Machine Learning for NLP 1

Exercise 6

Topic Modeling

University of Zurich

Contents

1. Problem Statement
2. Data Preprocessing
3. Vectorizer and Label Encoder
4. Indent the first line of each paragraph.

Problem Statement

The objective of this exercise is to explore topic modeling techniques, specifically Latent Dirichlet Allocation (LDA) and Combined Topic Models (CTMs). Using the "dblp" database, which includes metadata on computer science publications, the task involves performing topic modeling on publication titles to identify significant topics and trends within the field of computer science. The exercise aims to enhance understanding and application skills in topic modeling using LDA and PLM-based approaches.

Data Preprocessing

Part 1

For Part 1, we use basic preprocessing, i.e., we remove all non-alphabetic characters and lower-case all the characters.

Part 2

For Part 2, we use the NLTK library to remove stop words from the titles and then use the WhiteSpacePreprocessingStopwords function from the contextualized_topic_models.utils.preprocessing module that removes stop-words and white-spaces.

Vectorizer and Label Encoder

Part 1

After preprocessing our data, we used the count-vectorizer with the following parameters- *max_df=0.95, min_df=2, max_features=num_features, stop_words='english'* to vectorize our data.

Note: The vectorizer will omit stop-words because stop_words is set to ‘english’ (which is a “True” case).

Then, we use LatentDirichletAllocation function from sklearn.decomposition module to perform LDA on our preprocessed and vectorized data. The parameters for our LDA function are as follows- *n_components=num_lda_topics, max_iter=5, learning_method='online', random_state=42*.

Part 2

We vectorized our text using TF-IDF for both words and characters. We wanted to use the TF-IDF vectorization of characters as an additional feature along with TF-IDF vectorization of words. We choose the *ngram_range* parameter as (1,1), i.e., it considers one word or character for vectorization. While finding the optimal hyperparameters in grid search, we will experiment with this parameter.

We use Label Encoder from sklearn to encode our labels and to make them numeric for the models to be able to work with them.

Results and Discussions

Part 1

From 1990 to 2009

The output from the LDA model was-

1. Topic 0: based information new network systems model estimation modeling time approach fuzzy image

2. Topic 1: design method theory computing identification structure case digital application sets implementation search
3. Topic 2: using systems control analysis networks linear nonlinear algorithm adaptive models problem optimal
4. Topic 3: data model neural detection software learning development knowledge power codes prediction set
5. Topic 4: study dynamic graphs management scheme systems programming logic realtime space tracking properties

We concluded the following topics from these outputs-

1. Information Theory (Words like information, networks, modeling)
2. Algorithm Design (Words like theory, computing, implementation)
3. Linear and non-linear modeling of data
4. Neural Networks/Deep Learning (Words like prediction, learning, neural)
5. Dynamic Programming (Words like dynamic, realtime, tracking)

2010 onwards

The output from the LDA model was-

1. Topic 0: using networks model systems network algorithm detection neural efficient performance wireless time
2. Topic 1: optimization image application equations applications methods hybrid new identification smart digital sensing
3. Topic 2: based study information deep framework mobile classification prediction problem multiple management approach

4. Topic 3: control learning nonlinear estimation linear design distributed robust optimal power problems approach
5. Topic 4: analysis data method systems adaptive dynamic energy recognition finite selection images graphs

We concluded the following topics from these outputs-

1. Neural Networks and Algorithms
2. Image Classification and Analysis
3. Deep Learning and Classification
4. Algorithm Design
5. Incoherent (Since, we could not make any conclusive topic out of the output words)

Discussion

The topics do make sense to us as the top words for each topic appear to be correlated in most cases. Therefore, we can assume the topic from the list of top words for each topic.

We can observe a shifting trend in the interest of researchers in the field of computer science. The research interest shifted from logic and theory of computing in the papers before the 1990s to information theory and the onset of machine learning in the papers from 1990 to 2009 and then deep learning, image classification and prediction dominated the topics of the papers from 2010 onwards.

Part 2

Before the 1990s

The output from the CTM was-

1. Topic 1: 'digital', 'fault', 'analysis', 'design', 'error'

2. Topic 2: 'note', 'problems', 'technical', 'problem', 'linear'
3. Topic 3: 'network', 'communications', 'memory', 'digital', 'communication'
4. Topic 4: 'systems', 'model', 'decision', 'distributed', 'control'
5. Topic 5: 'code', 'probability', 'random', 'surface', 'generator'

We concluded the following topics from these outputs-

1. Digital Design
2. Incoherent (Since, we could not make any conclusive topic out of the output words)
3. Communication Networks
4. Control Systems
5. Incoherent (Since, we could not make any conclusive topic out of the output words)

From 1990 to 2009

The output from the CTM was-

1. Topic 1: 'power', 'low', 'high', 'phase', 'circuit'
2. Topic 2: 'models', 'model', 'markov', 'distribution', 'estimation'
3. Topic 3: 'uuml', 'der', 'und', 'de', 'von'
4. Topic 4: 'service', 'web', 'services', 'environments', 'management'
5. Topic 5: 'image', 'recognition', 'images', 'detection', 'segmentation'

We concluded the following topics from these outputs-

1. Power Modeling
2. Markov Models
3. Incoherent (Since, we could not make any conclusive topic out of the output words)
4. Web Services and Management
5. Image Recognition and Analysis

Discussion

There is a clear difference between the performance of LDA and CTM in our topic modeling task. The output from LDA is a mix of words that could belong to several topics. However, in situations like these, where there is not a clear winner, CTM's output is very incoherent.
Eg. in topic 2 from the 1990 to 2009 period, we had the following words from LDA- *using systems control analysis networks linear nonlinear algorithm adaptive models problem optimal*. There could be multiple topics within these words. For eg. Control Systems, Algorithm Design, etc. For this topic, the output from CTM was *'note', 'problems', 'technical', 'problem', 'linear'*. There is a clear lack of theme here.

However, in cases where there is more context at our disposal, CTM outperforms LDA and gives us very accurate, definitive and almost overfitting results. Unlike LDA, the words in the output of CTM are very closely linked together, for eg., *'image', 'recognition', 'images', 'detection', 'segmentation'*. This gives us a clear topic idea, i.e., Image Recognition and Analysis.

To conclude, given enough context, CTM outperforms LDA considerably. However, where there is not enough context or correlation between the data, where CTM gives us completely incoherent results, LDA gives us a mix of words out of which there could be several candidate topics.