Report of the Baseline Model : SalesMan for Webpage

Introduction : This baseline model finds the accuracy and precision of the model that is implemented based on the query user asks related to a particular feature of the mobile phone. The model uses two techniques : Naïve Bais Text Classification algorithm and Cosine similarity.

Methodology : The baseline model is implemented using python. These are the following steps involved in the implementation of the project.

1.) Collection of Train and test data:-

We have collected training data and test data for the project . Training data was collected from sources like kaggle and several websites from the google . We collected around 750 question asked about different features of mobile phone like battery , camera , processor, storage and many more. Some questions are similar but they are asked in a different way and all these questions will categorize to one category or one feature of the mobile phone.

Test data was collected from e commerce websites like Amazon and Flipkart . It consists of questions that users ask in the "question and answer" section of the particular product. We decided to collect those questions from ecommerce websites because the questions that are present there are asked in simple human language and therefore it will greatly benefit us in testing our training data.

2.) Data Preprocessing :

First we generated all the possible tokens from the user query then we converted all the tokens to the lowercase followed by removing the stopwords and that is how data processing was done.

3.) Distribution of data :

Distributed the data into train dataset and test dataset and implemented the Naïve bayes text classification algorithm on the dataset.

4.) Model applied on data:

We applied naïve text classification over our train and test data.

Naïve Bais Text Classification :

1.) Naive Bayes is a probabilistic algorithm used for classification, particularly in natural language processing and text classification.

2.) Naive Bayes assumes that each feature (or word) in the input is independent of every other feature, which makes it a "naive" algorithm.

3.) In text classification, each document is represented as a bag of words (or a set of word frequencies), and the Naive Bayes algorithm calculates the probability of each document belonging to a particular class.

4.) The algorithm works by calculating the prior probability of each class (i.e., the proportion of documents in the training set that belong to each class) and the conditional probability of each word given each class.

5.) To classify a new document, the algorithm calculates the probability of the document belonging to each class using Bayes' theorem, which takes into account both the prior probability and the conditional probability.

6.)Algorithm then assigns the document to the class with the highest probability. Also we have tried to match cosine similarity to match user query similarity with the word in our corpus.

Cosine Similarity:

1.) Cosine similarity is a measure of similarity between two non-zero vectors, typically used to compare documents or other text-based data.

2.) It calculates the cosine of the angle between the two vectors, which ranges from 0 (no similarity) to 1 (perfect similarity).

3.) In text-based applications, the vectors typically represent the frequency of words in the documents.

4.) To calculate cosine similarity, the dot product of the two vectors is divided by the product of their magnitudes.

5.) Cosine similarity is commonly used in information retrieval, text mining, and recommendation systems to identify similar items or documents.

Results : We got the following results after running the model on our train and test data.

Accuracy of model: 0.8838

Precision of model: 0.88

Conclusion : We got the the above mentioned accuracy and precision for the baseline model and in the coming days we will be going forward with our project and implementing all the other things we have planned like making an user interactive IR system which will give the relevant result for the query that user will ask for any particular product.