Assignment: Finding Similar Product in E-commerce

Marks: 100 Points

Due on: 18th Sept, 2025

Overview

In this assignment, you would LSH to find similar products in e-commerce sites. You will build a software that allows one to upload product information. Given a product, find similar other products using various similarity functions.

Instructions

- The assignment has to be done in groups of 3-4.
- You can use any programming language.
- Deliverables
  - Part A: Single zip file that contains all source code. Don't include any dataset or library in the zip file. Name it as GroupXY.zip (E.g. Group05.zip)
  - Part B: Report based on the assignment. It should be named as GroupXY.pdf (E.g. Group05.pdf)
  - Please make sure to submit **ONLY** according to the instructions given.

We may use software to detect plagiarism. If we find cases of copying, then all those who are involved will be given either FR grade or 0 marks for the assignment. Please don't share your code or report with anyone.

Dataset Description

You will use an Amazon dataset from here. To keep the dataset set size small we will use the **Appliances** data from Amazon Review Data (2018). Download the metadata for **Appliances** category, which has information about 30,459 products. The site also has code to help you with loading the data.

https://amazon-reviews-2023.github.io/

https://stackoverflow.com/questions/753052/strip-html-from-strings-in-python?noredirect=1&lq=1

https://stackoverflow.com/questions/9662346/python-code-to-remove-html-tags-from-a-string

Exercises

[20 Points] Exercise 1: You can use any existing open source e-commerce software or your own simple software to list all the products from the above given dataset.

In the report, include 2-3 screenshots showing how the products are shown by your software.

[40 Points] Exercise 2: Add a new UI component to your software that allows users to search for similar products using the following three text similarity functions:

1. Products having similar title (PST)
2. Products having similar description (PSD)
3. Products having similar title and description (PSTD). You can hybrid the title and description in your own preferred way.

For any product, the user should be able to select one of the three similarity functions: PST, PSD or PSTD. The system should show top-5 or top-10 similar products in decreasing order of similarity.

Consider any product that has non-empty similar products (attribute “similar_item”) in the dataset. Preferably consider a product with at least 5-10 given similar products. In the report,

include 3 screenshots showing the similar products shown by your software for the product. Also discuss how the similar products obtained using the three methods compare with the similar products given in the dataset. You can report precision@k (k=10) between the given set of similar products and the one obtained using your algorithms.

[40 Points] Exercise 3: In this exercise you will perform some basic evaluation of your algorithms and also study the effect of hyperparameters.

Evaluation Set: Top-100 products that have the highest number of similar products given in the dataset. In the report, mention the max and min given similar item set size for the products in the evaluation set.

Evaluation Metric: Use precision@k between the obtained similar items and the given set of similar items.

In the report, present your results in the form of tables or graphs for the following hyperparameters. You need to report MAP@10 (Mean Average Precision for all the 100 products). When you vary one of the parameters, you need to fix the other parameters based on your own choice. Ideally, the other parameters should be the one that you think gives optimal performance.

1. K-character shingles. Consider K=2, 3, 5, 7, 10 characters.
2. Number of permutation/hash functions during Minhashing phase. Consider 10, 20, 50, 100, 150 hash functions.
3. Parameters b & r in the LSH phase.

You can present the result for product title (PST) and product description (PSD) similarity separately. No need to consider the hybrid of title and description (PSTD) for Exercise 3.