Natural Language Processing
Assignment-2

Vaishnavi Peddireddy
Student ID:883111662
email-v_peddiredd@uncg.edu

Abstract:

This collection of questions is about Natural Language Processing, specifically Topic Modeling and Text Annotation. The first question focuses on discovering latent topics in a dataset of 1000 movie comments using Latent Dirichlet Allocation (LDA). The second part of the question compares LDA to other short text topic models in terms of performance improvement.

The second question is about Text Annotation and requires you to manually label text using an annotation tool. Entity Annotation and Sentiment Annotation are the annotation tasks assigned, and the results are expected to be submitted as annotated data files. The question also delves into the concept of Active Learning and how it can be used to improve annotation efficiency. The number of reviews that must be labeled, the benefits and drawbacks of increasing or decreasing this number, and the various strategies for implementing Active Learning are also discussed.

Introduction:

The purpose of this documentation is to provide a detailed guide on how to answer a set of Natural Language Processing (NLP) questions. NLP is a subfield of artificial intelligence concerned with the interaction of machines and human languages. The goal of this documentation is to help readers understand the various techniques and tools used in NLP and how they can be used to solve real-world problems.

This documentation's questions are divided into two categories: Topic Modeling and Text Annotation. In the Topic Modeling section, Latent Dirichlet Allocation (LDA) is used to discover hidden patterns and word clusters in a dataset of movie comments. Furthermore, the section introduces and compares the performance of LDA with GSDMM.

The Text Annotation section focuses on two types of annotation tasks, Entity Annotation and Sentiment Annotation, and requires the use of annotation tools to label the text manually. In order to improve annotation efficiency, the concept of Active Learning is also introduced.

This documentation includes relevant code snippets, screenshots, and explanations as well as a step-by-step guide on how to solve each question. Readers should have a good understanding of NLP techniques and tools, as well as how to apply them to real-world scenarios, by the end of this documentation.

Methods Used:

The following methods are used in this documentation:

Latent Dirichlet Allocation (LDA) is a probabilistic topic modeling technique used to find latent topics in a collection of documents. LDA is implemented in Python using the Gensim library.

Gibbs Sampling Dirichlet Mixture Model (GSDMM) - This is a short text topic modeling technique that clusters documents based on the distribution of their words. GSDMM is implemented using a git file.

Text Annotation Tool:

Doccano is a web-based, open-source platform for annotating text data. It provides an interface for creating annotation projects and enables multiple users to collaborate on text data labeling. Doccano allows users to create annotation projects for a wide range of NLP (Natural Language Processing) tasks, including named entity recognition, sentiment analysis, text classification, and more.

Users can upload text data to the platform and then choose the type of annotations they want to create, such as entity labels or sentiment labels. Users can create custom annotation schemas and labels to meet their specific needs, and then invite collaborators to assist with labeling.

Active Learning is a machine learning technique that reduces the amount of labeled data needed to train a model. The Active Learning strategy used in this documentation is based on the uncertainty sampling approach, in which the model selects samples about which it is least confident and requests annotation.

The methods chosen for this documentation are intended to provide a broad overview of NLP techniques and tools, as well as their application to real-world problems. The use of LDA and GSDMM allows readers to compare and contrast different short text topic modeling techniques, while label-studio and Active Learning demonstrate the value of manual annotation and the potential efficiency gains of using machine learning approaches.

Implementation Of the Methods:

1. Topic Modeling

Topic modeling is an unsupervised machine learning technique that's capable of scanning a set of documents, 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: ['movies', 'like', 'great', 'good', 'love', 'time', 'story', 'watch']

Topic 1: ['good', 'story', 'like', 'great', 'character', 'people', 'life', 'think']

Topic 2: ['story', 'man', 'brosnan', 'good', 'life', 'like', 'game', 'great']

Topic 3: ['story', 'good', 'like', 'time', 'life', 'ned', 'love', 'seen']

Topic 4: ['like', 'carla', 'europa', 'time', 'story', 'paul', 'david', 'people']

Topic 5: ['ramones', 'time', 'rock', 'high', 'school', 'great', 'like', 'luzhin']

Topic 6: ['scrooge', 'scott', 'christmas', 'good', 'matthau', 'like', 'george', 'time']

