Sentence clustering

Experiment settings

For each item and domain in the mobility datasets, we first apply a **word filtering**, followed by a **embedding mapping** of word groups to vectors which we then feed directly into a k-means clustering algorithm with a predefined target **number of clusters**. All bold values are settings for which we explored different configurations. We describe them all further below.

A first setting considered was to filter some words hypothesized to only add noise to the word groups. The different filters considered are (1) the absence of filter, (2) filtering all words with less than 4 letters and (3) filtering words contained in a stop-words public dictionary.

From a filtered word group we now needed to extract vectorial representations. This is done by using publicly available word-level vectorial mappings. The word-level mappings used are GLoVe with 50, 100, 200 and 300 dimensional embeddings. We also experimented with bio-medical vectorial embeddings, trained specifically on biomedical data. For those mappings, both word-level and sentence-level vectorial mappings are available and explored.

For the number of clusters to look for by our algorithm, we used values ranging from 4 to 8 included.

Results

Results are given in the form of a .csv file containing items and domains grouped by clusters. The clusters are listed in order of proximity to their respective clusters, meaning that the top word groups of each column should be strong indicators of their clusters' contents.

For the items clustering, the best found configuration was to use no word filtering, 50d GLoVe embeddings with 4 clusters. The cluster sizes are (87, 88, 302, 1083). While the clusters are heavily imbalanced, each cluster is clearly centered around a separate theme.

For the domains clustering, the best found configuration was to use no word filtering, 50d GLoVe embeddings with 4 clusters. The cluster sizes are (10, 11, 22, 87). Once again they are very imbalanced but inter-cluster correlation is strong.

Lastly, we also give a 2d representation of the obtained clusters obtained through PCA dimensionality reduction for both the items and domains.