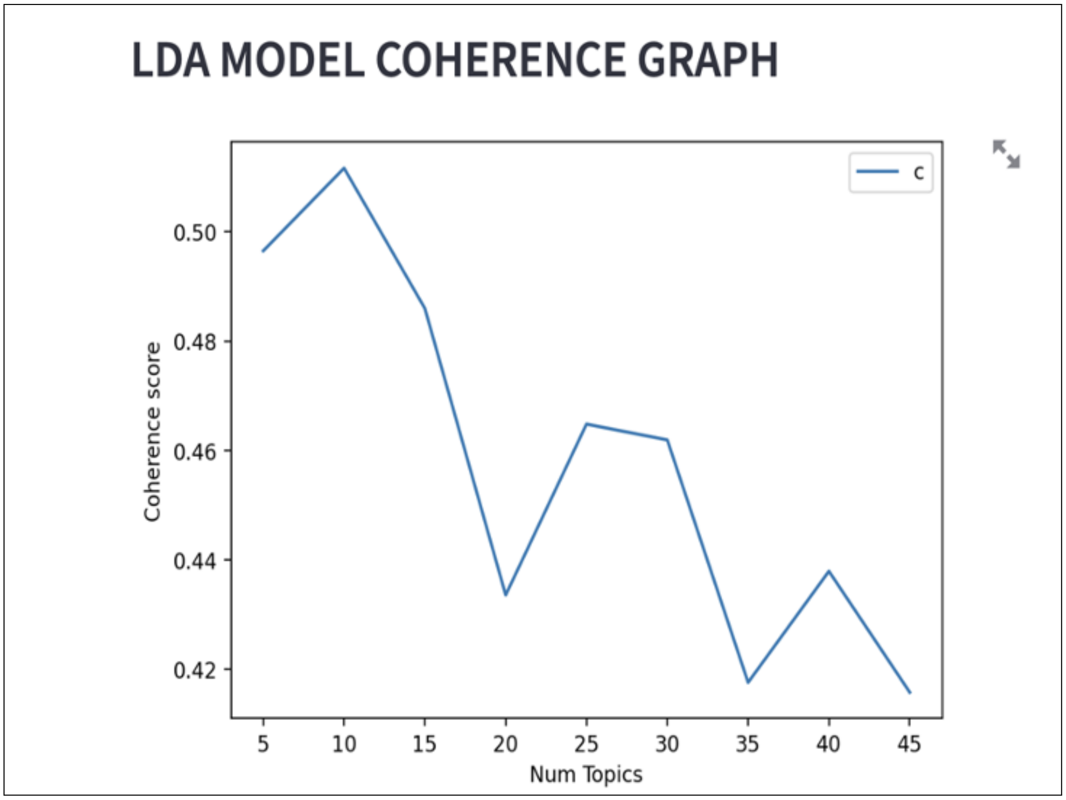
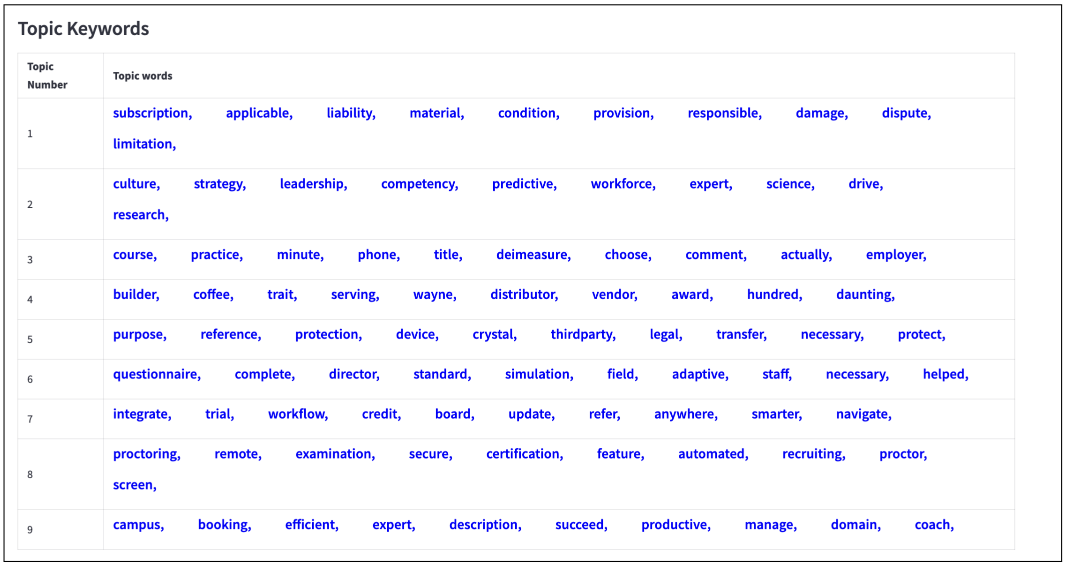


Figure 5: model coherence graph

3.3.2 Topic model

The topic model was run with different numbers of topics as per around obtained optimal number of topics from the model coherence graph (Figure 5) until the keywords for each topic were distinguishable. After experimentation, it was found that the optimal number of topics for the data was 9. Topic models were generated using this number of topics, and the resulting models had distinguishable keywords for each topic.

