Document Tagging System for Information retrieval and relevance

Shah Vansh, Dhwani Gandhi, Swara Desai, Aditya Bhat

Abstract

Document tagging is a key aspect of Information Retrieval (IR) systems, enabling efficient indexing and retrieval by assigning meaningful labels or keywords to documents. Our project explores tagging methodologies using the NIPS Papers dataset, which focuses on machine learning and artificial intelligence. Leveraging advanced Natural Language Processing (NLP) techniques such as RAKE and BERT embeddings, the method addresses challenges in multi-label classification for document abstracts.Comparative evaluation of models highlights the superiority of contextual embeddings and enhances tagging precision while addressing challenges of computational costs and scalability.

Keywords

Auto-tagging, cosine similarity, automatic keyword extraction, domain classification, graph-based method

1. Dataset

The NIPS Papers dataset was chosen for its domain-specific nature and high-quality content, which aligns well with the requirements of keyword extraction and topic modeling tasks.The dataset has 9700 rows with columns being paperid, author, title, abstract and full text. This data set comprises machine learning research papers with expert-curated abstracts that offer concise but informative text. The preprocessing steps included cleaning the dataset by removing rows with missing and irrelevant abstracts and normalizing the text by tokenizing, eliminating stop words and stemming.We manually annotated the keywords, extracted different sections such as abstract keywords. The extraction was done using spacy layout tool These steps ensured that the dataset was both clean and standardized, providing a strong foundation for training and evaluating tagging models. The specificity and high quality of the NIPS Papers dataset make it ideal for tasks that require precise and domain-relevant tagging.

2. Evaluation Metrics

To assess the performance of the tagging models, standard evaluation metrics such as precision, recall, and F1 score were used. Precision measures the proportion of relevant tags among those generated by the model, while recall evaluates the model's ability to identify all relevant tags. The F1-score provides a harmonic mean of precision and recall, balancing these two key aspects.

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}} \tag{1}$$

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}} \tag{2}$$

$$F_1 \text{ Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

Introduction to Information Retrieval (IT550), Autumn 2024, DAIICT, India

3. Results

The evaluation revealed huge difference in the performance of the models. RAKE, a simple keyword extraction algorithm relying on word frequency and co occurrence patterns, provided a straightforward approach but struggled with contextual accuracy, leading to lower precision and recall. BERT, while leveraging pre trained contextual embeddings to capture semantic nuances, exhibited high accuracy in generating relevant tags. However, its computational demands and slower processing times can lead to challenges for large-scale applications. While, YAKE struck a balance between efficiency and contextual understanding, outperforming BERT in practical scenarios by delivering precise and relevant keywords without the computational overhead. Examples of keyword extraction illustrated these differences, with YAKE effectively producing concise terms such as 'machine learning,' while BERT occasionally overgeneralized, and RAKE tended to generate overly extended phrases. YAKE's superior scalability and precision makes it an balanced choice for keyword extraction in large datasets.

Table 1
Precision, Recall, and F1 Scores for RAKE, YAKE, and BERT

Method	Precision	Recall	F1 Score
RAKE	0.50	0.33	0.40
YAKE	0.82	0.68	0.74
BERT	0.80	0.66	0.72

4. Key Challenges and Learnings

We faced many difficult challenges in document tagging. Capturing nuanced semantics in technical and interdisciplinary texts proved difficult for traditional models. The computational cost of advanced models like BERT poses scalability challenges, which limits their application in larger datasets. Furthermore, generalizing models trained on domain-specific datasets to other contexts remains a significant hurdle. Key outcomes of the project include the importance of selecting models based on task requirements. For instance, RAKE is suitable for straightforward tasks, YAKE offers a balance of scalability and accuracy, and BERT excels in precision-critical applications. Additionally, thorough preprocessing is critical for meaningful keyword extraction, while domain-specific datasets like NIPS significantly improves tagging outcomes.

5. Conclusion

Our project shows how important document tagging is for improving search systems and compares different models to solve current challenges. RAKE and YAKE stand out for being simple and scalable, but YAKE does a better job at balancing precision and speed, making it great for most tasks. BERT is much better at understanding context and giving highly accurate tags, especially for highly technical documents, but it's computationally heavy and difficult to scale. In future, we should focus on making models faster, improving tagging for different kinds of data, and finding a good balance between accuracy and efficiency.

References

[1] Thushara, M. G., Krishnapriya, M. S., & Nair, S. S. (2017). A model for auto-tagging of research papers based on keyphrase extraction methods. In Group Project (pp. 1695–1700). https://-doi.org/10.1109/icacci.2017.8126087

[2] Rios, P., & Hogan, A. (2018). PubTag: Generating Research Tag-Clouds with Keyphrase Extraction and Learning-to-Rank. IEEE/WIC/ACM International Conference on Web Intelligence (WI'04), 680–683. https://doi.org/10.1109/wi.2018.00-12

[3] Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A Pretrained Language model for Scientific text. https://doi.org/10.18653/v1/d19-1371

[4] Cai, L., Song, Y., Liu, T., & Zhang, K. (2020). A Hybrid BERT Model That Incorporates Label Semantics via Adjustive Attention for Multi-Label Text Classification. IEEE Access, 8, 152183–152192. https://doi.org/10.1109/access.2020.3017382