DATA AND
ARTIFICIAL INTELLIGENCE

Natural Language Processing

Course-End Project

simplilearn

DATA AND
ARTIFICIAL INTELLIGENCE

Topic Analysis of Review Data

Objectives

- To assist a major mobile brand in comprehending the voice of the consumer and the subjects that customers are discussing by examining the product reviews on Amazon
- To use machine learning and Python to comprehend customer voice and forecast the review rating based on Amazon's product
- To perform topic modeling on specific parts of speech
- To interpret the emerging topics using customer voice

Prerequisites

- Data exploration
- POS tagging
- Model creation
- Python
- Topic modeling
- Linear discriminant analysis (LDA)
- Natural language toolkit (NLTK)

Industry Relevance

- **Data exploration:** It is used to find trends and patterns or to check assumptions by analyzing data with visual tools.
- **POS tagging:** It is used in text analysis tools and algorithms as well as corpus searches.
- **Model creation:** It uses Python's backend to create a sequential model.
- **Python:** It is widely used to implement data analysis and machine learning.

Industry Relevance

- **Topic modeling:** It is used to discover the themes that run through a corpus by analyzing the words of the original texts.
- **LDA:** It is used as a preprocessing step in machine learning and it is mainly used for classification problems.
- **NLTK:** It is a platform used for building Python programs that work with human language data for applications in statistical natural language processing (NLP).

Problem Statement

A popular mobile phone brand, Lenovo, has launched its budget smartphone in the Indian market. The client wants to understand the voice of the customer (VOC) on the product. This will be useful to not just evaluate the current product but also to get some direction for developing the product pipeline.

The client is particularly interested in the different aspects that the customers care about. Product reviews by customers on a leading e-commerce site should provide a good amount of data.

Dataset

Dataset available to complete the project:

K8 Reviews v0.2.csv

Dataset Description

Variable	-	Description
Sentiment	-	The sentiment against the review (4- and 5-star reviews are positive, and 1-, 2-, and 3-star reviews are negative)
Reviews	-	The main text of the review

Tasks to Perform

Discover the topics in the reviews, and present them to the business in a consumable format by utilizing syntactic processing and topic modeling techniques.

Perform specific cleanup and POS tagging and add restrictions to relevant POS tags. Then, perform topic modeling using LDA. Finally, give business-friendly names to the topics, and make a table for the business.

1. Read the .csv file using Pandas, and look at the first few top records

2. Normalize the casing of the review text, and extract the text into a list for easier manipulation.

Tasks to Perform

3. Tokenize the reviews using NLTK's **word_tokenize** function

4. Perform parts-of-speech tagging on each sentence using the NLTK POS tagger

5. For the topic model, include nouns only
   a. Find all POS tags that correspond to nouns
   b. Limit the data to terms with these tags

Tasks to Perform

6. Lemmatize
    a. The different forms of the terms need to be treated as one
    b. For the time being, there is no need to provide a POS tag to the lemmatizer

7. Remove stopwords and punctuations (if any)

8. Create a topic model using LDA on the cleaned-up data with 12 topics
    a. Print the top terms for each topic
    b. Find the coherence of the model with the **c_v** metric

Tasks to Perform

9. Analyze the topics through the business lens
    a. Determine which of the topics can be combined

10. Create a topic model using LDA with what you think is the optimal number of topics
    a. Find the coherence of the model

11. Businesses should be able to interpret the topics.
    a. Name each of the identified topics
    b. Create a table with the topic names and the top ten terms in each to present to the business

Project Outcome

- This project is designed to help one perform exploratory data analysis and detect toxic comments using Python.
- This project also helps one performs topic modeling and use machine learning and Python to create a rating prediction model.

Submission Process

1. Complete the project in the Simplilearn Lab
2. Complete each task listed in the problem statement
3. Take screenshots of the results for each question and the corresponding code
4. Save it as a document, and submit using the **Assessment** tab
5. Tap the **Submit** button (this will present you with three choices)
6. Attach three files, and then click **Submit**

Note: Be sure to include screenshots of the output

DATA AND
ARTIFICIAL INTELLIGENCE

Thank You

simplilearn