英文CNN数据集是模拟浏览器行为从CNN上爬取下来的。首先每天从Twitter上的美国Trending列表（当前趋势，可以视为热点话题）中选取5个加入待爬取话题列表。选取话题的方法是每天从早上8点到晚上8点每隔一个小时爬取一次排名前20的Trending，每天结束时将所有Trending聚合去重，根据Trending的出现次数和每次排名加权得到当日最热门的5个Trending，最后得到61个话题。

使用话题作为关键字模拟浏览器行为在CNN中进行搜索，抓取和话题相关的新闻40条（只检索文章，并且按照与话题相关度高低排序），采集新闻时会将新闻本身的发布时间作为新闻时间文本保存在新闻内容中，可以作为缺少时间的文本作为时间。通过这种方式，最后一共收集到2377篇新闻文档，46855条英文新闻段落。