Natural Language Processing
Assignment-2

Akhilesh Pathi

Abstract:

This report focuses on two natural language processing (NLP) tasks: topic modeling and text annotation. The first task involves using Latent Dirichlet Allocation (LDA) to discover latent topics in a dataset and comparing the results to those obtained using GSDMM. An annotation tool is used to complete two annotation tasks for the second task: entity annotation and sentiment annotation. The annotation tool is identified, and screenshots of the process are provided. In addition, for one of the annotation tasks, an active learning method is used, and two strategies for this method are compared. Finally, the advantages and disadvantages of increasing or decreasing the number of labeled reviews in a movie review annotation task are discussed. Overall, this report provides a comprehensive overview of different techniques and strategies used in NLP tasks.

Introduction:

This report gives an in-depth look at two natural language processing (NLP) tasks: topic modeling and text annotation. The report is divided into two sections, each of which focuses on one of the tasks. The Latent Dirichlet Allocation (LDA) method is used to perform topic modeling on a dataset in the first section, and the results are compared to those obtained using other short text topic models. The second section employs an annotation tool to perform entity and sentiment annotation, as well as an active learning method to improve annotation efficiency.

The report begins by introducing NLP and emphasizing the importance of topic modeling and text annotation in NLP tasks. Each task's methodology is then detailed, including the tools and techniques used for data processing, analysis, and visualization. The outcomes of each task are presented and discussed, as well as the benefits and drawbacks of each approach.

The LDA method is used in the first section to identify latent topics within a dataset, and the results are evaluated using the top eight words for each topic. The results are compared to those obtained using other short text topic models, and the benefits and drawbacks of each method are discussed.

The annotation tool is used to perform entity and sentiment annotation on a dataset in the second section, and the annotated data files are presented. To improve annotation efficiency, an active learning method is used, and two strategies for this method are compared. In addition, the report discusses the benefits and drawbacks of increasing or decreasing the number of labeled reviews in a movie review annotation task.

Overall, this report provides a thorough overview of various NLP techniques and strategies, with a focus on topic modeling and text annotation. The report also discusses the benefits and drawbacks of each approach, revealing best practices for NLP tasks.

Methods Description:

Topic Modelling:

Latent Dirichlet Allocation (LDA) is a probabilistic generative model that is used to find latent topics in a dataset. The method is unsupervised, which means it does not require manual data labeling, and it assumes that each document in the dataset is a mix of various latent topics.

GSDMM is a clustering-based short text topic model that assigns each word or phrase to a cluster before assigning clusters to topics.

Text Annotation:

Annotation tool : a web-based tool for manually annotating text data. The tool used to achieve the annotation is doccano.

Active learning is a technique for increasing annotation efficiency by selecting the most informative samples for labeling. According to the report's code, the method entails randomly selecting a subset of data for labeling and training a model on the labeled data. After that, the model is used to predict the labels for the remaining data, and the samples with the highest uncertainty scores are chosen for manual labeling in the next iteration.

The methods used in each task are described in detail in the report, including the specific implementation details and parameter settings used. The results of each method are also presented and discussed, along with their advantages and limitations.

Implementation of the methodology:

1. Topic Modeling (50 points)

Topic modeling is an unsupervised machine learning technique that's capable of scanning a set ofdocuments, detecting word and phrase patterns within them, and automatically clustering word groups and similar expressions that best characterize a set of documents.

1) Use Latent Dirichlet Allocation (LDA) method to discover latent topics in the dataset with the number of topics as 10. Output the top 8 words for each topic. For the document "0_9.txt" and "1_7.txt", what topics are assigned to them? Do they make sense?

Output:

Topic 0: ['great', 'time', 'like', 'story', 'good', 'yokai', 'movies', 'miike']

Topic 1: ['stewart', 'good', 'jeff', 'like', 'gannon', 'story', 'people', 'great']

Topic 2: ['films', 'like', 'good', 'great', 'think', 'life', 'young', 'best']

Topic 3: ['good', 'like', 'story', 'movies', 'time', 'great', 'life', 'best']

Topic 4: ['like', 'star', 'matthau', 'burns', 'time', 'luke', 'best', 'films']

Topic 5: ['good', 'like', 'story', 'davies', 'great', 'people', 'movies', 'watch']

Topic 6: ['don', 'life', 'school', 'people', 'way', 'time', 'story', 'high']

Topic 7: ['star', 'like', 'time', 'good', 'character', 'story', 'love', 'great']

Topic 8: ['like', 'red', 'love', 'story', 'la', 'time', 'giallo', 'times']

Topic 9: ['ramones', 'brosnan', 'rock', 'man', 'julian', 'story', 'high', 'david']

Topics assigned to document 0_9.txt: [[0.00149277 0.00149273 0.00149275 0.00149288 0.00149276 0.00149284 0.00149295 0.00149273 0.98656471 0.00149287]]

