Recommendation System from word embeddings

Burak Han Bakir

Yacine Berbiche

# Problem description

Design a recommendation system that leverages Natural Language Processing (NLP) techniques to provide personalized suggestions to users based on their textual preferences.

A recommendation model based on Word2Vec can benefit various applications and industries, specifically in e-commerce platforms, platforms delivering articles, blog posts, or other textual content, job portals, social media networks, educational platforms, customer support chatbots, financial Services, real-time personalization. The applications of Word2Vec-based recommendation models are diverse and can be applied wherever understanding the semantic relationships between words or documents is crucial for making accurate and relevant suggestions. A recommendation model based on Word2Vec is interesting because it harnesses the power of semantic understanding, contextual similarity, and language nuances to deliver personalized and relevant recommendations across various domains, ultimately enhancing the user experience and system performance.

# Data

Dataset: Online Retail II - This Online Retail II data set contains all the transactions occurring for a UK-based and registered, non-store online retail between 01/12/2009 and 09/12/2011.The company mainly sells unique all-occasion gift-ware. Many customers of the company are wholesalers (Obtained From UC Irvine ML repository).

Dataset Characteristics: Multivariate, Sequential, Time-Series, Text

Dataset Subject Area: Business

The description of the fields in this dataset:

1. InvoiceNo: Invoice number. a unique number assigned to each transaction
2. StockCode: Product/item code. a unique number assigned to each distinct product
3. Description: Product description
4. Quantity: The quantities of each product per transaction
5. InvoiceDate: Invoice Date and time. The day and time when each transaction was generated
6. CustomerID: Customer number. a unique number assigned to each customer

The dataset contains 5,25461 transactions and purchase history of 4383 customers. This can be used to create a system that automatically recommends a certain number of products to the consumers on an E-commerce website based on the past purchase behavior of the consumers.

