DATA DESCRIPTION

Data contains an alternative phrase clustering of the MemeTracker dataset phrases, as defined in:

E. Omodei, T. Poibeau and J.-P. Cointet
"*Multi-Level Modeling of Quotation Families Morphogenesis*"
2012 ASE/IEEE Intl. Conf. On Social Computing (SocialCom), 2012

The MemeTracker dataset is a publicly available database downloadable from http://memetracker.org and introduced in:
J. Leskovec, L. Backstrom and J. Kleinberg, "Meme-tracking and the Dynamics of the News Cycle", ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining, 2009

For each phrase cluster the data contains all the phrases in the cluster and a list of URLs where the phrases appeared, encoded as in the MemeTracker dataset (the only difference is the absence of the root phrase).

The file *clust-omodeicointetpoibeau.txt* contains all the phrases, whereas the file *clust-omodeicointetpoibeau-filtered.txt* only contains clusters and phrases remaining after applying the filters described in section III.C of the article ("Since the dataset contains many quotations that are either not in English either too short to convey a real unit of meaning, we first decided to filter it by considering only quotations containing at least 5 words in English").