News aggregation

In this task you will have to perform news aggregation. You are provided with the corpus of news pieces gathered from different British websites (e.g. BBC News, Sky News, The Guardian, etc.). Your task is to find ``stories'' among them. A ``story'' is a group of news pieces covering one event, which may evolve when the time goes by (see example in the attached presentation).

We performed topic modeling on the corpus. A ``topic'' is some distribution over words; usually there is a small set of words with high probability (key words) and the rest with low probability. For each piece of news we computed a distribution over topics in it (``document''). These are your initial features for the task.

Data description:

news_formatted.txt: first line contains the number of news pieces M; M following lines contain a news piece, one per each line (punctuation eliminated);

time_stamps.txt: M lines with time stamps for each news piece;

wordmap.txt: first line contains the number of words in vocabulary V; next V lines consist of 2 strings per line: word [space] word_id;

news.phi: matrix phi of the size K x V (K=200), phi(k, w) corresponds to the frequency of assigning word_id w to the topic_id k;

news.twords: for each topic here are top-40 words with its frequency;

news.theta: matrix theta of the size M x K, theta(m, k) corresponds to the frequency of topic_id k in the piece m;

news.tassign: mth line contains information for the piece m in the form ``word_id1:topic_id1 word_id2:topic_id2, … word_idN:topic_idN''.