Data Mining - Homework 3 - Problem 1: Report on Clustering and Feature Engineering

Olga Sorokoletova
1937430
Sapienza Universitá di Roma

December 5, 2021

Abstract

Given the Steam Reviews Dataset 2021 on Kaggle, I performed clustering with k-means++, where Elbow Method has been used to find the most suitable number of clusters. The aim was to provide a comparative analysis of clustering without and with application of the feature engineering techniques, carefully chosen after data inspection procedure. First, clustering model for the topic extraction has been implemented based on the textual data only. Then, clustering of their entire dataset has been performed.

I. Solution Description

Two Jupyter Notebooks were handed in:

1. `DM_..._text.ipynb`;
2. `DM_..._entire.ipynb`.

It was noted that given dataset has fields of the different data types. Therefore, an **idea behind splitting** overall task of clustering for this dataset into two sub-tasks: Topic Extraction and Global Clustering was that instead of performing «blind» clustering («cluster everything what I have, see what happens and then try to improve») it is better to find an appropriate feature engineering for the most informative and specific field of the dataset – `review`, which is textual, and then apply gained knowledge to the framework for the clustering of the entire dataset.

Each of the Notebooks consists of the **four so-called experimental set-ups**, where the first of them describes clustering based on the raw data and the following three are three different approaches to the feature engineering grown up upon each other:

1. preliminary;
2. preliminary + key(something more «smart»),
3. preliminary + key + reduction(something helpful to reduce time and memory consumption).

Exact choices of the feature engineering techniques and motivation behind them is provided in the following two subsections.

i. Topic Extraction

Topic Extraction task has been formulated as an aim to find the difference in topics of reviews, authors of which recommended or purchased (`recommended` and `steam_purchase` fields, correspondingly) an app, compared to those, who didn't.

***k*-means** has been chosen as a clustering algorithm, because it has shown good results while being used for the Topic Extraction problem. Its `Scikit-Learn` implementation performs grouping of the similar reviews and produces a list of the common words. In fact, its *k*-means++ modification has been used in order to provide the better centroids initialization.

For the text vectorization of the raw data studied previously approach of creating a **TF-IDF** matrix has been employed.

A simple and representative method exploited to validate the number of clusters is the **Elbow Method**.

i.1 Text Preprocessing

Used As: Preliminary step of feature engineering;
Motivation: Lists of top-10 common words (centroids after back transformation from the vector space) for each of the 18 clusters obtained with *k*-means++ clustering based on the raw data are full of stop-words, which means that topic modeling has not really been performed.

NLP preprocessing **outline**:

1. normalization (lowercasing),
2. noise (punctuation, digits, special characters, extra white spaces) removal,
3. lemmatization,
4. stop-words removal.

i.2 Word Embeddings

Used As: Key step of feature engineering;
Motivation: Semantic information coming from the words co-occurrences is relevant and indeed important to capture the context of review.

Global Vectors model is chosen for the embeddings because it takes into consideration both local and global statistic and appears to be the best word embeddings in my experience so far. Embedding vectors are loaded as pre-trained on a large corpus. Dimensionality 100d was found by performing the hyperparameters detection procedure.

i.3 PCA

Used As: Reduction step of feature engineering;
Motivation: Even for a very small portion of 600MB of data (even less after dataset cleaning) first two experimental set-ups take around 20-30 minutes to execute on a local memory and the third set-up interrupts its execution due to the memory over-consumption. Even though re-loading computations on the GPU could help (but personally me had a problem with that due to being blocked and restricted in resources by Colab), it makes clear hint that dimensionality reduction is necessary to work on this dataset (especially if one would like to exploit the full amount of data provided by Steam in Kaggle). However, PCA has been used also in the other experimental set-ups for the visualization purposes.

Principal Component Analysis, which employees idea to find a more meaningful projection, is chosen as well studied during the lectures and powerful method for the dimensionality-reduction.

ii. Global Clustering

Global Clustering task considers studying of the separability of the entire dataset. Now the textual field of review is treated on an equal basis with other feature columns and focus is moved away from the Topic Extraction problem to the problem of capturing the global patterns in data, however the question about ability of the clustering model to be able to distinguish between the users of the platform who inclined and not inclined to recommend or purchase reviewed items is still inside the area of interest.

