

New 1 4

Home 1 2 3 4

Sales 1 2 3 4

Top 1

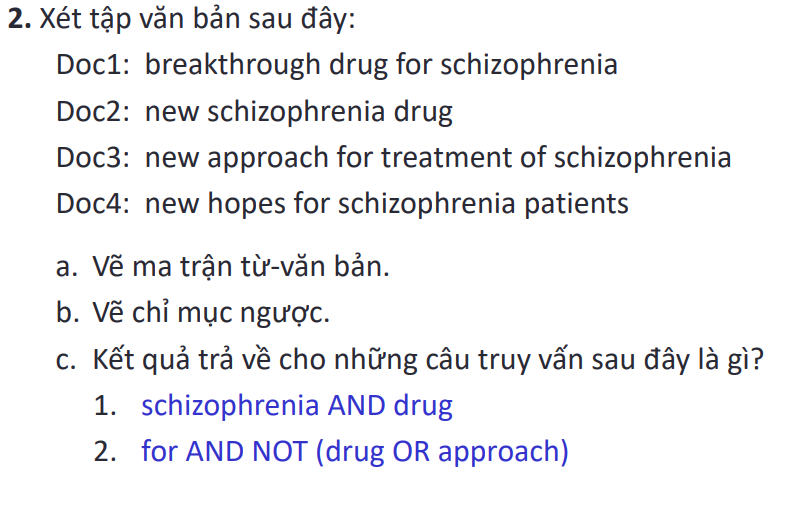
Forecasts 1

Rise 2 4

In 2 3

July 2 3 4

Increase 3



a, **Vẽ ma trận từ-văn bản**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Doc1** | **Doc2** | **Doc3** | **Doc4** |
| **breakthrough** | 1 | 0 | 0 | 0 |
| **drug** | 1 | 1 | 0 | 0 |
| **for** | 1 | 1 | 0 | 0 |
| **schizophrenia** | 1 | 1 | 1 | 1 |
| **new** | 0 | 1 | 1 | 1 |
| **approach** | 0 | 0 | 1 | 0 |
| **treatment** | 0 | 0 | 1 | 0 |
| **of** | 0 | 0 | 1 | 0 |
| **hopes** | 0 | 0 | 0 | 1 |
| **patients** | 0 | 0 | 0 | 1 |

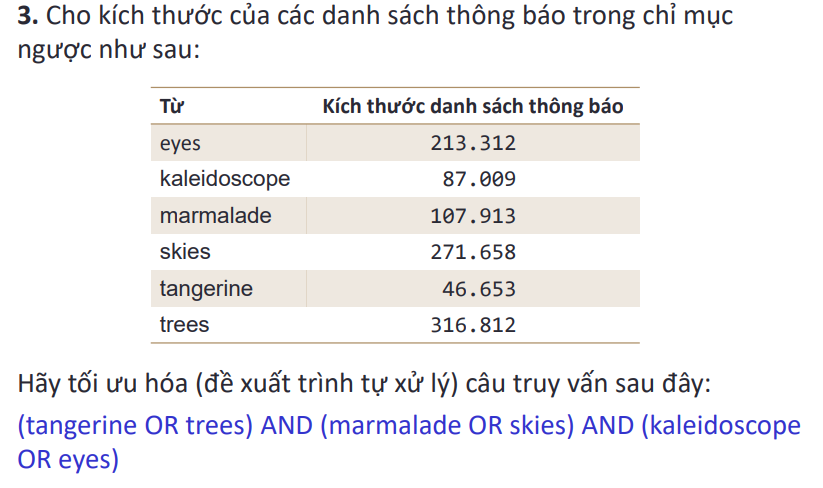
b, Vẽ chỉ mục ngược

|  |  |
| --- | --- |
| **breakthrough** | **1** |
| **drug** | **1 2** |
| **for** | **1 3 4** |
| **schizophrenia** | **1 2 3 4** |
| **new** | **2 3 4** |
| **approach** | **3** |
| **treatment** | **3** |
| **of** | **3** |
| **hopes** | **4** |
| **patients** | **4** |

c, schizophrenia AND drug = 1 , 2

for AND NOT (drug OR approach) = 4

Chỉ có “for” mà ko có “drug” hoặc “approach”



|  |  |
| --- | --- |
| **Từ** | **Kích thước danh sách thông báo** |
| tangerine | **46.653** |
| kaleidoscope | **87.009** |
| marmalade | **107.913** |
| eyes | **213.312** |
| skies | **271.658** |
| trees | **316.812** |

