عنوان مقاله:

Comparing k-means clusters on parallel Persian-English corpus

محل انتشار:

مجله هوش مصنوعی و داده کاوی, دوره 3, شماره 2 (سال: 1394)

تعداد صفحات اصل مقاله: 6

نویسندگان:

.A. Khazaei - *Electrical & Computer Engineering Department, Yazd University, Yazd, Iran*

.M. Ghasemzadeh - *Electrical & Computer Engineering Department, Yazd University, Yazd, Iran*

خلاصه مقاله:

This paper compares clusters of aligned Persian and English texts obtained from k-means method. Text clustering has many applications in various fields of natural language processing. So far, much English documents clustering research has been accomplished. Now this question arises, are the results of them extendable to other languages Since the goal of document clustering is grouping of documents based on their content, it is expected that the answer to this question is yes. On the other hand, many differences between various languages can cause the answer to this question to be no. This research has focused on k-means that is one of the basic and popular document clustering methods. We want to know whether the clusters of aligned Persian and English texts obtained by the k-means are similar. To find an answer to this question, Mizan English-Persian Parallel Corpus was considered as benchmark. After features extraction using text mining techniques and applying the PCA dimension reduction method, the k-means clustering was performed. The morphological difference between English and Persian languages caused the larger feature vector length for Persian. So almost in all experiments, the English results were slightly richer than those in Persian. Aside from these differences, the overall behavior of Persian and English clusters was similar. These similar behaviors showed that results of k-means research on English can be expanded to Persian. Finally, there is hope that despite many differences between various languages, clustering methods may be extendable to .other languages

کلمات کلیدی:

(Clustering, Mizan English-Persian Parallel Corpus, K-means, Principal Component Analysis (PCA