Natural Language Processing (NLP)
Assignment-2

Topic Analysis of Review Data

Objective: Help a leading mobile brand understand the voice of the customer by analyzing the reviews of their product on Amazon and the topics that customers are talking about. You will perform topic modeling on specific parts of speech. You'll finally interpret the emerging topics.

Problem Statement:
A popular mobile phone brand, Lenovo has launched their budget smart phone in the Indian market. The client wants to understand the VOC (voice of the customer) on the product. This will be useful to not just evaluate the current product, but to also get some direction for developing the product pipeline. The client is particularly interested in the different aspects that customers care about. Product reviews by customers on a leading e-commerce site should provide a good view.

Domain: Amazon reviews for a leading phone brand

Analysis to be done: POS tagging, topic modeling using LDA, and topic interpretation

Content:
Dataset: 'K8 Reviews v0.2.csv'
Columns:
Sentiment: The sentiment against the review (4,5 star reviews are positive, 1,2 are negative)
Reviews: The main text of the review

Steps to perform:
Discover the topics in the reviews and present it to business in a consumable format. Employ techniques in syntactic processing and topic modeling.
Perform specific cleanup, POS tagging, and restricting to relevant POS tags, then, perform topic modeling using LDA. Finally, give business-friendly names to the topics and make a table for business.

Tasks:
1. Read the .csv file using Pandas. Take a look at the top few records.
2. Normalize casings for the review text and extract the text into a list for easier manipulation.
3. Tokenize the reviews using NLTKs word_tokenize function.
4. Perform parts-of-speech tagging on each sentence using the NLTK POS tagger.
5. For the topic model, we should want to include only nouns.
    a. Find out all the POS tags that correspond to nouns.
    b. Limit the data to only terms with these tags.
6. Lemmatize.
    a. Different forms of the terms need to be treated as one.

    b. No need to provide POS tag to lemmatizer for now.
7. Remove stopwords and punctuation (if there are any).
8. Create a topic model using LDA on the cleaned up data with 12 topics.
    a. Print out the top terms for each topic.
9. Analyze the topics through the business lens.
    a. Determine which of the topics can be combined.
10. Create topic model using LDA with what you think is the optimal number of topics
11. The business should be able to interpret the topics.
    a. Name each of the identified topics.
    b. Create a table with the topic name and the top 10 terms in each to present to the business.

12. Plot using pyLdavis and using Facetgrid