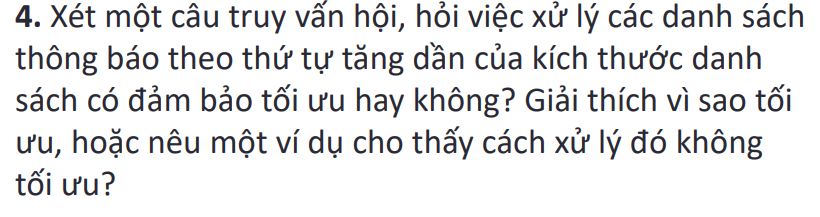
tangerine OR trees: 46653 + 316812 = **363465**

marmalade OR skies: 107913 + 271658 = **379571**

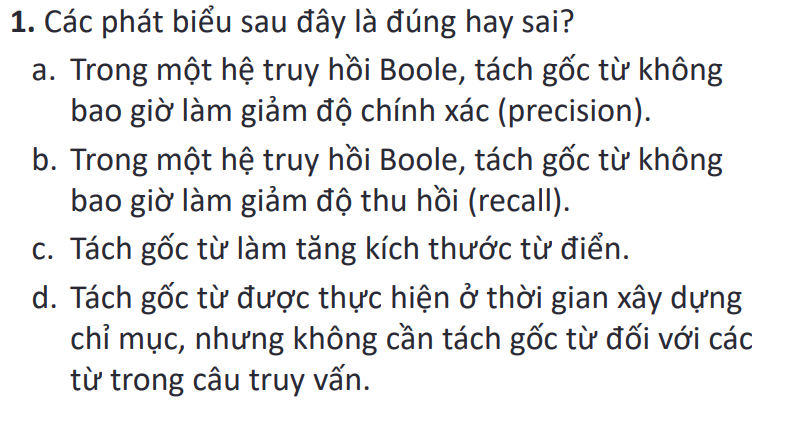
kaleidoscope OR eyes: 87009 + 213312 = **300321**

**Sắp xếp từ bé đến lớn:**

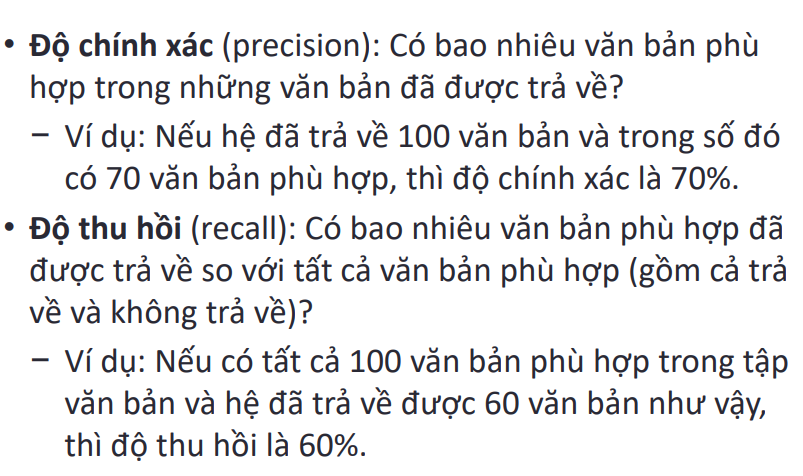
(kaleidoscope OR eyes) AND (tangerine OR trees) AND (marmalade OR skies)

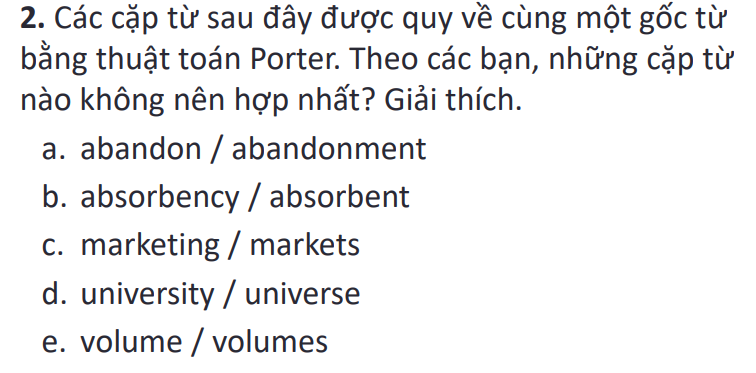


* Có
* Giao của hai danh sách không dài hơn danh sách ngắn hơn
* Nếu ta bắt đầu bằng cách lấy giao của hai danh sách thông báo ngắn nhất, thì tất cả các kết quả trung gian đều không dài hơn danh sách thông báo ngắn nhất.



