Natural Language Processing - Aspect-based or Topic-based Extraction of Text

Ritu Verma: University of Lucknow

February 8, 2022

1 Introduction of NLP

Natural language processing strives to build machines that understand and respond to text or voice dataand respond with text or speech of their ownin much the same way humans do. One of the most common goals with NLP is to analyze text and extract insights of the given data.

2 Technique

2.1 Introduction of NLP Spacy

Spacy is an open-source software library for advanced NLP, Spacy supports deep learning workflows that allow connecting statistical models trained by popular machine learning libraries like TensorFlow, PyTorch or MXNet through its own machine learning library Thinc.

2.2 Introduction of NLP TextBlob

TextBlob is a library that offers sentiment analysis out of the box. It has a bag-of-words approach, meaning that it has a list of words that have a sentiment score attached to them. It is also able to pick up modifiers and intensifiers that affect the sentiment score.

2.2.1 NaiveBayesClassifier Algorithm

NaiveBayesClassifier using a very simple syntax thats easy for use to improve our aspect-based text extraction.It is based on the Bayes Theorem for calculating probabilities and conditional probabilities

2.3 Methodology

We will be using spacy, a natural language processing library in Python along with Textblob which offers simple tools for analysis and text processing.
Our first goal is to split our sentences in a way so that we have the target aspects and their sentiment descriptions.
For each token inside our sentences, we can see the dependency and child tokens, so that were able to pick up intensifiers.
As for the sentiment analysis part, ideally, we want to label a lot of data so that we can create more advanced classifiers with a higher amount of accuracy. If we dont have pre-labeled data, We can create an initial analysis using a simple tool like TextBlob, and then instead of deciding on the sentiment for each phrase, we can choose if you agree with TextBlob or not which is a lot quicker than deciding the sentiment from scratch.

2.3.1 Topic Modeling and Latent Dirichlet Allocation (LDA)

Topic modeling is a type of statistical modeling for discovering the abstract topics that occur in a collection of documents. Latent Dirichlet Allocation (LDA) is an example of topic model and is used to classify text in a document to a particular topic. It builds a topic per document model and words per topic model, modeled as Dirichlet distributions.

3 Code Available

$https://github.com/rituvermaCS/NLP-Aspect-based-or-Topic-based-Extraction-of-Text$

4 References:

$https://www.ibm.com/cloud/learn/natural-language-processing$
https://en.wikipedia.org/wiki/SpaCy
$https://textblob.readthedocs.io/en/dev/$https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/
$https://towardsdatascience.com/topic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24$