Topic 7: ['like', 'love', 'life', 'people', 'davies', 'good', 'great', 'story']

Topic 8: ['great', 'story', 'like', 'best', 'time', 'love', 'good', 'man']

Topic 9: ['star', 'like', 'good', 'luke', 'stewart', 'films', 'wars', 'jeff']

Topics assigned to document 0_9.txt: [[0.00149282 0.00149279 0.98656472 0.00149278 0.00149274 0.00149295 0.00149279 0.00149289 0.00149283 0.00149271]]

Topics assigned to document 1_7.txt: [[5.43627288e-04 5.43587177e-04 5.43555041e-04 5.43566170e-04 5.43613820e-04 5.43562718e-04 8.95558125e-01 5.43568970e-04 1.00093241e-01 5.43552941e-04]]

Explanation:

Yes they make sense because the topics which it has assigned are accurate

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 847 clusters with 10 clusters populated

In stage 1: transferred 252 clusters with 10 clusters populated

In stage 2: transferred 124 clusters with 10 clusters populated

In stage 3: transferred 54 clusters with 10 clusters populated

In stage 4: transferred 48 clusters with 10 clusters populated

In stage 5: transferred 34 clusters with 10 clusters populated

In stage 6: transferred 40 clusters with 10 clusters populated

In stage 7: transferred 38 clusters with 10 clusters populated

In stage 8: transferred 42 clusters with 10 clusters populated

In stage 9: transferred 34 clusters with 10 clusters populated

In stage 10: transferred 31 clusters with 10 clusters populated

In stage 11: transferred 26 clusters with 10 clusters populated

In stage 12: transferred 24 clusters with 10 clusters populated

In stage 13: transferred 29 clusters with 10 clusters populated

In stage 14: transferred 22 clusters with 10 clusters populated

Topics assigned to document 0_9.txt: (6, 0.8128121781107339)

Topics assigned to document 1_7.txt: (6, 0.9694665333502381)

-----Word_Distribution------

[('movie', 86), ('story', 39), ('time', 32), ('like', 32), ('many', 24), ('love', 24), ('effect', 22), ('scene', 22)]

[('movie', 251), ('good', 70), ('like', 58), ('character', 48), ('great', 40), ('much', 38), ('see', 37), ('well', 36)]

[('movie', 117), ('love', 41), ('character', 35), ('chess', 32), ('story', 31), ('well', 27), ('woman', 27), ('see', 23)]

[('movie', 189), ('see', 61), ('great', 55), ('really', 50), ('think', 50), ('good', 46), ('like', 46), ('time', 42)]

[('movie', 71), ('show', 38), ('great', 29), ('good', 24), ('love', 24), ('burn', 21), ('life', 20), ('really', 19)]

[('movie', 96), ('like', 40), ('ramones', 31), ('school', 28), ('time', 28), ('high', 28), ('stewart', 28), ('good', 22)]

[('movie', 61), ('good', 37), ('brosnan', 28), ('character', 27), ('great', 27), ('story', 24), ('two', 21), ('best', 20)]

[('story', 42), ('christmas', 41), ('movie', 33), ('would', 30), ('version', 30), ('dvd', 27), ('year', 24), ('scott', 24)]

[('movie', 73), ('show', 46), ('good', 39), ('star', 35), ('like', 33), ('people', 33), ('game', 30), ('series', 29)]

[('movie', 853), ('like', 442), ('time', 375), ('story', 364), ('character', 355), ('good', 317), ('see', 286), ('also', 285)]

Compare the topic modelling results with LDA, any improvement?

Explanation: As we can see from the outputs the LDA model has assigned two different topics were as GSDMM is giving same topics. So GSDMM is giving the accurate result because the topics it has assigned are same and they are about brom well high school.