***k*-means++** learning algorithm remains unchangeable except for the fact that now it's used with a fixed number of clusters, found during the procedure of optimal hyperparameters detection (the same **Elbow Method**, but it's not plotted in the second Notebook) and equal to 20, in all four experimental set-ups.

Wherever is not specified, **TF-IDF** vectorizer is used for the textual fields as default.

ii.1 Feature Drilldown and Normalization

Used As: Preliminary step of feature engineering;
Motivation: Feature Selection – poorly performing or not relevant for a given problem features are better to be dropped to simplify the model training. Feature Invention – smart combinations/variations of the existing features prompts to the model the better learning direction. Data Normalization – transforming data to have similar distributions over feature vectors allows the model to not be misled by scale of the data.

Features Selected to drop:

1. unique identifiers,
2. duplicates,
3. timestamps.

Features Invented to add:

1. the number of words in review,
2. binary flag denoting if author is an experienced user of the platform.

Data Normalization is approached by means of Standardization (less affected by outliers than other popular normalization methods):

$$x' = \frac{x - \bar{x}}{\sigma}$$

ii.2 Word Embeddings

Used As: Key step of feature engineering;
Motivation: The step where the knowledge gained about best feature engineering for the textual field is exploited (combining step: «stack» the best found so far feature engineering for the textual data and the best found so feature engineering for the other fields).

ii.3 PCA

Used As: Reduction step of feature engineering;
Motivation: The reason to seek for a representation of the input features in a lower dimensional space is essentially the same as before, but this time performed for an entire dataset.

II. Q&As

The scatter plots of the clusters in 3D for the all the **four stages of the Global Clustering** subtask are represented by the Figure 1, Figure 2, Figure 3 and Figure 4. For the sake of fast computations reduced dataset consisting of a 5000 randomly sampled rows of dataset has been used to produce plots.

1) The scatter plots of the clusters in 2D for the Global Clustering; 2) Scatter plots of the clusters in 2D and 3D for the Topic Extraction; 3) Heatmaps for the `recommended` and `steam_purchase` slices of interest; 4) Elbow curve plots; 5) Information on clustering running times; 6) Observation on the results obtained – are provided directly in the Notebooks together with useful notes and utilities to re-execute them if desired.

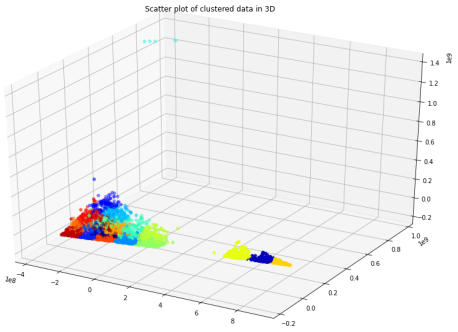

Figure 1: *Without feature engineering.*

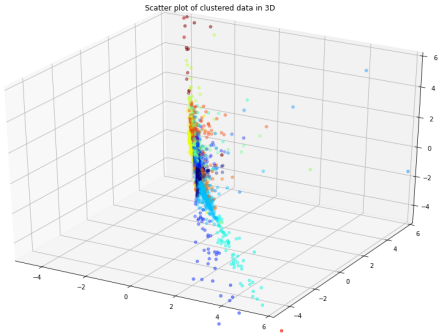

Figure 2: *Feature Drilldown and Data Normalization.*

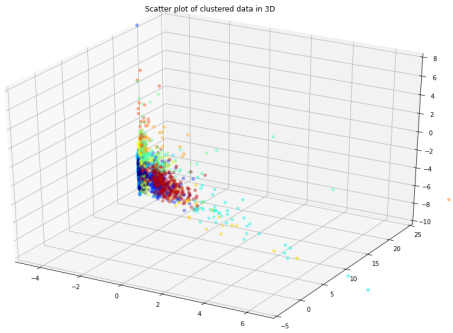

Figure 3: *Word Embeddings of the textual data.*

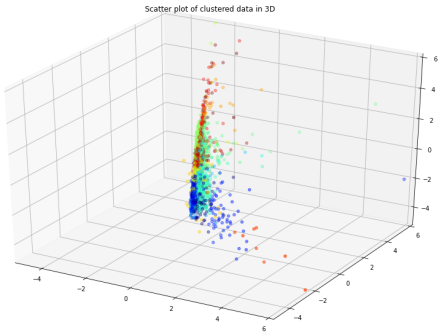

Figure 4: *Dimensionality Reduction with PCA.*