Keyword Extraction for Ecommerce Review Data

- ## Goal

  So the main goal is to extract the **meaningful keywords** from the ecommerce review data so that it can be used in our **Customer Experience Platform** to understand the **Customer Experience** in a better way.

- ## Dataset

  The dataset is about the customer reviews which is taken from an ecommerce website. The dataset contains 12 columns and 23k rows.The types of columns that it contains are **Clothing ID, Age, Title, Review Text, Rating, Recommended IND, Positive Feedback Count, Division Name, Department Name, Class Name.** For this assignment we will be only considering 1024 rows which is from **Clothing ID ==1078 .** We will be mainly focusing on two columns **Title** and **Review Text**.

  Dataset link:-

  https://www.kaggle.com/nicapotato/womens-ecommerce-clothing-reviews/version/1?select=Womens+Clothing+E- Commerce+Reviews.csv

- ## Assumptions

  1. Given that the nouns and adjectives within the dataset are indicative of the products being reviewed, we can assume that these words play a crucial role in describing and evaluating the reviewed products.
  2. Considering review text which has Clothing ID == 1078.

- ## Procedure

  The steps are as follows:-

  1. Import Libraries
  2. Load Data using Pandas
  3. Select rows which has a **Clothing ID == 1078**
  4. Removing Un-necessary column from dataframe
  5. Handling **NaN** values in the dataframe
  6. Converting all values to **lowercase**
  7. Handling **apostrophes** values in the dataset
  8. Remove **Extra Space** from the dataset
  9. **Punctuation And Emojis** removal
  10. **Stopword** removal

11. Understanding data by forming word clouds**(Unigram, Bigram, Trigram)**
12. Understanding POS patterns in the data**(Unigram, Bigram, Trigram)**
13. Approach-1:- Applying Heuristic approach related to **POS tags** used by the words
14. Approach-2:- Applying Heuristic approach related to **dependency parsing** between words.

- ## Throught Process

So as you know our main task here is to extract meaningful features i.e to extract keywords which will have some context in to it. So everything here is related to words and their arrangement and what context are they delivering. So I followed two approaches, both are heuristic approaches and these approaches are in following ways:-

1. My initial approach revolves around the concept of POS, an acronym for Parts Of Speech. This concept serves as a cornerstone for defining the grammatical structure of English text, enabling us to categorize words into various classes such as Verb, Noun, Pronoun, Adjective, Conjunction, Adverb, and more. Our primary objective is to uncover diverse POS patterns that yield valuable insights.

   To achieve this, I employed a data-driven technique, starting with the creation of word clouds (both Bigram and Trigram) after removing stopwords from the dataset. By eliminating stopwords, we ensure that our analysis focuses on words that carry more substantial meaning.

   During this process, I identified several common patterns that emerged:

   Noun-Adjective (NOUN ADJ)

   This combination is often used to describe the qualities or characteristics of products or features. Here are some of the examples for Noun-Adjective pattern "Great camera," "poor battery life," "comfortable fit".

   Noun-Verb (NOUN VERB)

   These combinations often indicate actions or behaviors of products or features. Here are some of the examples for Noun-Verb pattern "The phone performs well," "the software crashed," "the camera captures".

   Adjective-Noun-Noun(ADJ NOUN NOUN)

   It creates a descriptive phrases and convey more specific information about the nouns being discussed. It helps provide a clearer and more vivid picture of the objects or concepts being described. Here are some of the examples for Adjective-Noun-Noun pattern “I bought a comfortable leather sofa”.

Adjective-Adjective-Noun(ADJ ADJ NOUN)

Describes multiple qualities of a product or service.Here are some of the examples for Adjective-Adjective-Noun pattern “Durable and lightweight design” “quick and responsive service”.

Noun-Verb-Adjective (NOUN VERB ADJ)

Describes actions and their qualities. Here are some of the examples for Noun-Verb-Adjective pattern "The app runs smoothly," "the phone feels sturdy," "the software is buggy"

Noun-Adjective-Noun (NOUN ADJ NOUN)

Specifies attributes of nouns. Here are some of the examples for Noun-Adjective-Noun pattern "Camera with high resolution," "phone with excellent battery life".

