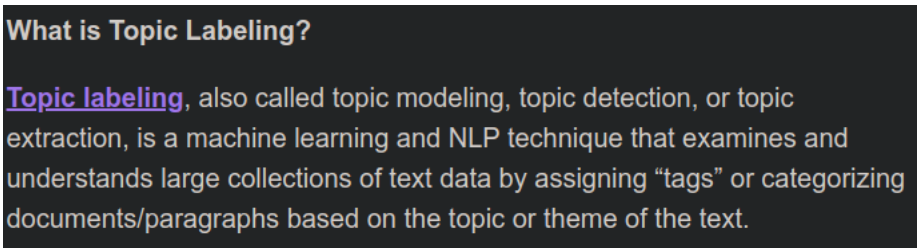
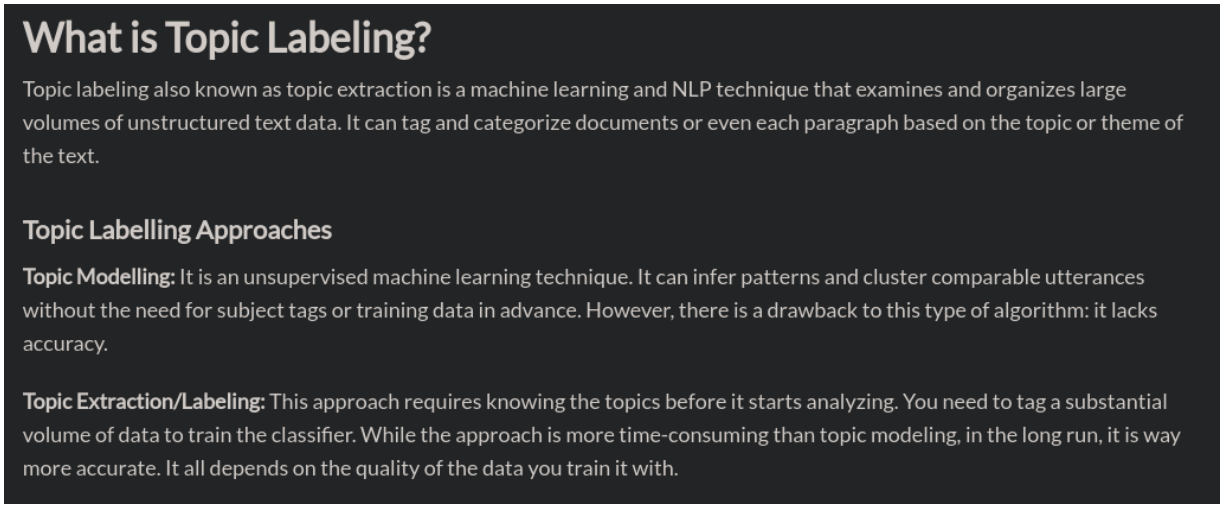
Topic Labeling Literature

▼ Link 1: https://dev.to/shreyasahani/topic-labeling-a-beginners-guide-3jn6



▼ Link 2: https://www.bytesview.com/blog/topic-labeling/



▼ Link 3: https://www.researchgate.net/publication/220874747_Automatic_La

Abstract

We propose a method for automatically labelling topics learned via LDA topic models. We generate our label candidate set from the top-ranking topic terms, titles of Wikipedia articles containing the top-ranking topic terms, and sub-phrases extracted from the Wikipedia article titles. We rank the label candidates using a combination of association measures and lexical features, optionally fed into a supervised ranking model. Our method is shown to perform strongly over four independent sets of topics, significantly better than a benchmark method.

Use of Wikipedia

- Single word lable, based on what meaning it gives, that should make the label of the topic.

> The task of automatic labelling of topics is a natural progression from the best topic term selection task of Lau et al. (2010). In that work, the authors use a reranking framework to produce a ranking of the top-10 topic terms based on how well each term – in isolation – represents a topic. For example, in our *stock market investor fund trading ...* topic example, the term *trading* could be considered as a more representative term of the overall semantics of the topic than the top-ranked topic term *stock*.

- Combining the top N words to make a label.

> While the best term could be used as a topic label, topics are commonly ideas or concepts that are better expressed with multiword terms (for example STOCK MARKET TRADING), or terms that might not

- Wikipedia extraction.

$$|(\cup_{p \in O(a)} C(p)) \cap (\cup_{p \in O(b)} C(p))|$$

- A label will only be considered a good label candidate if the outlinks it has belong to the list of categories the article itself belongs to.

Finally, we add the top-5 topic terms to the set of candidates, based on the marginals from the original topic model. Doing this ensures that there are always label candidates for all topics (even if the Wikipedia searches fail), and also allows the possibility of labeling a topic using its own topic terms, which was demonstrated by Lau et al. (2010) to be a baseline source of topic label candidates.

- Concatenated the top 5 words extracted by the LDA/PAM model for a given topic, to be used as a label condidate or label itself, if search results fail to yield anything.

the topic terms. To learn the association of a label candidate with the topic terms, we use several lexical association measures: pointwise mutual information (PMI), Student's t-test, Dice's coefficient, Pearson's

- Find the frequency of occurance of the label candidate terms and topic terms from wikipedia results, average them, get search engine score for the terms and candidates to find the association between them, and hence, that would be the label for the topic.

We took a standard approach to topic modelling each of the four document collections: we tokenised, lemmatised and stopped each document,[5] and created a vocabulary of terms that occurred at least ten times. From this processed data, we created a

- Topic modeling techniques.

▼ Link 4: https://python.gotrained.com/lexical-resources-nltk/

▼ Link 5: https://www.pluralsight.com/guides/topic-identification-nlp