Problem

1. For most cases number is found in the headline.
2. The word “নিহত” appears before or after the number within 3 words.
3. Sometimes “মৃত্যু” is used instead of “নিহত”.
4. The word “সহ” appears before the number (if there is women or children included in the death count).
5. Sometimes “জন” can be found after the number.
6. Sometimes there is no number (nor verbal neither numeral). Instead there is relation between people or occupation. Ex: দম্পতি, মা-মেয়ে, ভাই-বোন, কৃষক, যুবক, শিশু, মামা-ভাগ্নে, সভাপতি, গৃহবধূ etc. when a relation is used there is **usually** 2 casualties and for occupation or age type there is 1.
7. Sometimes age of the victim is provided.
8. Some news provides statistics of a certain time period of time.
9. Some news includes both news of current death and statistics.
10. “দগ্ধ”, “আহত” words used to count injured people.
11. There are some international news.
12. Combined news of “বজ্রপাত ও ঝড়”! Details can be found in the detailed news, but the title arises confusion.
13. Some news come with news that provide different number from different area, and that news may not provide a total number. Ex: বজ্রপাতে মাদারীপুরে তিনজন ও শরীয়তপুরে একজনের মৃত্যু হয়েছে.
14. More confusion! #12 is the detailed part of the situation #11☹.

Solution

1. Extract information
2. 7-gram the corpus
3. Keep lines having number (e.g. “নিহত”, “মৃত্যু”, “জন” or number list) at the 4th (middle element for 7-gram).
4. Search for numbers stored in a list (numeral or verbal).
5. Result that contain max valued number is the total death count.
6. Max valued number can be found by element’s list index.
7. Other numbers contain information about victim type.
8. From other numbers and their neighboring word we can determine the number of male, female or child casualties (we may be needing larger n-gramming). But for getting detailed information I need further study and implementation. We will need to identify names and gender from name.