The obtained topic keywords are as below:



Topic Keywords

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 6: Topic model result

3.4 Data visualization

Figure 7 shows a word cloud of the entire corpus after the data was cleaned. Figure 8 displays an intertopic distance map visualization that depicts the relationships between topics in the corpus. This visualization helps to explore how the clusters of related topics are connected.

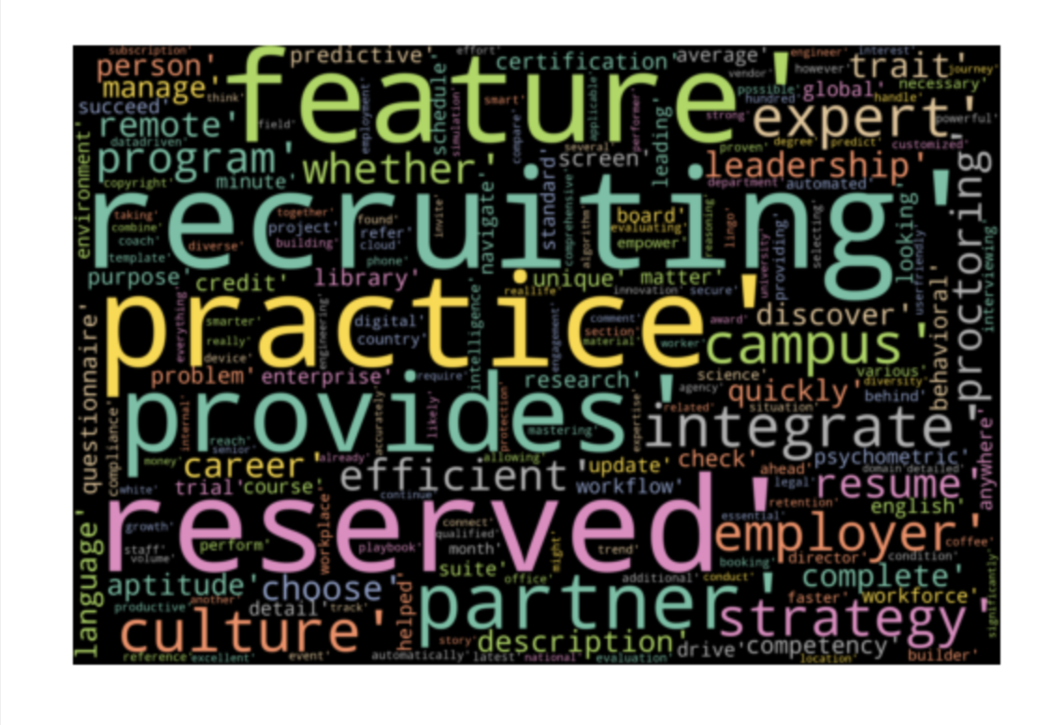
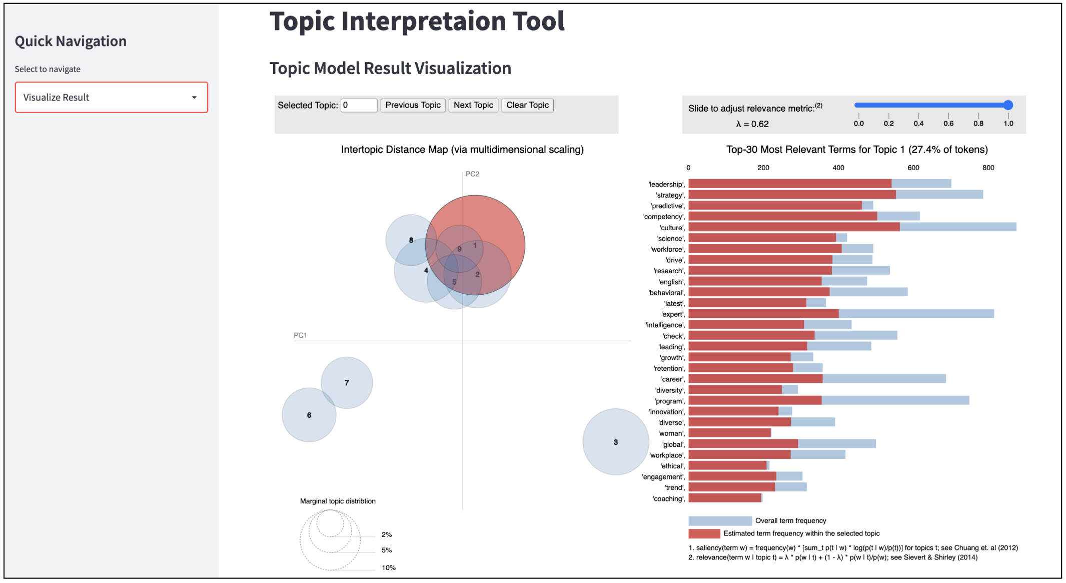


Figure 7: Wordcloud



Quick Navigation

Select to navigate

Visualize Result ▾

Topic Interpretaion Tool

Topic Model Result Visualization

Selected Topic: 0 | Previous Topic | Next Topic | Clear Topic

Slide to adjust relevance metric:(2)

λ = 0.62

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 8: intertopic visualization

References

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

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

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
https://doi.org/10.1016/j.hrmr.2022.100926

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–.

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

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

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.

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.

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

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