Using Text Analytics for Customer II

September 14, 2018

Using Text Analytics for Customer Feedback Analysis in the Airline World – Part II

Sounds Super Cool! But what is actually under the hood, how is it actually done?

In the last blog post we described how we used text analytics to visualize the diversity of topics covered in a publicly available dataset of airline customer reviews. In this topic we would like to get down to the nitty gritty technical details involved to perform latent semantic analysis (LSA). The outline for this procedure is the following:

- Creation of a term document matrix
- Dimensionality reduction through truncated singular value decomposition
- Multidimensional scaling for graphical visualization in 2D

LSA (Latent Semantic Analysis) is preceded through standard text preprocessing including tasks like character conversion to lower case as well as the removal of any punctuation. Now we are ready to create the td-matrix.

Creating the term document matrix*

Assuming m documents an exemplary term document matrix would look as follows.

	word 1	word 2	word 3	...	word n
Rev 1	2	5	29		5
Rev 2	0	4	32		0
Rev 3	0	0	27		
...					
Rev m	0	1	35		f

Here each row is linked to one distinct reviews while the columns label individual words. The set of words that enter is derived by creating the union of words from all documents. Thus, each row in the term document matrix represents the distribution of words used in the document and vice versa each column represents a distribution for a given word in each of the documents. To compensate for words that occur often but carry little meaning (e.g., “the”, ”a”, ”is”) multiply each entry with a weight called inverted document frequency.* Although a neat representation of the whole text corpus, it is still an ineffective way of representing the information content due to the sparse nature of these large matrices. This is the point where LSA comes into play.

Dimensionality reduction through truncated singular value decomposition

At the heart of LSA lies the Singular Value Decomposition (SVD). SVD TD-Matrix gives a factorization according to