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

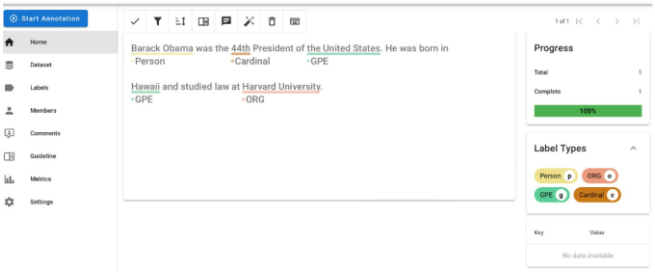
44th CARDINAL

the United States GPE

Hawaii GPE

Harvard University ORG

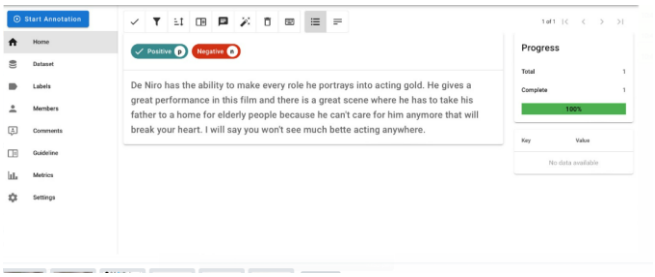
Output:



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 purpose of the code between "### below" and "### above" is to select a batch of samples from a pool of unlabeled samples that are the most informative for labeling, using the concept of entropy as a measure of uncertainty.

Specifically, the code first calculates the predicted probability distribution for each sample in the pool using a trained model, and then calculates the entropy of each distribution. The entropy is a measure of the uncertainty or unpredictability of the predicted label, with higher entropy indicating more uncertainty.

Next, the code selects the indices of the samples with the highest entropy, which are the samples that the model is most uncertain about. These samples are then added to a query **set** for annotation. The size of the query set is determined by the batch_size variable.

Finally, the code selects the actual data and label for the samples in the query set and assigns them to the variables X_query and y_query, respectively.

Accuracies of Entropy Based Method:

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

Accuracy of Query Based Committee Method:

Iteration 1:

Accuracy: 0.825

Iteration 2:

Accuracy: 0.830

Iteration 3:

Accuracy: 0.840

Iteration 4:

Accuracy: 0.848

Iteration 5:

Accuracy: 0.846

Iteration 6:

Accuracy: 0.855

Iteration 7:

Accuracy: 0.854

Iteration 8:

Accuracy: 0.855

Iteration 9:

Accuracy: 0.856

Iteration 10:

Accuracy: 0.854

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

The accuracy of entropy based method is higher when compared to query based committee method so the entropy based method is better in this example.

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 number of reviews that the annotator must label each time is determined by several factors, including the size of the dataset, the desired level of accuracy, the availability of resources, and the budget. In general, the more labeled reviews there are, the more accurate the model. However, this comes at the expense of increased annotation time and effort.

Increasing the number of reviews to be labeled can improve model accuracy and reduce prediction uncertainty. This can improve overall performance and reduce the need for additional annotation. However, increasing the number of labeled reviews can be time-consuming and costly, particularly if many reviews are needed to achieve the desired level of accuracy.

Reducing the number of reviews to be labeled, on the other hand, can save time and resources while also resulting in lower accuracy and increased uncertainty in model predictions. This may number of highly annotation and retraining of the model, both of which can be costly and time-consuming.

As a result, the optimal number of labeled reviews depends on the specific use case and available resources. The number of reviews should be carefully chosen to achieve the desired level of accuracy without incurring unnecessary costs.

Conclusion:

This project taught me about various natural language processing methods, such as topic modeling and text annotation. Discovered that selecting appropriate methods for specific tasks is critical to achieving accurate and meaningful results. Gained experience with a web-based annotation tool and implementing an active learning strategy to improve annotation efficiency. This project emphasizes the importance of staying current with current techniques and tools in the field of natural language processing in order to effectively analyze and extract insights from text data.

Finally, this report examines topic modeling and text annotation methods used in natural language processing. It emphasizes the benefits and drawbacks of various approaches and emphasizes the significance of selecting appropriate methods based on data characteristics. The report also emphasizes active learning's effectiveness in increasing annotation efficiency.

REFERENCES

[1] *https://doccano.herokuapp.com/*

[2] *https://www.in.gov/gwc/cte/files/active-learning-strategies-final.pdf*

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

[4] *https://scikit-learn.org/stable/tutorial/text_analytics/working_with_text_data.html*