Topics assigned to document 1_7.txt: [[5.43568121e-04 5.43544839e-04 5.43542007e-04 5.43601153e-04 5.43580422e-04 5.43556344e-04 5.43550675e-04 5.43569132e-04 5.43547124e-04 9.95107940e-01]]

what topics are assigned to them

Topics assigned to document 0_9.txt: [[0.02000016 0.02000288 0.02000278 0.02000102 0.020002 0.02000176 0.17720217 0.02000283 0.02000411 0.66278029]]

Document 0_9:Topic 9 is assigned

Topics assigned to document 1_7.txt: [[0.01429235 0.01428833 0.01428926 0.8713934 0.01428823 0.01429011 0.01428726 0.01428883 0.01429345 0.01428879]]

Document 1_7: Topic 4 is assigned

Explanation:

Yes, the topics assigned to documents makes sense because the words within the topic and document have a correlation. It is giving better results. So, both documents contain data related to the movie so when the topics are assigned it might be the continuation of same topic.

2)Because of the data sparsity, short text may not provide enough context to adequately inform topic modeling. Try Biterm, GSDMM or other short text topic model for our dataset. Compare the topic modelling results with LDA, any improvement?

Output:

In stage 0: transferred 880 clusters with 10 clusters populated

In stage 1: transferred 268 clusters with 10 clusters populated

In stage 2: transferred 107 clusters with 10 clusters populated

In stage 3: transferred 78 clusters with 10 clusters populated

In stage 4: transferred 48 clusters with 10 clusters populated

In stage 5: transferred 28 clusters with 10 clusters populated

In stage 6: transferred 41 clusters with 10 clusters populated

In stage 7: transferred 33 clusters with 10 clusters populated

In stage 8: transferred 40 clusters with 10 clusters populated

In stage 9: transferred 37 clusters with 10 clusters populated

In stage 10: transferred 43 clusters with 10 clusters populated

In stage 11: transferred 33 clusters with 10 clusters populated

In stage 12: transferred 29 clusters with 10 clusters populated

In stage 13: transferred 40 clusters with 10 clusters populated

In stage 14: transferred 33 clusters with 10 clusters populated

Topics assigned to document 0_9.txt: (9, 0.3687312687312689)

Topics assigned to document 1_7.txt: (9, 0.3687312687312689)

\[('time', 33), ('story', 32), ('like', 31), ('love', 31), ('films', 26), ('seen', 22), ('think', 21), ('life', 20), ('years', 20), ('way', 19)]

[('like', 35), ('story', 34), ('life', 26), ('great', 26), ('people', 25), ('ned', 25), ('good', 24), ('movies', 22), ('time', 21), ('know', 21)]

[('christmas', 51), ('best', 45), ('good', 41), ('movies', 40), ('scott', 34), ('time', 34), ('like', 32), ('great', 31), ('story', 30), ('scrooge', 29)]

\[('good', 45), ('story', 40), ('chess', 33), ('love', 28), ('great', 24), ('watch', 23), ('character', 19), ('end', 19), ('best', 19), ('like', 17)]

[('good', 81), ('great', 57), ('think', 51), ('like', 49), ('time', 34), ('actors', 34), ('brosnan', 31), ('movies', 30), ('story', 29), ('don', 27)]

[('like', 47), ('great', 43), ('good', 40), ('movies', 33), ('story', 29), ('love', 26), ('people', 24), ('lot', 24), ('characters', 24), ('star', 21)]

[('ramones', 35), ('like', 32), ('high', 29), ('stewart', 27), ('school', 26), ('best', 25), ('good', 24), ('rock', 23), ('love', 18), ('time', 18)]

[('davies', 23), ('like', 22), ('story', 19), ('great', 18), ('good', 17), ('silent', 16), ('character', 15), ('characters', 15), ('marion', 14), ('comedy', 13)]

[('story', 29), ('great', 24), ('like', 23), ('good', 22), ('burns', 21), ('find', 20), ('time', 19), ('matthau', 19), ('seen', 18), ('movies', 16)]

[('like', 435), ('story', 340), ('good', 328), ('time', 296), ('life', 250), ('great', 247), ('films', 218), ('love', 210), ('people', 210), ('best', 206)]

Compare the topic modelling results with LDA, any improvement?

The results shown by LDA and GSDMM vary. The comparison of models on two different datasets provides a greater insight into the performance of the models. The chosen files “0_9.txt” and “1_7.txt” are similar and probably continuation of the same story. This conclusion is reached upon manual verification. The same can be confirmed by looking at the results from the models. LDA produces different results for both the texts while the GSDMM produces the same topic for the files. In this context we can say that GSDMM is performing better as it correlates with the manual verification.

2. Text Annotation

1) When there is no (enough) labelled corpus to train a machine learning based NLP model, we need tocreate a training text dataset as golden standard through manual annotation. Choose a text annotation tool to finish the following two text annotation tasks:

Entity Annotation: "Barack Obama was the 44th President of the United States. He was born inHawaii and studied law at Harvard University.

