# SimilarNews Optimization

written by Haiqin Huang(z5295162)

The codes implementation is basically following the guidance in project3 help session.

The first stage is to sort every headline in the least frequent which could help to filter lots of not similar pair in second stage. The second stage is to select the token according to the input ‘similarity threshold t’ and construct similar pairs. For every headline, the first ((1-t) * headline +1) will be select as key to construct pairs, because a headline1 has t similarity with headline2, headline must have a same word in the part of (1-t)*headline. The last stage is to filter some pairs with a similarity lower than t and remove duplicated key.