Adjective-Preposition-Noun (ADJ)

Expresses feelings or sentiments toward nouns. Here are some of the examples for Adjective-Prepositon-Noun pattern "Satisfied with the product," "disappointed in the service".

Noun-Verb-Preposition-Noun(NOUN VERB ADP NOUN)

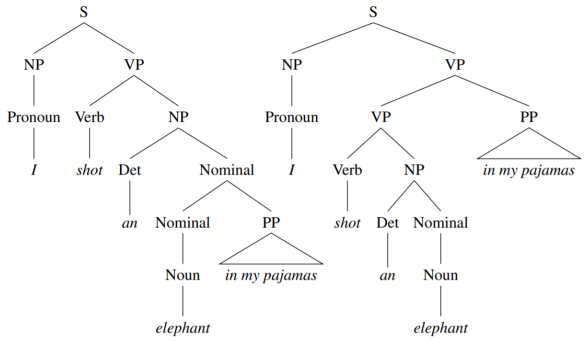
Shows how products or features contribute to specific purposes.Here are some of the examples for Noun-Verb-Preposition-Noun pattern "The device helps with productivity," "the software is essential for work".

Noun-Verb-Verb-Adjective (NOUN VERB VERB ADJ)

In this pattern, the noun serves as the subject of the sentence, followed by two verbs that indicate actions or behaviors related to that noun, and finally, an adjective that describes a quality or attribute of the noun. Here are some of the examples for Noun-Verb-Verb-Adjective patterns "The technician skillfully repaired the faulty television".

Following the discovery of these patterns, they were systematically compared within the text dataset to extract valuable keyword values. This comparison process allowed us to pinpoint and extract keywords that carry significance, enriching our dataset with meaningful information.

2. The second method pertains to leveraging dependency parsing tree techniques, unveiling the intricate relationships between words. I harnessed the power of the Spacy library to construct these enlightening sentence structures. Below, you'll find an illustrative depiction of this tree formation.



In this context, we adhere to a fundamental principle of dependency parsing: a root node serves as the foundation, branching into child nodes to construct the tree. My strategy involves commencing at the root node and systematically traversing the entire structure. Whenever I encounter a noun, I merge it with its associated children to craft a purposeful keyword. While this approach excels in extracting meaningful keywords, it's important to note that it may not offer a one-size-fits-all solution, as certain exceptional cases could present challenges.

- **Drawback for above approaches**

1. In approach-1 we have to find all the possible POS combinations to make this algorithm work by covering all the edge cases and creating POS patterns, now this can take time because also you need to have dataset knowledge to form patterns. In approach-2 again there can be same problem where you need to have good understanding of english language to create some out of it, it is also a time taking process. Although both the approaches can give you good results to some extent but if you want to improve the model, we can try to go for Deep learning or Machine Learning models.

2. In Approach-1 and Approach-2 , I have taken some assumptions like noun values will be all related to product over which customer review is done and adjective value will be related to status of the product. If we don't take these assumptions then it will be difficult to use these approaches because noun can be person name also or place name or many things, so there it might cause problem.

- ## What's Next?

  We have the option to explore alternative approaches to keyword extraction. Some promising avenues include:
  Keyword Extraction Libraries: Utilizing existing libraries such as Rake_NLTK, YAKE, KeyBert, or leveraging APIs like MonkeyLearn and Textrazor. These tools offer pre-built solutions for keyword extraction and can provide valuable insights.

  Word Embedding Comparisons: Comparing word embeddings and selecting words with high similarity scores for keyword formation. This method relies on semantic similarity to identify relevant keywords.

  Deep Learning Models: Training deep learning models specifically designed for keyword extraction. These models can potentially capture complex patterns and context in the data to yield meaningful keywords.

  Exploring these alternative approaches can help us refine our keyword extraction process and uncover more valuable insights from the dataset.