Project 3: Mining E-Commerce Customer Reviews

Version: 0.1

In this project, you are expected to perform various document modeling techniques to model an E-Commerce customer review dataset for the purpose of information retrieval. You are also required to devise a research question of your choice based on the dataset given and solve it.

Outline

- Evaluation
- Dataset
- Questions
    - Question 1: Preprocess the corpus of customer reviews dataset
    - Question 2: Remove stopwords, standardize tokens
    - Question 3: Build the Term-Frequency Inverse-Document-Frequency (TF-IDF) matrix and apply the Latent Semantic Analysis (LSA) method
    - Question 4: Compare the performance of Information Retrieval (IR) using both TF-IDF and LSA methods
    - Question 5: Create and solve a problem of your choice
- Report
- References

Evaluation

Your score depends upon the correctness of implementation and how you analyze your results.

- Question 1-5 (10 Marks)
- Question 6 (5 Marks)
- **No more than two-page project report** (5 Marks): It should be prepared using the latex format distributed before. No other format is allowed. See the report section, before you start

Dataset

The dataset provided consists of the information regarding the Women's E-Commerce Clothing Reviews and Ratings in CSV format. The Dataset consists of the following columns:

COLUMN NAME	**DESCRIPTION**
Clothing ID	- Unique ID of a product
Age	- Age of the reviewer
Title	- Title of the review
Review Text	- Customer review in text format
Rating	- Product rating by reviewer
Recommended IND	- Whether the product is recommended or not by the reviewer

Positive Feedback Count	- Number of positive feedback on the review
Division Name	- Name of the division product is in
Department Name	- Name of the department product is in
Class Name	- Type of product

Questions

Question 1: Preprocess the corpus of customer reviews dataset

- It is recommended to select a common cloth-type (i.e. Clothing ID) and work on its records for your project to speed-up execution.
- Tokenize the corpus of customer reviews, build a dictionary

Question 2: Remove stopwords, standardize tokens

- Come up with a list of stop words based on the reviews or use a standard stop-words list available online
- Standardize word tokens by using stemming and lemmatization, see, e.g., NLTK (Natural Language Processing Kit)

Question 3: Build the Term-Frequency Inverse-Document-Frequency (TF-IDF) matrix and apply the Latent Semantic Analysis (LSA) method

- Perform LSA using Singular Value Decomposition (SVD). Consider the TF matrix for SVD. You can also perform SVD on the TF-IDF matrix.
- Plot documents in the LSA/TF-IDF space
- Use the Python packages sklearn (see the Jupiter notebook we discussed in the class) or Gensim

Question 4: Compare the performance of Information Retrieval (IR) using both TF-IDF and LSA methods

Solve the IR problem for the query of your choice to retrieve important documents. You can use cosine similarity to measure the similarity between each review (document) vector and the query vector in the TF-IDF/LSA space.

Question 5: Create and solve an interesting problem for this dataset

Form an interesting research problem for this dataset and solve it bases on the methods you have learned in the course

Report

It's worth reading this: http://hallertau.cs.gsu.edu/~mweeks/project.html

- You are required to explain the things you learned from this project (primarily from Questions 4 and 5) in clear English sentences/paragraphs
- You may include no more than 2-3 important plots or tables that you have created for this project. Each figure and table should be formatted well and provided with a proper caption.
- The report should not exceed more than two pages
- It should be prepared using the latex format distributed before. No other format (e.g. Jupyter notebook) is allowed. You must upload a PDF report only.

Submission Requirements

- You must upload your python source as a zip file and show your results to the course TAs.
- Upload your project report as a separate PDF file. It should not be included in the source code zip file.

References

1. Women's E-Commerce Clothing Reviews Dataset
2. Sample Code
3. Stemming and Lemmatization
4. Natural Language Toolkit (NLTK)
5. LSA