1. Sai
2. Đúng
3. Sai
4. Sai





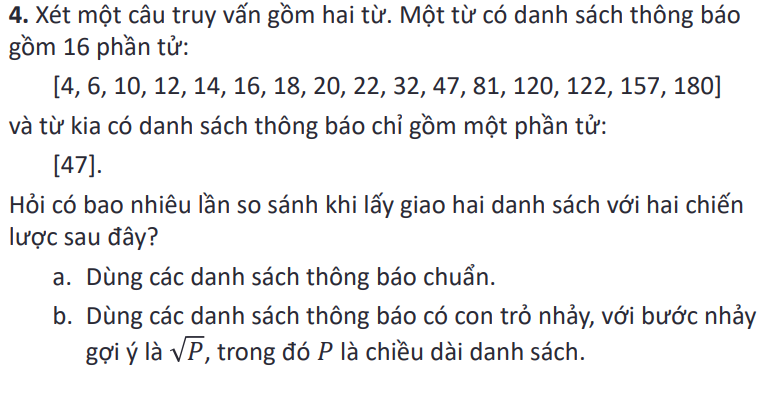
a, bỏ rơi / bỏ rơi

b, thấm hút / hấp thụ

c, tiếp thị / thị trường

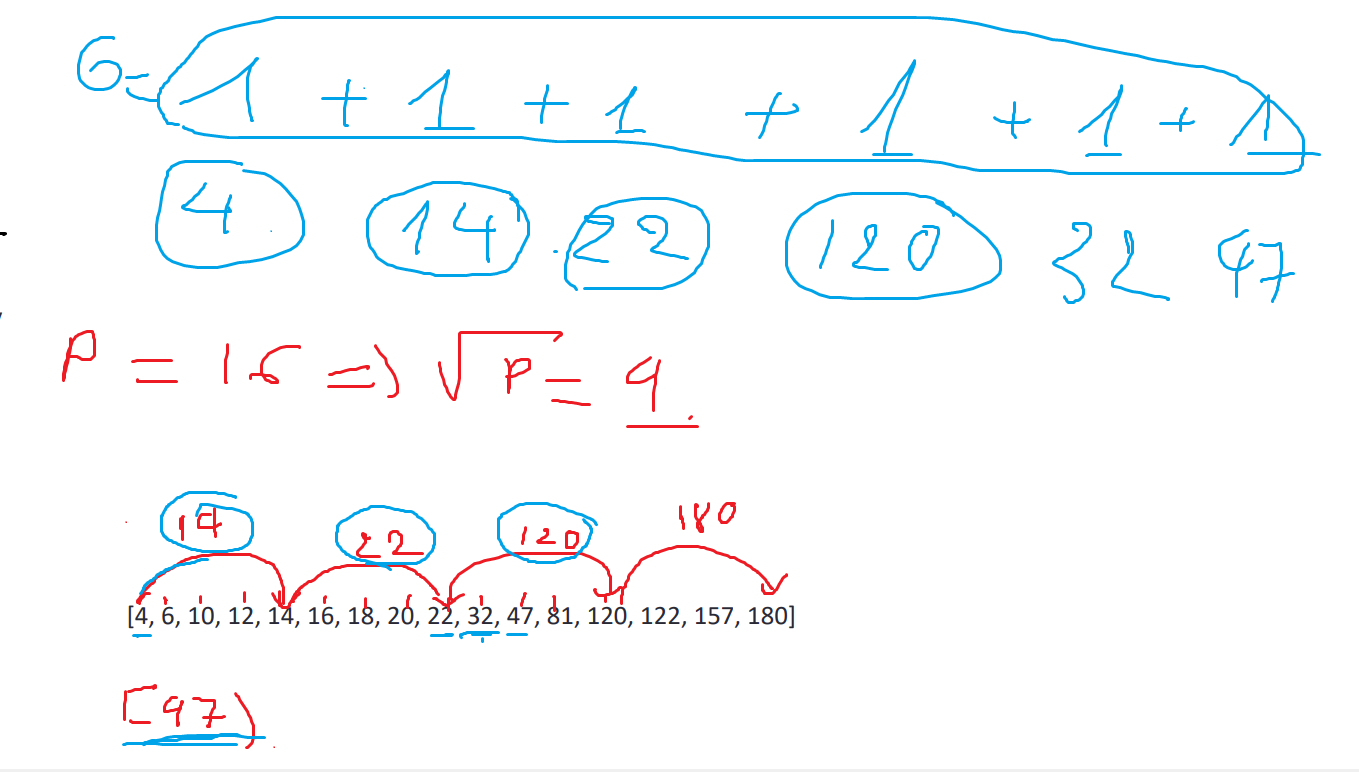
