Text Embeddings Based on Synonyms

Searching for text representation is one of the main tasks in information retrieval domain. The appropriate model has an impact on sentiment analysis also known as opinion mining. Take for instance books review sentiment studies. The goal is to assess people's opinions or emotions towards the book. Obviously, it may be applied in various fields such as recommendation systems. However, the quality of text representation affects the performance of this type of tasks.

In general, the problem of text model is to provide a vector representation of text document. The popular Bag of Words approach describes a document by a sparse vector. The vector elements are associated with the term frequency. Another technique is Term frequency - Inverse document frequency (TF-IDF) [1], where the importance of a word in the corpus is explored. Despite the fact that the methods have gained a lot of attention, they suffer some drawbacks. First of all, one of the disadvantages of the methods is high dimensionality of vectors. It causes few problems, especially in classification and clustering tasks. What is more, those models do not take word semantic similarity into consideration. For example, consider the distances among the three documents in Table 1. The first two documents (d_1 and d_2) express almost the same information. However, the third document should be located in another place in vector space. This is the reason why the semantics of documents has to be explored.

We construct a simple technique to represent text. Figure 1 provides an overview on the architecture of our model called the Bag of Centroids Model. It groups the semantically close words and causes the separation of documents which are expected to be far away from each other. The construction process includes two parts. After the first steps the collection of clusters with synonyms is obtained. To find the representation of the words, we use K-means algorithm [2]. Experimental results on UCI Amazon book reviews dataset [3] shows that our model performs favorably.

Table 1: A comparison of documents and their similarities.

Sentence	
d_1	She was smiling while reading an interesting book.
d_2	She was smiling while reading an impressive book.
d_3	She was smiling while reading an unexciting book.

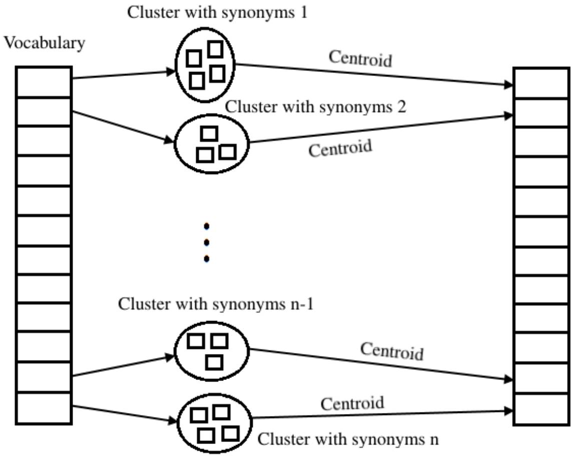

Figure 1: Architecture of Bag of Centroids

References

[1] R. Garreta and G. Moncecchi. *Learning Scikit-learn: Machine Learning in Python*. Packt Publishing, 2013.

[2] J. A. Hartigan and M. A. Wong. A k-means clustering algorithm. *JSTOR: Applied Statistics*, 28(1):100–108, 1979.

[3] M. Lichman. UCI machine learning repository, 2013.