Stage 1

The 2 challenges below can be found on Hackerrank. We have simply copied them here for convenience.

To submit your results, make a public repo on Github called "NLP Challenges" on where you should upload your code in different folders for different challenges. Share the link with us and we would review the code and your approach and get back to you.

If it is found that the code was plagiarised, the applicant would get disqualified immediately.

1. Guess the flipkart Query (Information Retrieval - Reverse)

Flipkart is a popular Indian e-commerce portal. One of the most common actions performed by users of the portal, is to use the search box and query for a brand, product or product-line. The search box then returns the best matching products which it can find - along with their prices, details, descriptions, etc.

We tried out twenty different search queries (specified below), and made a list of some of the product names which were returned in response to them. You are provided with a list of N names of products from this list. Your task is to guess, which search query each of them was returned in response to.

The twenty search queries we made were:

```
axe deo
best-seller books
calvin klein
```

```
camcorder
camera
chemistry
chromebook
c programming
data structures algorithms
dell laptops
dslr canon
mathematics
nike-deodrant
physics
sony cybershot
spoken english
timex watch
titan watch
tommy watch
written english
```

Here's a small example of the task at hand:

In response to which of these queries, was the product 'Dell Vostro 2520 Laptop (2nd Gen PDC/ 2GB/ 320GB/ Linux...' (most likely) returned?

Answer:dell laptops

In response to which of these queries, was the product 'Calvin Klein One Eau de Toilette - 200 ml' (most likely) returned?

Answer:calvin klein

Input Format

The first line contains an integer N.

This is followed by N lines each containing the name of a product.

Input Constraints

1 <= N <= 200

The product names will not exceed 200 characters in length. Sometimes, when the product name is long and descriptive, after the first 55 characters, there are likely to be a series of dots, such as the examples below. Please handle them appropriately (strip them off, or ignore them).

```
Laptops: AMD Mobile Platform, AMD Vision, Barebook, Cen...
Dell Vostro 2520 Laptop (2nd Gen PDC/ 2GB/ 320GB/ Linux...
Dell Inspiron 15R 5521 Laptop (3rd Gen Ci7/ 8GB/ 1TB/ W...
```

Output Format

The output should contain exactly N lines.

The i^{th} line should contain the query (your best guess) which returned the i^{th} product name in the input file. The query should strictly be from one of the twenty queries specified above, as is. Please do not add any leading or trailing spaces or any extra punctuation. Also ensure that the case remains the same.

Sample Input, Output and Training Files

The sample input, output and training files can be accessed at the following links:

Training File

Sample Input

Sample Output

The training file can also be opened from your program, during execution. It can be opened using the name "training.txt" and is available in the same directory where the program is being run.

Sample Input

```
60
Data Structures and Algorithms with Object- Oriented Design Patterns in C++ 1 Edition (Paperback)
God Moments: Stories That Inspire, Moments to Remember (Paperback)
The Ultimate C: Concepts, Programs and Interview Questions (Paperback)
Canon EOS 1100D SLR (Black, with Kit (EF S18-55 III))
A Textbook of Organic Chemistry for JEE Main & Advanced and Other Engineering Entrance Examinations (Paperback)
Test your C ++ Skills 1 Edition (Paperback)
IIM Ahmedabad Business Books: Day to Day Economics (Paperback)
Calvin Klein One Eau de Toilette  -  200 ml
..........
```

Sample Output

```
data structures algorithms
written english
c programming
dslr canon
chemistry
c programming
best-seller books
calvin klein
............
```

Explanation

The first product in the sample input is a book 'Data Structures and Algorithms with Object- Oriented Design Patterns in C++ 1 Edition (Paperback)' which was returned in response to the query 'data structures algorithms'. The second product in the sample input is a paperback book 'God Moments: Stories That Inspire, Moments to Remember (Paperback)' which was returned in response to the query 'written english'. Please note, that as in the real world, there are always cases like the second one, where it is

nearly impossible - to identify which is the most appropriate search query which led to this product: that is fine - you can answer with your best guess in such situations.

Training File

A small training file with a few examples of products returned for the various search queries is available. Please note, that this is only a small training file, and it is expected that a mix of simple and creative ideas from machine learning, string matching and information retrieval will be used in the submitted solution.

The format of the training file is as specified:

The first line contains an integer N.

This is followed by N lines each containing the product name, and the search query, separated by a tab character.

```
N
productName_1   query_1
productName_2   query_2
productName_3   query_3
....
```

Here's a quick look at what the training file looks like:

```
111
Calvin Klein IN2U Eau de Toilette  -  150 ml (For Men)  calvin klein
For The Love of Physics (Paperback) physics
Nike Fission Deodorant Spray  -  200 ml (For Men)   nike-deodrant
Spoken English (With CD) 2nd Edition (Paperback)    spoken english
The C++ Programming Language 3 Edition (Paperback)  c programming
......
```

Scoring

Your score for a test case = C/N * M where:

M = Maximum Score for the test case C = search queries correctly identified N = Total number of product names in the test case

The hidden test case carries thrice as much weightage as the sample test case (which is visible on hitting 'Compile and Test').

Libraries

Libraries available in our Machine Learning/Real Data challenges will be enabled for this contest and are listed here. Please note, that occasionally, a few functions or modules might not work in the constraints of our infrastructure. For instance, some modules try to run multiple threads (and fail). So please try importing the library and functions and cross checking if they work in our online editor in case you plan to develop a solution locally, and then upload to our site.