d, trường đại học / vũ trụ

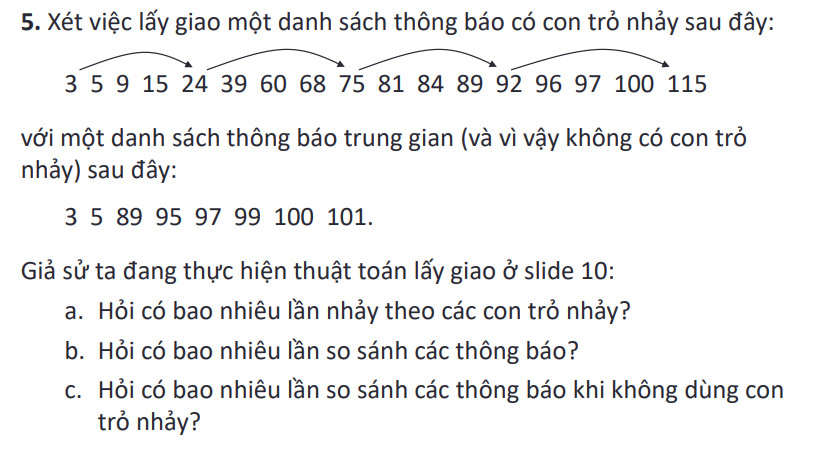
e, âm lượng / tập

**Các từ khác nghĩa không nên gộp**

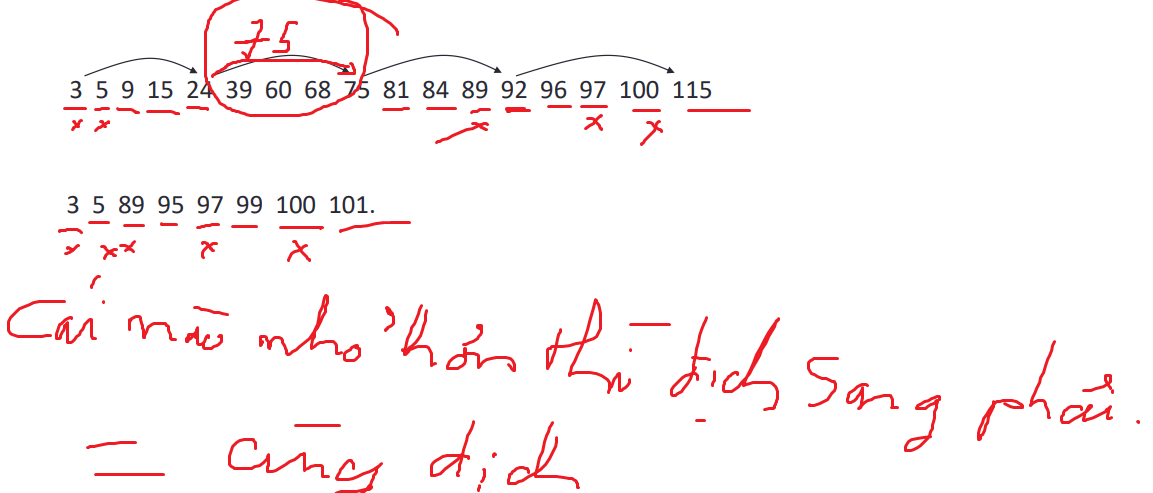
a, 11 (số lần so sánh đến 47 = 11 lần)

