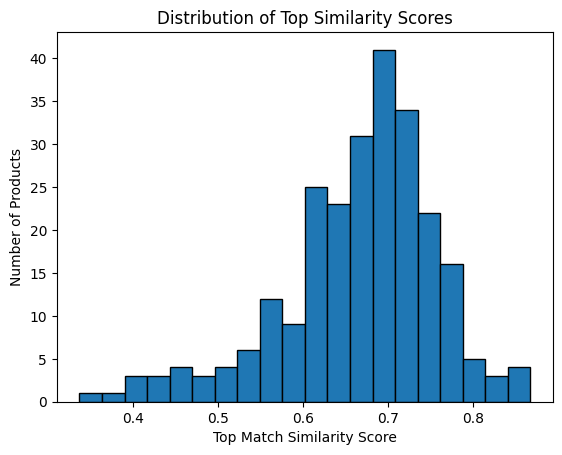
**Product Catalog Matching System Project**

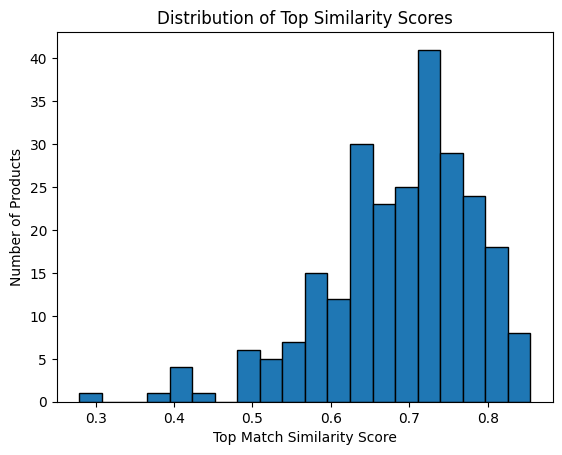
**Approach**

* **Data Preview / Observations :** Observe some samples of the given data in order to understand the typology of the request and think of the appropriate approach.
* **Data Curation :** The process of cleaning the data and bring them to the desired format to work with included several steps:
  1. Filter out filler words that do not represent any attribute but are parts of everyday speech.
  2. Normalize sizes in descriptions (e.g. extra small → xs)
  3. Create a dictionary using all words from the catalogs
  4. Compare every word from each description with the words from the dictionary and correct any typos / misspellings (this comparison is done using the Levenshtein distance which computes the letter-by-letter distance of words)
  5. Compare every word from each (typo-corrected) description with the words from the dictionary. Substitute the word with its match if the cosine similarity is over 0.75. (For words already existing in the dataset of course their best match will be themselves).
  6. Extract specific attributes (color, brand, size etc.) from the free-text descriptions.
* **Compare descriptions with catalog items:** After curating the descriptions, create embeddings of each description and each catalog item and for each description calculate the top-3 matches (based on cosine similarity score).
* **Model Evaluation :** All contextual sentence comparisons where made using 3 different models from the Sentence Transformers package :   
  **all-MiniLM-L6-v2,** **all-mpnet-base-v2** and **BAAI/bge-large-en-v1.5**For each model I created the distribution of the scores that correspond to each description’s top match (see distributions below)

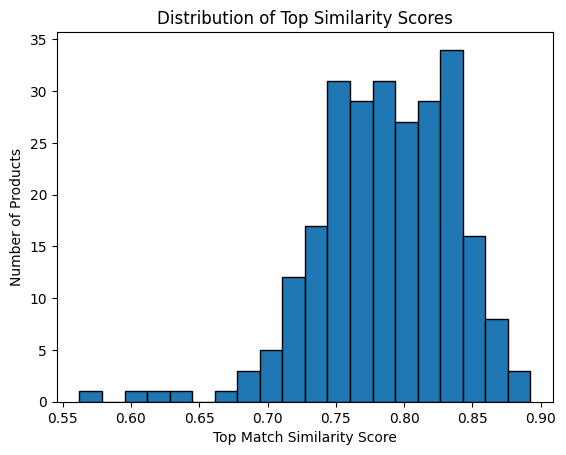
all-MiniLM-L6-v2



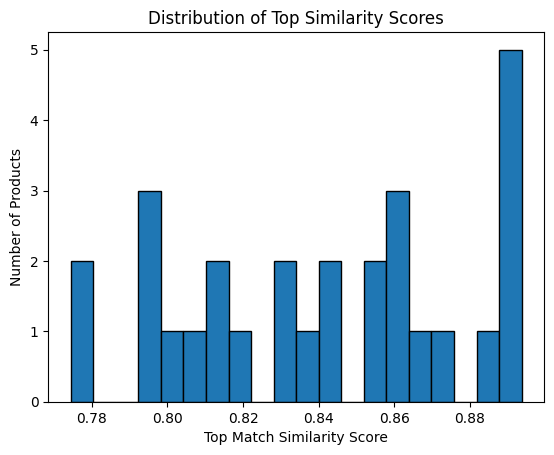
all-mpnet-base-v2



BAAI/bge-large-en-v1.5



Test Dataset:



Total Descriptions: 28  
Correct Matches: 26  
Accuracy: 92.86%  
  
Total Descriptions: 28  
Correct Top-3 Matches: 27  
Recall (Top-3 Accuracy): 96.43%