Topic Modeling and Labeling of NIPS Papers

Monirul Islam & Eric Korshun

Introduction

In a world generating vast amounts of unstructured data by the second, it is necessary to differentiate this information for human comprehension. The approach of topic modeling provides a means to uncover hidden patterns portrayed in a collection of documents. It is invaluable as it enables document classification, content recommendation, and trend identification – applications that can take long time periods for a human to complete.

This project aims to perform the natural language processing task of topic modeling in regards to the NIPS papers dataset. The dataset comprises papers from the Neural Information Processing Systems conference, a leading conference covering innovative topics in machine learning and artificial intelligence. This dataset consists of 1500 papers, with six columns highlighting information about the titles, the year of publication, abstracts that offer a brief overview, the full content of the documents, and other details.

The project is designed into two parts for analysis. The first section intends to treat the entire NIPS papers dataset as one corpus and identify the top *n* topics with their corresponding word distributions through various topic modeling algorithms. In the second section, the dataset is split into three time periods and analyzes the evolution of machine learning topics based on the topics and word distributions generated from the algorithms.

Part I

Data Preparation

Before training an algorithm, the NIPS papers dataset required preprocessing. A new data frame was created by combining the "title" and "paper_text" columns from the original dataset. A function was then applied to preprocess the text. It carried out several steps, including converting the text to lowercase, replacing newline characters and multiple spaces with a single space, removing digits, punctuation, and stopwords, tokenizing the text, discarding tokens with fewer than two characters, and lemmatizing the text using the spaCy library. The spaCy package offers a Part-of-Speech (POS) tagger and a morphological analyzer to make the lemmatization process more efficient and accurate.

NLP Algorithms, Model Evaluation, Results

The Latent Dirichlet Allocation (LDA) model — one of the most popular and foundational topic modeling algorithms — was applied to identify the topics and word distributions within the corpus. However, before an LDA model can be trained, it is necessary to establish a dictionary using the Gensim library. The dictionary maps each word in the processed text to a unique ID. The original dictionary contained 98,496 words but was condensed to 14,214 words by filtering out rare words appearing in fewer than five documents and generic terms that appeared in more than forty percent of documents. Since the LDA model expects a numerical input, the condensed dictionary was converted into a Bag-of-Words (BoW) corpus, representing the raw counts of

word occurrences, and a Term Frequency-Inverse Document Frequency (TF-IDF) corpus, which assigns values to each word given its importance across the corpus. Although standard LDA operates on the BoW representation, TF-IDF was also incorporated to compare evaluation metrics.

Separate LDA models were trained using the BoW and TF-IDF corpora, each configured to produce ten topics with ten words per topic using Gensim's LdaMulticore. Coherence metrics, c_v and u_mass, were computed to evaluate model quality. The range for c_v is from 0.0 to 1.0, and for u_mass it is -14 to 14, with higher scores reflecting better model performance. The model trained on the BoW corpus achieved a c_v score of 0.5660 and a u_mass score of -1.4513, whereas the model trained on the TF-IDF corpus yielded a c_v score of 0.4624 and a u_mass score of -10.7953. These evaluation scores reveal that the BoW-driven LDA model achieved higher topic coherence than its TF-IDF counterpart.

It is important to note that various hyperparameter combinations were tested, both during dictionary creation and in the LDA model settings. Different numerical values were explored in filtering rare and generic words for the dictionary, as well as for adjusting the number of topics and chunk size in the LDA settings. Ultimately, the combination of ten topics with a chunk size of ten produced the highest evaluation scores across both LDA models.

An intertopic distance map was created using pyLDAvis to visualize the topic relationships from the BoW-based LDA model. According to the visualization, topics 1, 2, 5, 6, and 8 are clustered together, suggesting overlapping themes. Topics 3 and 7 also exhibit overlap, indicating thematic similarity. In contrast, topics 4, 9, and 10 are isolated from one another and from the other topics, likely covering distinct content areas. Terms including "graph," "kernel," "cluster," "layer," "object," "label," "neuron," and "inference" are frequently used across the corpus and are highly salient. Topic 1 has the largest circle, accounting for 16.4% of all tokens with key terms like "bayesian," "likelihood," "posterior," "inference," "variational," and "sampling." Since one of the limitations of the LDA algorithm is that it cannot generate a descriptive topic label, topic 1 seems to cover **Bayesian inference and probabilistic modeling** based on human judgment. Topic 2, covering 15.3% of the tokens, appears to relate to **optimization techniques**, with top terms such as "convex," "sparse," "norm," "rank," "penalty," and "pca." Topic 3 comprises 12.9% of the tokens and includes terms like "object," "human," "detection," "face," "visual," "cnn," and "pixel," indicating a focus on **visual processing and object detection**. Topic 10 has the smallest share of tokens with 3.8%, but is distinct from other topics. It is characterized by terms like "word," "language," "text," "embed," "hash," and "lsh," – a theme focused on **natural language processing and language modeling**.

