Generating summary for product Reviews

A data scientist approach

Introduction:

- With the growing information on web, online product reviews are becoming a significant information resource for Internet users. However, online users post thousands of reviews on daily basis even for a single product and it is hard for users to manually summarize the reviews. Online Product review mining and summarization is one of the challenging tasks in natural language processing. Therefore, an automatic approach is desirable to summarize the lengthy product reviews, and it will allow users to quickly recognize the positive and negative aspects of a product.

  “ Our main aim is to develop an automatic system that takes all the positive and negative reviews for a particular product and generates an overall summary of a few lines that is easy to read and make final decision regarding the overall user experience of product . “

Why we need to address this issue

1) **Helping User Make an informed decision**
   With the growing availability and popularity of opinion-rich resources such as review forums for the product sold online, choosing the right product from a large number of products have become difficult for the user. For trendy product, the number of customers' opinions available can be in the thousands. It becomes hard for the customers to read all the reviews and if he reads only a few of those reviews, then he may get a biased view about the product.

2) **Saving time to generalize the overall user experience for a particular product:**
   Makers of the products may also feel difficult to maintain, keep track and understand the customers' views for the products.

Methodology:

Dataset used: Amazon_Consumer_Reviews_of_Amazon_Products dataset available on Kaggle

The dataset consists of 5000 consumer reviews records and 24 variables,

This is an unsupervised machine learning task for summary generation.

Model Used: NLTK, TF-IDF Algorithm

IDE : anaconda (jupyter notebooks)

1) Data Cleaning:

1. remove duplicate and missing values
2. Convert everything to lowercase
3. Eliminate punctuations and special characters
4. Remove stopwords
5. Lemmatize/ Stem

Other optional Steps:

1. Remove HTML tags
2. Contraction mapping: removing words or combinations of words that are shortened by dropping letters and replaced by an apostrophe
3. Remove (‘s)
4. Remove any text inside the parenthesis ( )
5. Remove short words ...words less than certain length (ex word>=3..the, be, as , an..)
6. Remove Numbers

2) Data Modelling

1. Tokenize the sentences We'll tokenize the sentences here instead of words. And we'll give weight to these sentences.
2. Create the Frequency matrix of the words in each sentence. We calculate the frequency of words in each sentence. Here, each sentence is the key and the value is a dictionary of word frequency.
3. Calculate TermFrequency and generate a matrix We'll find the TermFrequency for each word in a product review. TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document)
4. Creating a table for idf matrix IDF(t) = log_e(Total number of documents / Number of documents with term t in it)

Term frequency is how common a word is, inverse document frequency (IDF) is how unique or rare a word is.

5. Calculate TF-IDF and generate a matrix we are multiplying the values from both the matrix and generating new matrix.

6.. Score the sentences add tf-idf score of each word in a sentence.

7. Generate the summary Select a sentence for summary if the sentence score is more than the average score. For the threshold, we've used 1.3x of the average score

Conclusion

we mounted a study to generalise the thousands of reviews for a particular product in a few lines of summary.

“Online customer reviews is considered as a significant informative resource which is useful for both potential customers and product manufacturers. In web pages, the reviews are written in natural language and are unstructured-free-texts scheme. The task of manually scanning through large amounts of review one by one is computational burden and is not practically implemented with respect to businesses and customer perspectives. Therefore it is more efficient to automatically process the various reviews and provide the necessary information in a suitable form.This Data science approach addresses this problem and gives a convenient solution for summary generation.