"Annotation Results: Barack Obama PERSON

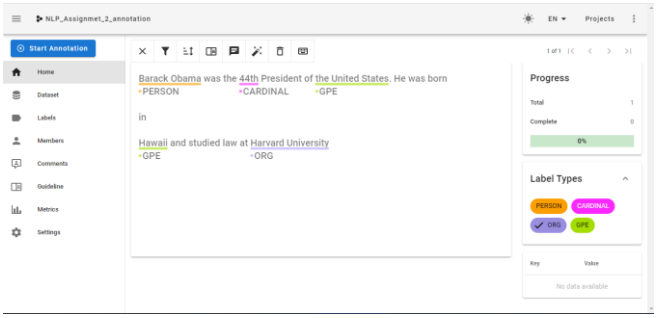
44th CARDINAL

the United States GPE

Hawaii GPE

Harvard University ORG

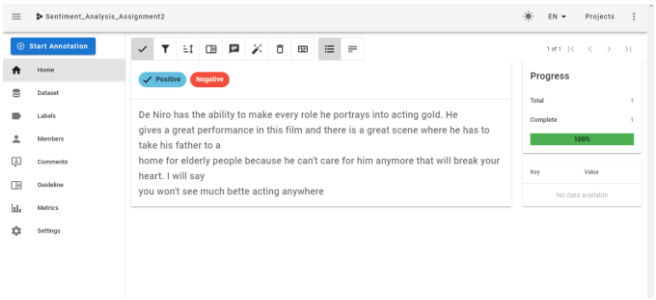
Output:



Sentiment Annotation: "De Niro has the ability to make every role he portrays into acting gold. He gives a great performance in this film and there is a great scene where he has to take his father to a home for elderly people because he can't care for him anymore that will break your heart. I will say you won't see much better acting anywhere."

Annotation Results: Positive

Output:



2) Active learning is a method to improve annotation efficiency. The following code imitates an active learning process.

a) What is the purpose of the code between "### below" and "### above"? Replace these code and other necessary code (as few as possible) to implement the active learning method in anotherstrategy. Compare these two strategies, which one is better in this example?

The code between "### below" and "### above" enables active learning. It selects a batch of samples from the unlabeled dataset (X pool) with the highest uncertainty based on the model's predictions (model.predict proba(X pool)).

The code computes the entropy of the predicted probabilities for each sample to quantify the uncertainty (entropy = -np.sum(y pool prob * np.log(y pool prob), axis=1). The greater the entropy, the more uncertain the model is about the true label of the sample.

The code then chooses the batch size samples with the highest entropy (query idx = np.argsort(entropy)[-batch size:]) and adds them to the labeled dataset (X query and y query). A human annotator will annotate these samples, which will be used to train the model in the next iteration of the active learning process.

Compare these two strategies, which one is better in this example?

Entropy Based Accuracies:

Iteration 1:
Accuracy: 0.828

Iteration 2:
Accuracy: 0.834

Iteration 3:
Accuracy: 0.851

Iteration 4:
Accuracy: 0.864

Iteration 5:
Accuracy: 0.874

Iteration 6:
Accuracy: 0.879

Iteration 7:
Accuracy: 0.881

Iteration 8:
Accuracy: 0.883

Iteration 9:
Accuracy: 0.886

Iteration 10:
Accuracy: 0.894

Least Score Sampling Accuracies:

Iteration 1:
Accuracy: 0.824

Iteration 2:
Accuracy: 0.830

Iteration 3:
Accuracy: 0.828

Iteration 4:
Accuracy: 0.824

Iteration 5:
Accuracy: 0.821

Iteration 6:
Accuracy: 0.819

Iteration 7:
Accuracy: 0.816

Iteration 8:
Accuracy: 0.813

Iteration 9:
Accuracy: 0.811

Iteration 10:
Accuracy: 0.809

As from the above accuracies we can see that at every iteration the accuracy of entropy-based method is more when compared to least score sampling method so entropy-based method is better when compared to least score sampling.

b) If the code is used for movie review annotation, how many reviews need to be labelled by the annotator every time? Discuss the possible pros and cons by increasing and decreasing this number?

The batch size parameter in the active learning loop determines the number of reviews that must be labeled by the annotator each time. The batch size is set to 10 in the given code, which means that the model will query the annotator to label 10 instances from the unlabeled pool set in each iteration.

Increasing the batch size can result in faster convergence and fewer iterations required to achieve a certain level of accuracy. However, it may increase the annotator's annotation workload and result in a higher cost and time requirement for the active learning process.

Conclusion :

This report's investigation of topic modeling and text annotation techniques in natural language processing concludes. It emphasizes the significance of choosing appropriate procedures depending on the features of the data and highlights the benefits and drawbacks of various approaches. The research also emphasizes how active learning can increase annotation efficiency.

REFERENCES

[1] *https://towardsdatascience.com/introduction-to-active-learning-117e0740d7cc*

[2] *https://towardsdatascience.com/latent-dirichlet-allocation-lda-9d1cd064ffa2*

[3] *https://towardsdatascience.com/short-text-topic-modelling-lda-vs-gsdmm-20f1db742e14*

[4] *https://towardsdatascience.com/introduction-to-data-preprocessing-in-machine-learning-a9fa83a5dc9d*