b, 6

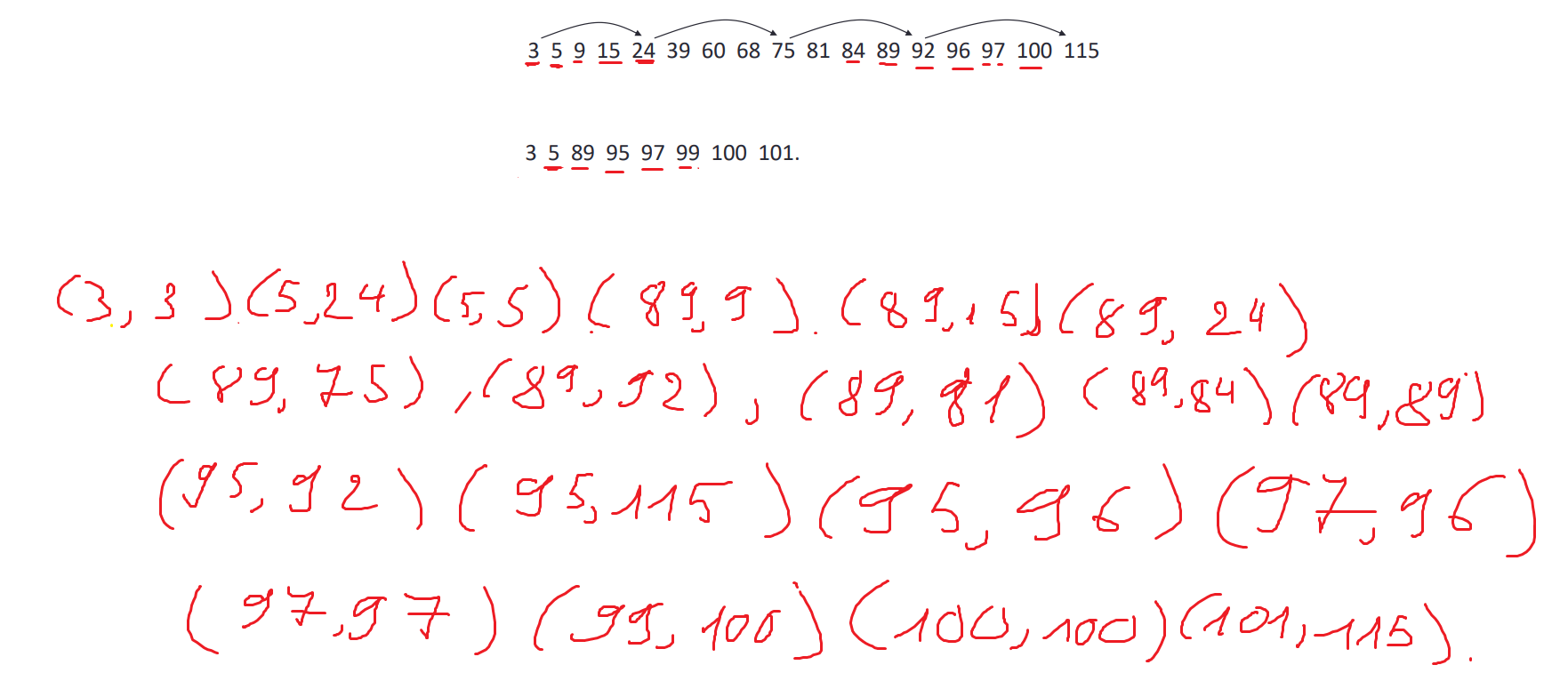


****

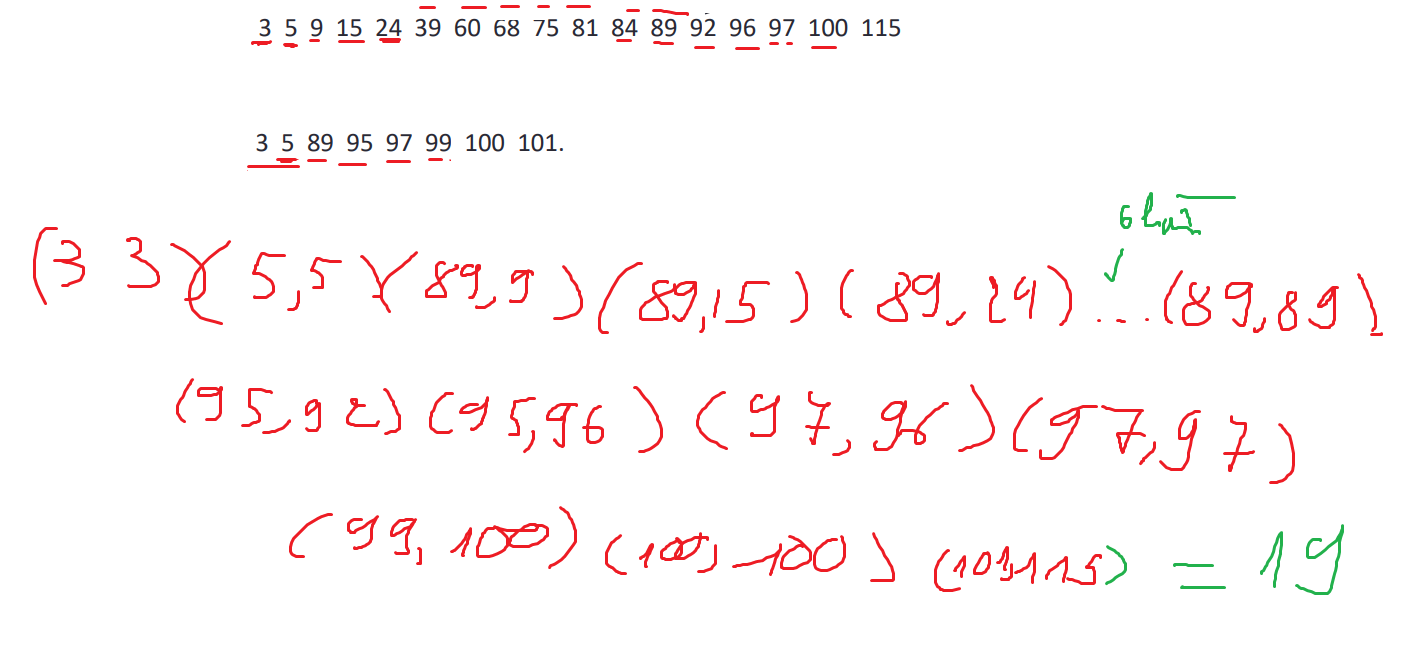
**a, 1**

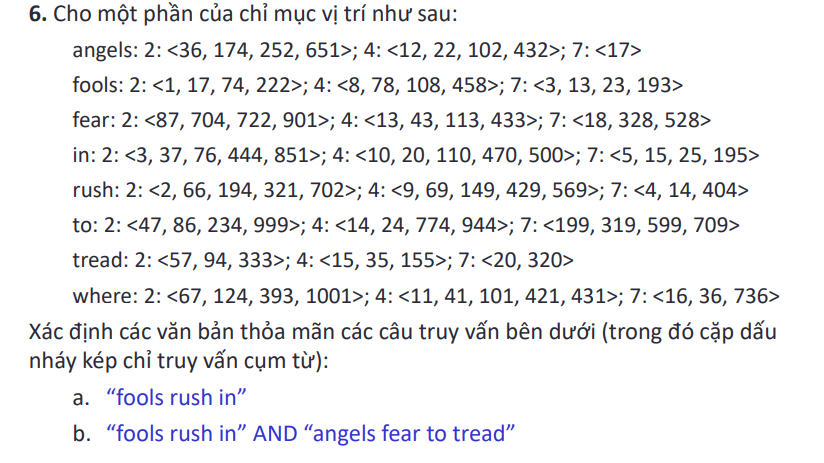
****

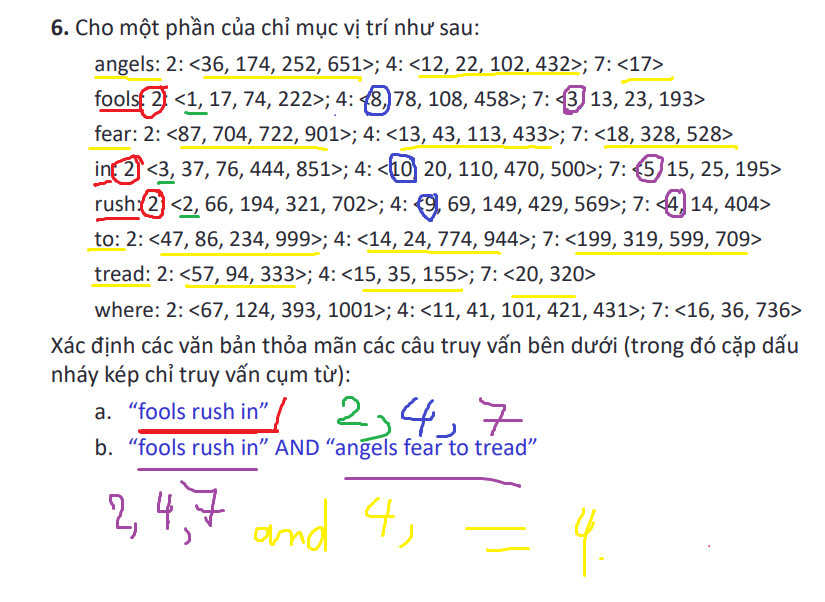
**b, 19**

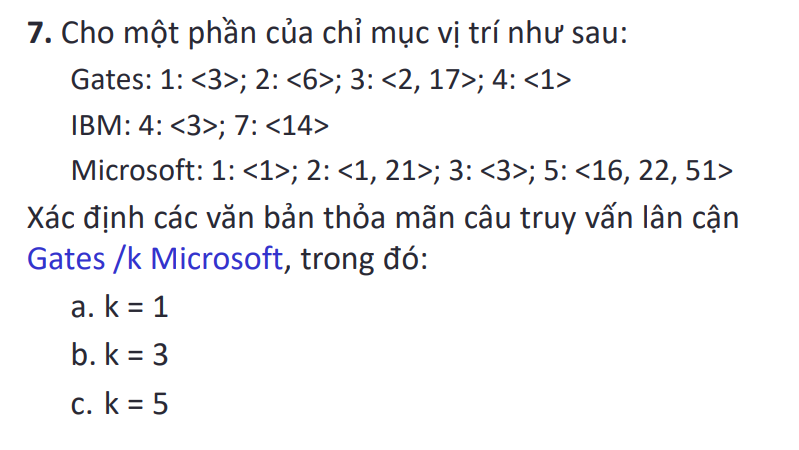
****

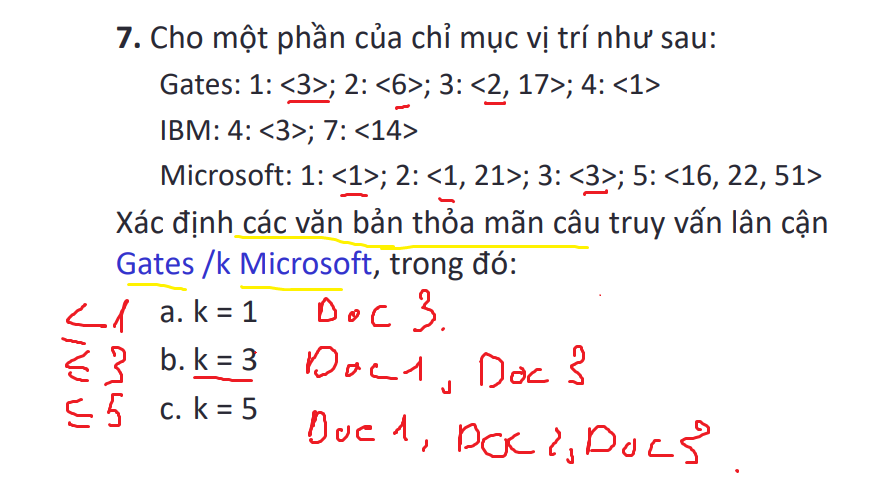
**c, 19**

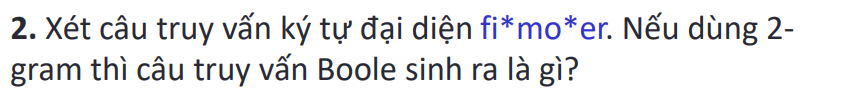
****











$f

fi

i\*

\*m

mo

o\*

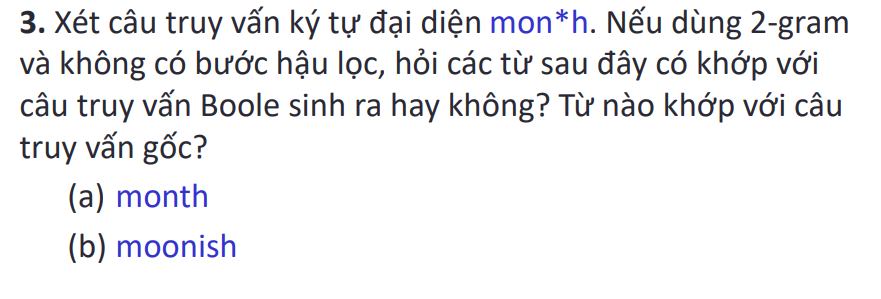