2. To be or to not be (Predicting)

The verb "to be", in its different forms, is one of the most commonly used building blocks of the English language. Many examples and different forms of "to be" are demonstrated here.

You are provided a paragraph of text in which some of the derivates of this verb have been blanked out. At various points in the text, occurences of 'am','are','were','was','is','been','being','be' have been blanked out and replaced with a series of four consecutive hyphens (----). Your task is to identify which of these words can appropriately fill up these blanks.

Input Format

The input will contain two lines. The first line will contain only one integer N, which will equal the number of blanks in the text. The second line contains one paragraph of text. Several occurences of the words mentioned previously have been blanked out and replaced by four consecutive hyphens (----). These are the blanks which you need to fill up with one of the following words: 'am','are','were','was','is','been','being','be'

Output Format

The output should contain exactly N lines. This is followed, by the appropriate words which need to be filled up in the N blanks in the provided paragraph of text, in the same order as the blanks which they are intended for, respectively.

Sample Input

```
6
When the modern Olympics began in 1896, the initiators and organizers ---- looking for a great popularizing event, recalling the ancient glory of Greece. The idea of a marathon race came from Michel Breal, who wanted the event to feature in the first modern Olympic Games in 1896 in Athens. This idea was heavily supported by Pierre de Coubertin, the founder of the modern Olympics, as well as by the Greeks. The Greeks staged a selection race for the Olympic marathon on 10 March 1896 that ---- won by Charilaos Vasilakos in 3 hours and 18 minutes (with the future winner of the introductory Olympic Games marathon coming in fifth). The winner of the first Olympic Marathon, on 10 April 1896 (a male-only race), was Spyridon "Spyros" Louis, a Greek water-carrier, in 2 hours 58 minutes and 50 seconds. The women's marathon ---- introduced at the 1984 Summer Olympics (Los Angeles, USA) and ---- won by Joan Benoit of the United States with a time of 2 hours 24 minutes and 52 seconds. Since the modern games were founded, it has become a tradition for the men's Olympic marathon to be the last event of the athletics calendar, with a finish inside the Olympic stadium, often within hours of, or even incorporated into, the closing ceremonies. The marathon of the 2004 Summer Olympics revived the traditional route from Marathon to Athens, ending at Panathinaiko Stadium, the venue for the 1896 Summer Olympics. Since the modern games ---- founded, it has become a tradition for the men's Olympic marathon to be the last event of the athletics calendar, with a finish inside the Olympic stadium, often within hours of, or even incorporated into, the closing ceremonies. The marathon of the 2004 Summer Olympics revived the traditional route from Marathon to Athens, ending at Panathinaiko Stadium,
```

```
the venue for the 1896 Summer Olympics. The Olympic men's record ---- 2:06:32.
```

Input Constraints

1 <= N <= 20 The text chunk will not contain more than 5000 characters.

Sample Output

```
were
was
was
was
were
is
```

Explanation

The blanks have been filled up with appropriate words as shown below:

When the modern Olympics began in 1896, the initiators and organizers **were** looking for a great popularizing event, recalling the ancient glory of Greece. The idea of a marathon race came from Michel Breal, who wanted the event to feature in the first modern Olympic Games in 1896 in Athens. This idea was heavily supported by Pierre de Coubertin, the founder of the modern Olympics, as well as by the Greeks. The Greeks staged a selection race for the Olympic marathon on 10 March 1896 that **was** won by Charilaos Vasilakos in 3 hours and 18 minutes (with the future winner of the introductory Olympic Games marathon coming in fifth). The winner of the first Olympic Marathon, on 10 April 1896 (a male-only race), was Spyridon "Spyros" Louis, a Greek water-carrier, in 2 hours 58 minutes and 50 seconds. The women's marathon **was** introduced at the 1984 Summer Olympics (Los Angeles, USA) and **was** won by Joan Benoit of the United States with a time of 2 hours 24 minutes and 52 seconds. Since the modern games were founded, it has become a tradition for the men's Olympic marathon to be the last event of the athletics calendar, with a finish inside

the Olympic stadium, often within hours of, or even incorporated into, the closing ceremonies. The marathon of the 2004 Summer Olympics revived the traditional route from Marathon to Athens, ending at Panathinaiko Stadium, the venue for the 1896 Summer Olympics. Since the modern games **were** founded, it has become a tradition for the men's Olympic marathon to be the last event of the athletics calendar, with a finish inside the Olympic stadium, often within hours of, or even incorporated into, the closing ceremonies. The marathon of the 2004 Summer Olympics revived the traditional route from Marathon to Athens, ending at Panathinaiko Stadium, the venue for the 1896 Summer Olympics. The Olympic men's record **is** 2:06:32.

Corpus of Text

You are provided with a corpus of text, which might assist you with the task at hand. This will also be available with the name “corpus.txt” in the same folder as the one where your program is executed, when you submit your solution. The corpus is located here.

Scoring

Each test carries a weightage proportional to the number of blanks in it. Score for each test case = M * C/N

Where C = Number of correct answer N = Number of total blanks M is the maximum score for the test case.

After the contest, submissions will be re-run after adding five hidden test cases as well.