BDC: Homework 2

Amazon Reviews Clustering

Alexandru Melnic 1692625

Preprocessing

The processing (and import) is all executed from the **`TextProcessing`** class. For each text, all the punctuation is removed, the text is made lower case, texts with length less than `min_text_len` are discarded , words with length less than `min_word_len` and words containing at least one number are also discarded. In the final step the lemmatization is performed. Optionally it is also possible to keep only the nouns with the boolean parameter `select_nouns`. The choice to keep only the nouns can be a reasonable one since amazon sells objects and since the expected clusters from this dataset are about different (object) categories. If the purpose was the one of performing a sentiment analysis it would be more appropriate to keep the adjectives instead.

Since the dataset is big it is possible to sample only a `frac_review` of the original texts. This sampling is made by keeping track of the reviews' indices, hashing them with the built-in python `hash()` function and selecting only the rows for which the condition `hash(row_idx) % 100 < frac_review` is verified.

To perform the vectorization of the texts it is used the `TfidfVectorizer` from the sklearn package with the choice of removing the stopwords and with the hyperparameters `max_df = 0.25` and `max_features = 20000`. Furthermore, in the literature, for texts, usually is used the cosine distance, for this reason before the clustering the vectors are normalized (since the cosine distance is proportional to the euclidean distance for normalized vectors) .

Analysis

The analysis is performed by clustering the texts and trying to get from the clusters meaningful information about the categories of the reviews. As first step the clustering is conducted on the dataset without any dimensionality reduction with the *Mini Batch K-Means;* as a second step it is performed an SVD (with 2 components) with a standard *K-Means* clustering and finally , again after SVD, a clustering with *Gaussian Mixtures* (interpreting the means of the gaussians as the representative values for each cluster).

Results

In this section the results of the analysis are presented. Tables containing metrics of the clustering and wordcloud pictures are presented. From the pictures it is possible to understand which are the two reviewed categories: products for **babies** and products for **pets**.

	Mini BatchK-Means	K-Means with SVD	GM with SVD
Accuracy (%)	85.6	86.3	89.0
Homogeneity	0.479	0.505	0.473
Completeness	0.506	0.508	0.491

Table 1: Clustering metrics with sample rate = 15%, and SVD with 2 components.

Figure 1: wordcloud of the GM after SVD clustering showing the top 30 terms of the means with size proportional to their importance.

From Figure 1 it is possible to distinguish the two categories. In the first one distinctive terms as “baby”, “seat” appear meanwhile in the second one appear “cat”, “dog”, “food”, “toy”.

To further get the most relevant terms, in the second part of this section results of the analysis keeping *only the nouns* are shown.

	Mini BatchK-Means	K-Means with SVD	GM with SVD
Accuracy (%)	64.2	79.4	85.5
Homogeneity	0.144	0.329	0.422
Completeness	0.248	0.366	0.431

Table 2: Clustering (only nouns) metrics with sample rate = 10%, and SVD with 2 components.

Figure 2: wordcloud of the GM after SVD clustering with only nouns.

In the second case the metrics of the clustering are worse than the previous ones, but the class of babies is slightly more understandable, with also the comparison of “toy” and “car” (probably meant as car seat) and the more relevance of the word “son”.