Week 2: Vectorizers and Sentiment Analysis

Topics covered in the week

- Introduction to Word representation
- Bag of word
- Countvectorizer
- TF-IDF Vectorizer
- Sentiment Analysis
- Text Classification using Machine learning

Session Agenda

- Steps to do after text pre processing
- Bag of words Representation
- TF-IDF Vector
- N-gram
- Sentiment analysis
- Text Classification using ML
- Case Study

Word Representation using Bag of word

- Machines cannot directly process text data in raw form.
- They are required to break down the text into a numerical format that’s easily readable by the machine.
- A bag of words is a representation of text that describes the occurrence of words within a document.
- The model is only concerned with whether known words occur in the document, not where in the document.
- We use the tokenized words for each observation and find out the frequency of each token.

Word Representation using TF-IDF vectorizer

- Words like “And”, “For”, “The”, “a” etc. appears many times in a corpus and their large counts will not be very meaningful in the encoded vectors.
- TF-IDF captures how important a word is to the document (without looking at other documents in the dataset).
- TF-IDF (term frequency inverse document frequency) is a scheme to weight individual tokens.
- One of the advantages of TF-IDF is to reduce the impact of tokens that occur very frequently, hence offering little to nothing in terms of information.

N-gram:

- It’s a sequence of N-words.
- Bi-gram is a special case of N-grams where we consider only the sequence of two words.
- In N-gram models we calculate the probability of Nth words given the sequence of N-1 words.
- We do this by calculating the relative frequency of the sequence occurring in the text corpus.

Text Classification using Machine Learning

- Text classification is a machine learning approach that assigns a set of predefined tags to open-ended text.
- We can assign a category to different forms of text i.e. Articles, news, paragraphs, books, web pages etc.

Applications of Text classification:

- Text classification is one of the fundamental tasks in natural language processing with broad applications such as sentiment analysis, topic labeling, spam detection, intent detection etc.
- Text based analysis in Marketing, Product Management.

Sentiment Analysis/Classification

- Sentiment classification is described as automating the process of identifying opinions in text and labeling them as positive, negative, neutral or in different category, based on the emotions customers express at different platforms.
- Natural language processing (NLP) and machine learning (ML) approach are combined to assign a sentiment scores to the topics, categories, entities within a phrase, sentence.
- The key feature of sentiment analysis is to analyze a body of text for understanding the opinion expressed by it.
- Used extensively to get overall end user feedback & gain insights on areas of improvement in product management.

Sentiment analysis tool: VADER & TEXTBLOB

- VADER (Valence Aware Dictionary for Sentiment Reasoning) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media.
- VADER uses a combination of sentiment lexicon which is a list of lexical features (e.g. words) which are generally labelled according to their semantic orientation as either positive or negative.
- VADER is a model used for text sentiment analysis that is sensitive to both polarity (positive/negative) and intensity (strength) of emotion.
- It is available in the NLTK package and can be applied directly to unlabeled text data.

Textblob

- The sentiment function of textblob returns two properties, polarity and subjectivity.

- Polarity is a float which lies in the range of [-1,1] where 1 means positive statement and -1 means a negative statement.

- Subjective sentences generally refer to personal opinion, emotion or judgment whereas objective refers to factual information.

- Subjectivity is also a float which lies in the range of [0,1].

Questions?

Thank You.
Happy Learning!