Since the LDA algorithm does not inherently produce topic labels, the keyword extraction technique, KeyBERT, was utilized to generate them automatically for improved interpretability.

The word distribution output from each LDA topic was treated as a "mini-document," and KeyBERT was used to summarize the words into a single key phrase. A for-loop was implemented to extract the words from the LDA output using regex, which were then concatenated into a single string and passed to KeyBERT. KeyBERT returned a summary key phrase and a similarity score, expressed as a percentage, indicating how representative the label is of the topic's overall word set. The model yielded high similarity scores, with two topics exceeding 70%, suggesting strong alignment between the key phrase and the topic words. Most other labels achieved scores above 50%, reflecting moderate validity. Despite the decent similarity scores, the KeyBERT-generated labels did not appear sufficiently accurate from a human perspective to differentiate the topics. This limitation stems from each "mini-document" consisting of only ten topic words, providing limited context. Nevertheless, the KeyBERT process was incorporated to enhance the overall analysis.

A final and more recent topic modeling technique, BERTopic, was included to compare the top *n* topics with those produced by the BoW-based LDA model. BERTopic discovers topics from a large corpus through transformer-driven embeddings, dimensionality reduction, and clustering. The model was trained on the processed text and reduced to the top five topics. Upon comparison between both models, the topic words in the LDA model appear more technical. For instance, LDA topics include terms such as "convolutional," "bayesian," and "graph," whereas the BERTopic model yields more general terms like "algorithm," "model," and "neuron." Additionally, some overlap and redundancy were observed in BERTopic's topic keywords, suggesting the need for hyperparameter tuning.

Part II

For the second part of the project, the focus was shifted to finding the progression of topics in the NIPS papers over time. Many of the processes used are similar to those seen in Part I, each being applied to specific sections of the dataset.

Data Preparation

Since the main component of this section involves binning the dataset into predetermined year ranges, we began data preparation by determining the overall year span of all documents in the corpus. We discovered the data spans over a 30-year time interval (31 years inclusive), from 1987 to 2017. Naturally, we chose to bin the data by 10-year intervals and examine the document distribution, starting by allocating the extra year to the 2007-2017 range:

Year Range	Count
2007–2017	843
1997–2006	396
1987–1996	261

Since the results yielded a skewed distribution of documents across the bins, we opted to take the extra year from the 2007-2017 bin and compile it to the 1987-1996 bin, resulting in the following new bins and distributions:

Year Range	Count
2008–2017	799
1998–2007	400
1987–1997	301

Despite still not having a uniform distribution of documents, the counts across bins have improved, and the substantial number of tokens within the documents should yield sufficient results.

Each of the newly created bins was placed into its respective data frames, where the “title” and “paper_text” columns were combined, and the results were preprocessed using the same cleanup function as in Part I.

Training, Model Evaluation, Results

LDA models were trained separately for each bin and its respective data frames. We began by creating new dictionaries for each data frame, resulting in the following initial word counts:

- 1987-1997: 29,392 words
- 1998-2007: 36,216 words
- 2008-2017: 64,645 words

When condensing the dictionary sizes, we decided to increase the “no_above” parameter in the dictionary filter from 0.4 to 0.6 to retain more information from the documents, producing the following post-filter word counts:

- 1987-1997: 2,647 words
- 1998-2007: 3,400 words
- 2008-2017: 6,445 words

These dictionaries were then converted into Bag-of-Words and TF-IDF corpora for evaluation after model training. Average coherence scores were calculated to get an overall understanding of how each representation technique performed. BoW generated an average c_v score of 0.49 across the three DataFrames compared to TF-IDF’s score of 0.50, indicating acceptable results on both fronts. The u_mass score, however, was significantly better for BoW than TF-IDF, with respective average scores of -1.07 and -9.80. Given better topic coherence with BoW, we opted to continue with this representation.

With the model trained and the word-topic distributions prepared, visualizations were created using pyLDAvis to illustrate the relationships between topics. Below is an overview of our findings:

1987-1997:

Topics 1, 2, 3, 4, 5, 7, and 8 are clustered together, indicating that the majority of topics in this document set share similarities, with topics 1, 2, and 3 accounting for 51.6% of total tokens.

- Topic 1 featured terms like “distribution,” “probability,” “estimate,” “gaussian,” and “noise,” pointing to work in **probabilistic modeling**.
- Topic 2 included words such as “cell,” “neuron,” “signal,” “stimulus,” and “response,” reflecting themes in **computational neuroscience**.
- Topic 3 was centered around terms like “hide,” “net,” “layer,” “node,” and “backpropagation,” hinting at early developments in **neural networks**.

