

# A Text-Processing Warmup

This problem will help you warm up and practice basic text and string processing techniques. This will be a first step towards more complex Text and Natural Language Processing and Analysis tasks.

You will be given a fragment of text.

1. In this fragment, you need to identify the articles used (i.e., 'a', 'an', 'the').
2. And you also need to identify dates (which might be expressed in a variety of ways such as '15/11/2012', '15/11/12', '15th March 1999', '15th March 99' or '20th of March, 1999').

You can make the following assumptions 1) In the date, year and day will always be in numeric form. Which means, you don't have to worry about "fifteenth" or "twentieth" etc. Month, could be either numeric form (1-12) or with its name (January-December, Jan-Dec).

2) This is a bit open ended, and somewhat intentionally so. The aim is for you to try to write something which figures out as many common patterns as possible, in which dates are present in text.

3) Most of the test cases are Wikipedia articles. Having a look at the common formats in which dates occur in those, will help.

4) Dates could either be in the form: Month followed by Day followed by Year, or Day followed by Month followed by Year.

5) The day could be in the form of either (1,2,3,...31) or (1st, 2nd, 3rd...31st).

A fragment is a valid date if it contains day, month and year information (all three of them should be present). To extract date information, you will need to try detecting different kinds of representations of dates, some of which have been shown above. The more patterns you match and identify correctly, the greater your score will be.

## Input Format

First line contains the number of test cases T. This is followed by T test fragments (each fragment will be in one line and each will have a blank line after it) . Each line contains a paragraph of text in which you need to identify the articles and dates. There will be a blank line after each paragraph.

So, totally there are  $2T+1$  lines in the input file. The last one is a blank line after the last text fragment.

## Output Format

4T lines, four lines of output for each test case. First line -> number of occurrences of 'a'. Second line -> number of occurrences of 'an'. Third Line -> number of occurrences of 'the'. Fourth Line -> number of occurrences of date information.

## Sample Input

5

Delhi, is a metropolitan and the capital region of India which includes the national capital city, New Delhi. It is the second most populous metropolis in India after Mumbai and the largest city in terms of area.

Mumbai, also known as Bombay, is the capital city of the Indian state of Maharashtra. It is the most populous city in India, and the fourth most populous city in the world, with a total metropolitan area population of approximately 20.5 million.

New York is a state in the Northeastern region of the United States. New York is the 27th-most extensive, the 3rd-most populous, and the 7th-most densely populated of the 50 United States.

The Indian Rebellion of 1857 began as a mutiny of sepoys of the East India Company's army on 10 May 1857, in the town of Meerut, and soon escalated into other mutinies and civilian rebellions largely in the upper Gangetic plain and central India, with the major hostilities confined to present-day Uttar Pradesh, Bihar, northern Madhya Pradesh, and the Delhi region.

The Boston Tea Party (referred to in its time simply as "the destruction of the tea" or by other informal names and so named until half a century later,[2]) was a political protest by the Sons of Liberty in Boston, a city in the British colony of Massachusetts, against the tax policy of the British government and the East India Company that controlled all the tea imported into the colonies. On December 16, 1773, after officials in Boston refused to return three shiploads of taxed tea to Britain, a group of colonists boarded the ships and destroyed the tea by throwing it into Boston Harbor. The incident remains an iconic event of American history, and other political protests often refer to it.

### Sample Output

```
1
0
4
0
1
0
5
0
1
0
6
0
1
0
6
1
4
1
13
1
```

### Line Wise Explanation of the Output

- 1 #Number of "a" in first text segment
- 0 #Number of "an" in first text segment
- 4 #Number of "the" in first text segment
- 0 #Number of dates in first text segment
- 1 #Number of "a" in second text segment
- 0 #Number of "an" in second text segment
- 5 #Number of "the" in second text segment
- 0 #Number of dates in second text segment
- 1 .....
- 0 .....

Delhi, is[a]metropolitan and[the]capital region of India which includes[the]national capital city, New Delhi. It is[the]second most populous metropolis in India after Mumbai and[the]largest city in terms of area.

Mumbai, also known as Bombay, is[the]capital city of[the]Indian state of Maharashtra. It is[the]most populous city in India, and[the]fourth most populous city in[the]world, with[a]total metropolitan area population of approximately 20.5 million.

New York is[a]state in[the]Northeastern region of[the]United States. New York is[the]27th-most extensive,[the]3rd-most populous, and[the]7th-most densely populated of[the]50 United States.

[The]Indian Rebellion of 1857 began as[a]mutiny of sepoysof[the]East India Company's army on[10 May 1857], in[the]town of Meerut, and soon escalated into other mutinies and civilian rebellions largely in[the]upper Gangetic plain and central India, with[the]major hostilities confined to present-day Uttar Pradesh, Bihar, northern Madhya Pradesh, and[the]Delhi region.

[The]Boston Tea Party (referred to in its time simply as [the]destruction of[the]tea" or by other informal names and so named until half[a]century later,[2]) was[a]political protest by[the]Sons of Liberty in Boston,[a]city in[the]British colony of Massachusetts, against[the]tax policy of[the]British government and[the]East India Company that controlled all[the]tea imported into[the]colonies. On[December 16, 1773], after officials in Boston refused to return three shiploads of taxed tea to Britain,[a]group of colonists boarded[the]ships and destroyed[the]tea by throwing it into Boston Harbor.[The]incident remains[an]iconic event of American history, and other political protests often refer to it.

## SCORING

Your score will be proportional to the number of lines in your output which match with the expected and correct output.

Score = MaxScore \* (correctly computed values in your output) / (4 x Number of Text Fragments)