Big Data Computing – Homework 2

Paolo Mandica 1898788
Data Science, "Sapienza" University of Rome

Due Date:
December 23rd, 2020

1 Design choices and motivations

The design of the process for the completion of the homework can be divided into 3 main steps: 1) Pre-processing of the textual data; 2) Vectorization and SVD; 3) Clustering; plus the extra step of visualizing the clusters using WordCloud.

1.1 Pre-processing of textual data

The corpus is composed of reviews which include different types of characters. In order to perform the subsequent steps there is the need to pre-process the reviews. Here are the pre-processing steps:

1. converting all the characters to **lower-case**;
2. **removing** every character which isn't a letter of the alphabet or a number (which means **punctuation, special characters**, etc...);
3. applying **stemming and lemmatization** to each word (after some tests I found that lemmatization actually reduced both SVD explained variance and clustering accuracy, so I decided to just apply stemming);

1.2 Vectorization and Dimensionality Reduction

The next step consists in vectorizing the corpus in order to obtain a matrix, which will be then decomposed using Singular Vector Decomposition, to reducing the dimensionality of it.
For the purpose of this homework, I decided to try 3 different types of vectorization.

1. **Tfidf Vectorization** from scikit learn;
2. **Count Vectorization** from scikit learn;
3. **Tfidf Vectorization** implemented from scratch;

To reduce the dimensionality of the vectorized corpus, the choice was between PCA and SVD. The problem with PCA is that it works poorly with the matrix obtained from the vectorization, since it needs a sparse matrix which in this case can barely fits in RAM. For this reason I decided to use SVD.

After performing the steps above with the different types of vectorization, both the CountVectorizer and the Tfidf-from-scratch return results that, after being subjected to SVD, have higher explained variance when compared to the one from Tfidf Vectorizer. But Tfidf-from-scatch requires a lot of RAM to operate, so I decided to use the two Vectorizers from scikit learn.

1.3 Clustering

The algorithm used to cluster the reviews in the dataset is the basic K-Means with "*k-means++*" initialization and number of clusters equal to 2. The results obtained using the matrices from Tfidf Vectorizer and CountVectorizer, plus SVD, are:

- **CountVectorizer + SVD (2 components)** clustering accuracy $= 0.821$
- **Tfidf Vectorizer + SVD (7 components)** clustering accuracy $= 0.85$

2 Findings

Some interesting findings and results can be extrapolated from the previous section:

- **Lemmatization** doesn't help with the task at hand, while **stemming** leads to higher explained variance and clustering accuracy.
- **CountVectorizer + SVD** leads to higher explained variance but lower clustering accuracy.
- **Tfidf Vectorizer + SVD** leads to lower explained variance but higher clustering accuracy.
- **Tfidf-from-scratch** leads to higher explained variance compared to Tfidf Vectorizer but is not usable due to the amount of RAM it requires.

2.1 Relevant terms and WordCloud

Visualizing the most relevant terms in the clusters returned by KMeans we can see that the first cluster seem to have grouped reviews related to babies and family, while the second one grouped reviews regarding domestic animals.
Here is a better look at these results using the WordCloud library.

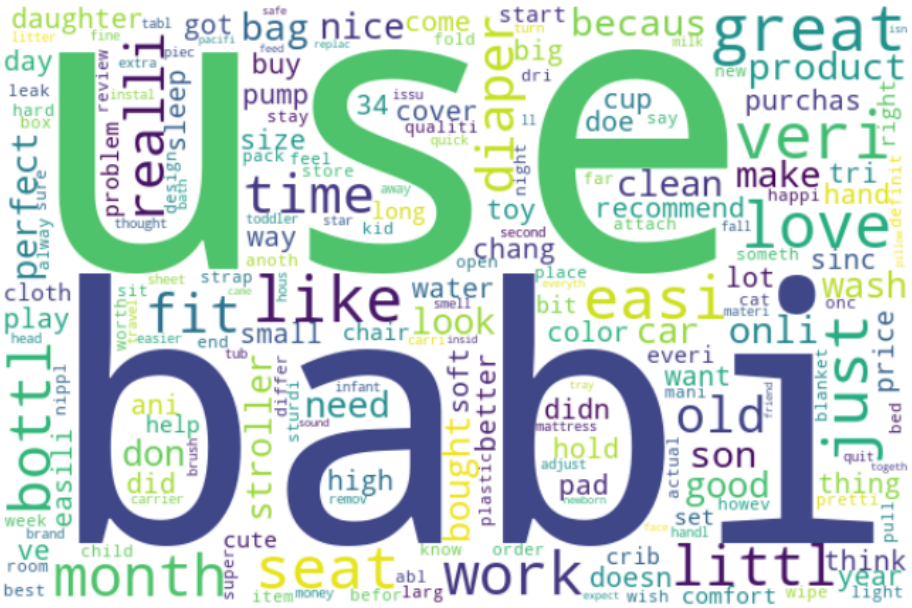

Figure 1: WordCloud visualization of cluster 1.

Figure 2: WordCloud visualization of cluster 2.