Clustering Documents with Kernel Density Estimation on Skip-gram graphs.

Levi Nicklas

FIRST DRAFT

2/7/2020

**Abstract:**

Through using graphical representations of documents, in the form of skip-gram graphs, clustering can be performed by using graph kernels and kernel density estimation. The data used will be a sample of comments from reddit (reddit.com), where users are discussing mental health and related topics.

**Introduction:**

Many methods used in text mining use *bag-of-words* techniques, but these methods do not always preserve the context and relationships between words. The graph representations for text, used here, aim to preserve the relationship and context between words. The idea of bigrams (pairs of two words) can be extended to “skip-grams”; skip-grams are words which appear within some width ***w*** of each other. This leads to more connections, and a richer graph representation of the text. This skip-gram method is similar to that used in the popular machine learning package *Word2Vec*, which uses the “continuous bag of words” or “continuous skip-grams”. In this paper, a graph representation of the text is constructed, using skip-gram methods, for each comment in the dataset. The graph representations are compared with a graph kernel, which is a measure of the similarity of two graphs. In this study, the graph kernel of choice is a “edge histogram kernel”, because the kernel utilizes the labels of the vertices to assess the similarity of two graphs– not just the topology. For each graph, the kernel produces a measure of similarity to the other ***n-1*** graphs in the dataset. We can use this similarity measure between two graphs, to assess how similar other graphs in the dataset may be. That is, we can compare the similarity of graphs ***B*** and ***C*** through their similarity to ***A***. To do this, a KDE (Kernel Density Estimation) to estimate a kernel density curve, the curve is then used to partition the rest of the dataset into clusters. The KDE bandwidth is modified to produce more or less local minima/maxima that can be then used to identify potential clusters. Basic calculus can be used to locate to inflection points that will serve as intervals for which a cluster will be defined.

**Methods:**

**Analysis:**

**Conclusion:**