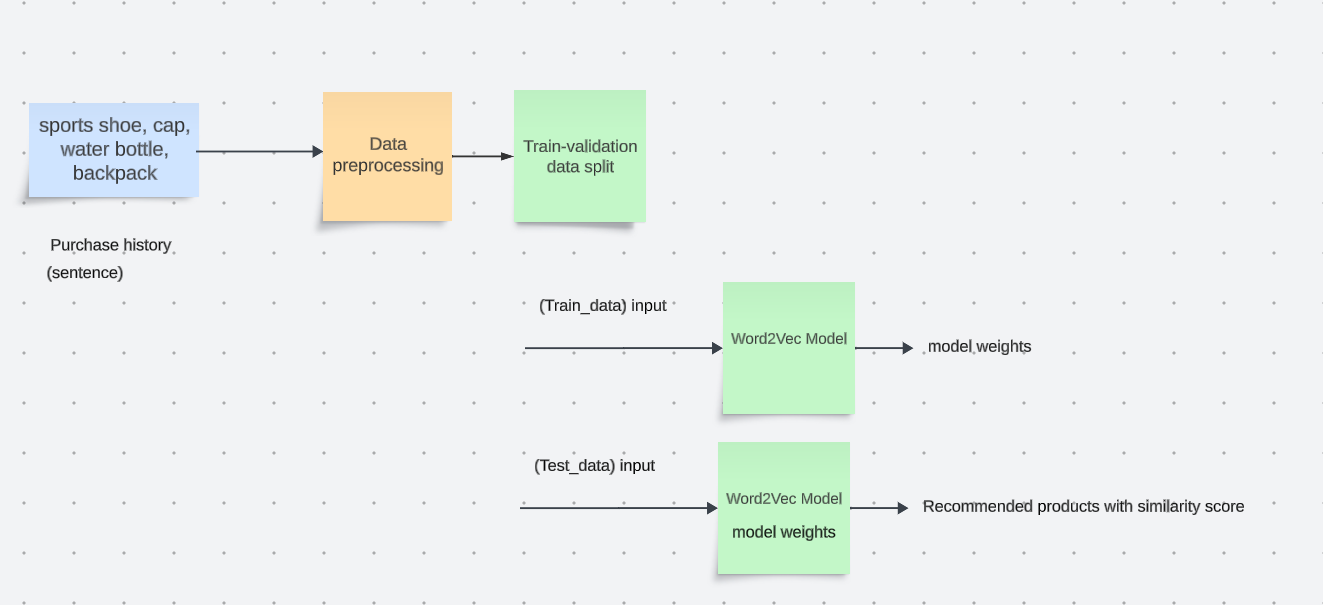
# Solution

Most of the time there is a pattern in the buying behavior of the consumers. Each of these products are represented by a vector, and similar products can be predicted using the Word2Vec model. Similar products can be recommended to the customer by using the vector similarity score between the products.

Approach:

The buying history of a consumer is considered as a sentence and the products as its words. A word2vec model is a simple neural network model with a single hidden layer. Its input is a text corpus and its output is a set of vectors: feature vectors that represent words in that corpus. The task of this model is to predict the nearby words for each and every word in a sentence. The Weights learned by the hidden layer of the model can be used as the word embeddings to find similar products for recommendation.

Project Stages:



# Model evaluation and discussion

The Model generated has a vocabulary of 3479 unique words and their vectors of size 100 each. The dimensions of the product embeddings were reduced from 100 to 2 by using the UMAP algorithm for visualization and clear clusters of similar products were visible when plotted. (UMAP is used for dimension reduction based on manifold learning techniques and ideas from topological data analysis.)



Inorder to learn about the model ranking quality, the test data was cleaned by removing the customers with less than 50 purchases. And then, the first 40 purchases of each resultant customer was used to predict the next 10 items purchased, which is already known in the test set. Out of the 10 recommended products, the relevant products were found out using set intersection of the recommended products and the overall relevant product list generated from the known next 10 products which were actually ordered. This was done by finding 5 most similar item vectors from the model for each product in the 10 actual ordered items. When evaluating the recommender system model offline, the following set of metrics were observed:

* Precision: Measures the proportion of relevant recommendations out of all the recommended items.
* Recall: Measures the proportion of relevant recommendations out of all the relevant items.
* Normalized Discounted Cumulative Gain (NDCG) :Evaluates the ranking quality by assigning higher importance to relevant items appearing at the top of the recommendation list.

After evaluation, it was found that 17 of the customers among 206, who have greater than 50 purchases, are served with recommendations with precision > 0.5 and the precision of the model lies between 0.0 and 0.9, which can result in increase in the revenue. Recall of the model lies between 0.0 and 0.18 which shows moderate performance. NDCG values of the model lie between 0.0 and 0.33 and 36 customers obtained positionally correct recommendations with NDCG > 0.2 which also implements a performing model.

# Summary

The goal of a Word2Vec-based recommendation system is to enhance the accuracy and relevance of recommendations by leveraging semantic relationships between words in a given dataset. In the context of recommendation systems, Word2Vec can be applied to represent items (such as movies, products, or articles) as vectors based on the words associated with them in user reviews, descriptions, or other textual data. Moreover, this modeling can capture semantic relationships, address the cold start problem, understand user preferences, incorporate contextual understanding, and generate meaningful item embeddings. That being said, there are few limitations to the model. Word2Vec models may not capture complex contextual relationships well. Also, Word2Vec may struggle to disambiguate between different meanings of a word, as it represents each word with a single vector. If the dataset is not representative or lacks diversity, the model may not capture the full range of semantic relationships, leading to biased or less accurate recommendations. Also, the quality of the word embeddings can be sensitive to text preprocessing techniques. Moreover, rare words or items with limited occurrences in the dataset may not have well-defined representations in the Word2Vec model. The model is limited in its capability to scale, and if done, it can be computationally expensive and time-consuming. Despite these limitations, Word2Vec remains a powerful tool for capturing semantic relationships in textual data. Addressing these challenges often involves combining Word2Vec with other techniques or using more advanced models that can handle some of these issues, such as contextual embeddings or attention mechanisms.