\*e

er

r$

Bỏ các cái có dấu \*

$f and fi and mo and er and r$



$m

mo

on

n\*

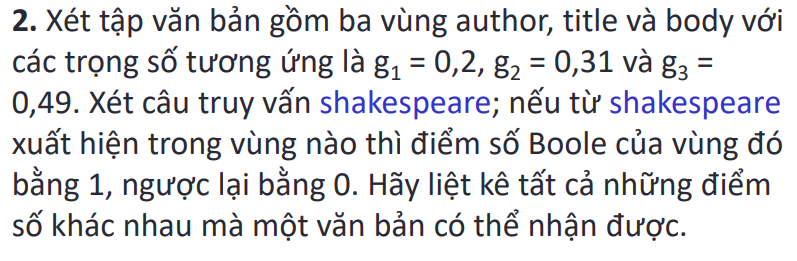
\*h

h$

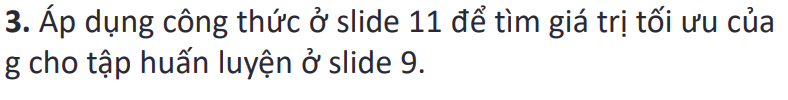
* $m and mo and on and h$

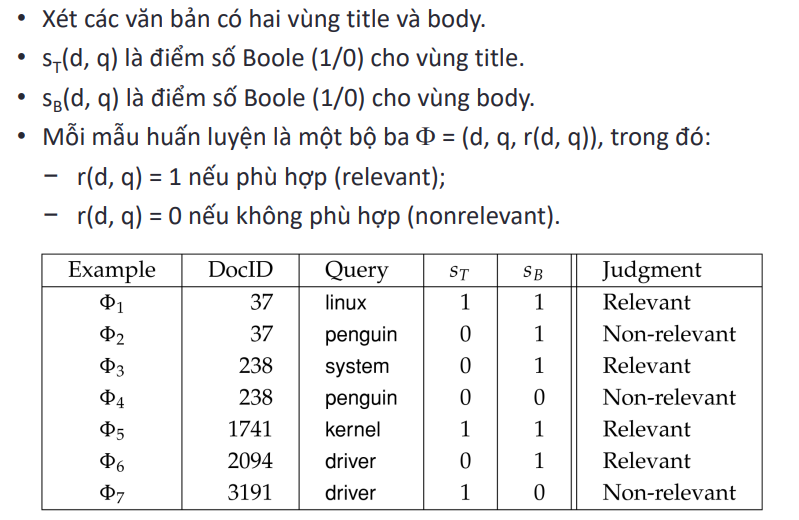
a, month => Khớp

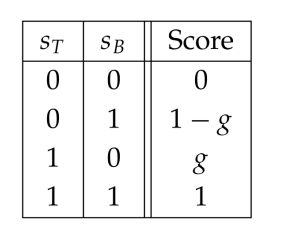
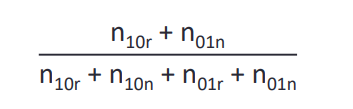
b, moonish => Khớp



|  |  |
| --- | --- |
| **1** | **0.2** |