Since these three topics represent the majority of this dataset, we can conclude based on our labels that the years 1987-1997 were likely focused on **neural networks.** By examining the top salient terms for this timeframe, we can find more insights into the research involving neural networks. Some of the prominent words being “image”, ”recognition”, ”speech", and “face” suggest an emphasis on **multi-modal pattern recognition.**

1998-2007:

Topics 1, 2, and 3 form a cluster, with topics 5 and 8 displaying overlap. Topics 1, 2, 3, and 5 had the highest relevance during this time range, accounting for 54.7% of total tokens.

- Topic 1 included terms like “tree,” “bayesian,” “inference,” “markov,” and “probabilistic,” focusing on **probabilistic models.**
- Topic 2 featured “loss,” “approximation,” “gaussian,” “variance,” and “proof,” indicating research in **statistical learning.**
- Topic 3 centered on terms like “classification,” “training,” “regression,” “svm,” and “label,” representing **supervised learning methods.**
- Topic 5 highlighted words such as “noise,” “component,” “covariance,” “scale,” and “density,” showing attention to **model validation**.

Based on these labels, NIPS papers from 1998–2007 appear to focus on **predictive modeling,** with the salient terms including “neuron”, “state”, “action”, and “policy” highlighting **representation learning**.

2008-2017:

Topics 1, 2, 3, 4, 5, and 7 have the closest distance, with topics 1, 2, 3, and 4 comprising 57.4% of total tokens.

- Topic 1 produced terms like “sparse,” “estimator,” “regression,” “component,” and “dimension,” reflecting work in **high-dimensional inference.**

- Topic 2 featured “layer,” “deep,” “object,” “image,” and “cnn,” pointing to the rise of deep learning for **visual representation.**
- Topic 3 brought in terms like “bound,” “loss,” “theorem,” “gradient,” and “convergence,” suggesting progress in **optimization methods**.
- Topic 4 included “inference,” “posterior,” “bayesian,” “variational,” “latent,” and “gradient,” emphasizing ongoing developments in **probabilistic machine learning.**

These trends imply that research from 2008 to 2017 focused heavily on building the **foundations of modern machine learning**. The salient terms remained consistent with those from previous decades, indicating a continued refinement of earlier topics.

KeyBERT

KeyBERT was utilized once again in attempts to label the topic output from lda.print_topics(). The primary goal remained to use LDA for the core project while exploring if we can potentially automate more human-type labels for our topics. Since the outputs of KeyBERT were not to our liking in Part I, we fine-tuned the following hyperparameters to generate a better output: keyphrase_ngram_range, use_mmr, and diversity. Keyphrase_ngram_range controls the number of tokens extracted from the input text. We raised this value to (1,3) from the previous (1,2) to increase the accuracy and interpretability of the labels. Use_mmr (maximal marginal relevance) was set to True since the input text consisted of topic keywords and not actual sentences, allowing KeyBert to focus on co-occurrence meaning. Diversity is necessary when use_mmr is enabled, with values ranging from 0-1, where lower values target relevance and higher values promote diversity. And since we wanted to produce a non-redundant but semantically similar output, we opted for a value of 0.7 for this hyperparameter.

This model yielded better results than in Part I, with outputs being more semantically sound. For instance, KeyBERT generated “gaussian noise mixture” for topic 1 in the years 1987-1996, while our manual label was “probabilistic modeling.” Although this approach is easier to digest than an entire output of tokens, it is important to note that KeyBERT does not consider the token weights generated by LDA. KeyBERT can return more interpretable results, but it can potentially lose the scope of the underlying topic.

Conclusion

Overall, we are satisfied with the findings and believe that LDA successfully identified and grouped topics appropriately in both parts of this project. Upon reflection, we found potential opportunities for improvement. Beginning with LDA, implementing a section for either Grid Search or Random Search could further optimize hyperparameters. KeyBERT could have been more effective had we run the model on a subset of complete documents. While our application of KeyBERT for this project was somewhat unorthodox, the results were nevertheless insightful.

The results of topic modeling offer meaningful insights with real-world applications. The process of identifying latent topics across the NIPS papers corpus demonstrates how large, unstructured datasets can be classified. Organizations with vast text repositories can use various techniques to categorize information and improve retrieval efficiency. Topic modeling is extremely relevant for product and service personalization. For example, in the realm of e-commerce, businesses can leverage topic modeling algorithms to analyze user-generated data and cluster it by overarching themes, enabling more targeted consumer support. With customer feedback, it can enhance the analysis of reviews, surveys, and other social media data to detect emerging trends.

Work Cited

Grootendorst, M. *Quickstart – KeyBERT*. https://maartengr.github.io/KeyBERT/guides/quickstart.html

Tran, K. (2025, April 7). *pyLDAvis: Topic Modeling Exploration Tool Every NLP Data Scientist Should Know*. Neptune.ai. https://neptune.ai/blog/pyldavis-topic-modeling-exploration-tool