Problem description:

1. Find 100 CNN news articles online (try to find them in different categories, e.g., sports and finance). You need to find the new article by yourself. Pls ignore picture or other non-text data in the new article.
2. Convert them to data matrix (each row is an article and each column is a unique term).
3. Compute the similarity between each pair of articles with Euclidean distance (you need to convert the Euclidean distance to similarity), Cosine and Jaccard.
4. Sort all pairs from most similar to least similar based on each of the three types of similarity measurements
5. Compare the sorted pairs and discuss which one is more accurate (i.e., close to your own judgment)

Deadline: please submit your results by 02/13 . You may compress your 100 news articles, data matrix, sorted pairs and your own discussion into a zip file to submit.