| **2** | **0.31** |
| **3** | **0.49** |
| **12** | **0.51** |
| **13** | **0.69** |
| **23** | **0.8** |
| **123** | **1** |
| **Rỗng** | **0** |

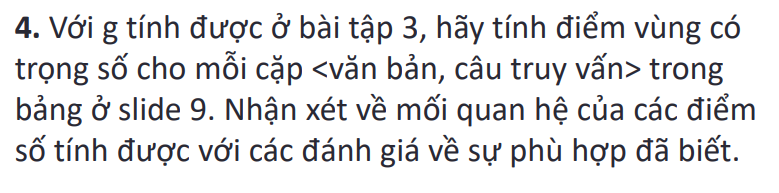


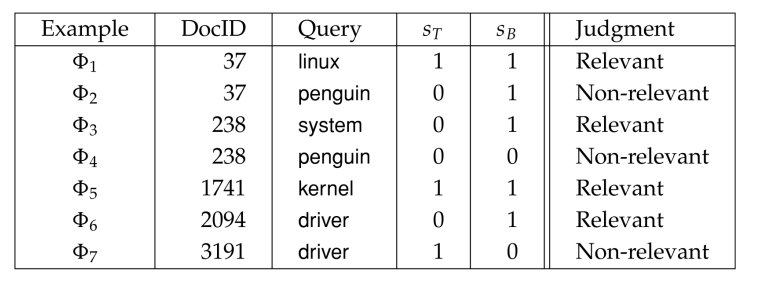




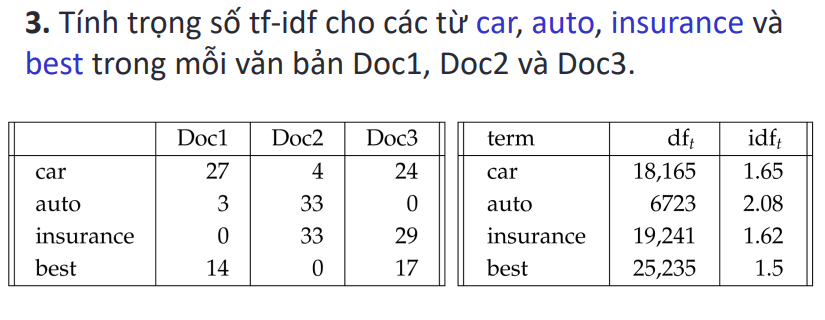
g = = 0.25 (title)

1 - 0.25 = 0.75 (body)





|  |  |  |  |
| --- | --- | --- | --- |
| query | Title | Body | Total |
| linux | 0.25 | 0.75 | 1 |
| penguin | 0 | 0.75 | 0.75 |
| …….. | ……….. | ………. | ………. |



tf-idf = tf . idf

Doc1:

Car: 27 \* 1.65 = 44.55

Auto: 3 \* 2.08 = 6.24

Insurance: 0 \* 1.62 = 0

Best: 14 \* 1.5 = 21

Doc2:

Car: 6.6

Auto: 68.64

Insurance: 53.46

Best: 0

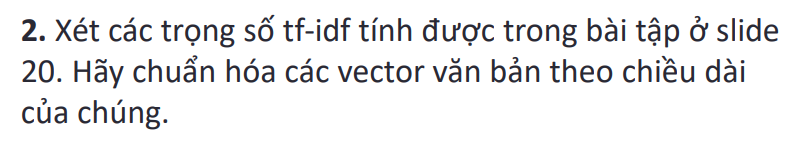
Doc3:

Car: 39.6

Auto: 0

Insurance: 46.98

Best: 25.5



Doc1:

Car: 44.55 => = 0.897

Auto: 6.24 => …..

Insurance: 0 =>0

Best: 21 => ……

