**1.** Vẽ chỉ mục ngược cho tập văn bản sau đây:

Doc1: new home sales top forecasts

Doc2: home sales rise in july

Doc3: increase in home sales in july

Doc4: july new home sales rise

Forecast -> 1

In ->2,3

Increase ->3

July -> 2,3,4

New -> 1,4

Home -> 1,2,3,4

Rise -> 2,4

Sale -> 1,2,3,4

Top -> 1

**2.** Xét tập văn bản sau đây:

Doc1: breakthrough drug for schizophrenia

Doc2: new schizophrenia drug

Doc3: new approach for treatment of schizophrenia

Doc4: new hopes for schizophrenia patients

a. Vẽ ma trận từ-văn bản.

b. Vẽ chỉ mục ngược.

c. Kết quả trả về cho những câu truy vấn sau đây là gì?

1. schizophrenia AND drug

2. for AND NOT (drug OR approach)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Doc1 | Doc2 | Doc3 | Doc4 |
| breakthrough | 1 | 0 | 0 | 0 |
| drug | 1 | 1 | 0 | 0 |
| for | 1 | 0 | 1 | 1 |
| 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 |

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

**schizophrenia AND drug:**

Schizophrenia -> 1,2,3,4

Drug ->1,2

KQ = 1,2

2. For AND NOT (drug OR approach)

Drug ->1,2

Approach -> 3

Kq1 = 1,2,3

For -> 1,3,4

Kq2 = 4

3. Xét câu truy vấn ký tự đại diện fi\*mo\*er. Nếu dùng 2-

gram thì câu truy vấn Boole sinh ra là gì?

$f AND fi AND mo AND er AND r$

4.Xét câu truy vấn ký tự đại diện mon\*h. Nếu dùng 2-gram

và không có bước hậu lọc, hỏi các từ sau đây có khớp với

câu truy vấn Boole sinh ra hay không? Từ nào khớp với câu

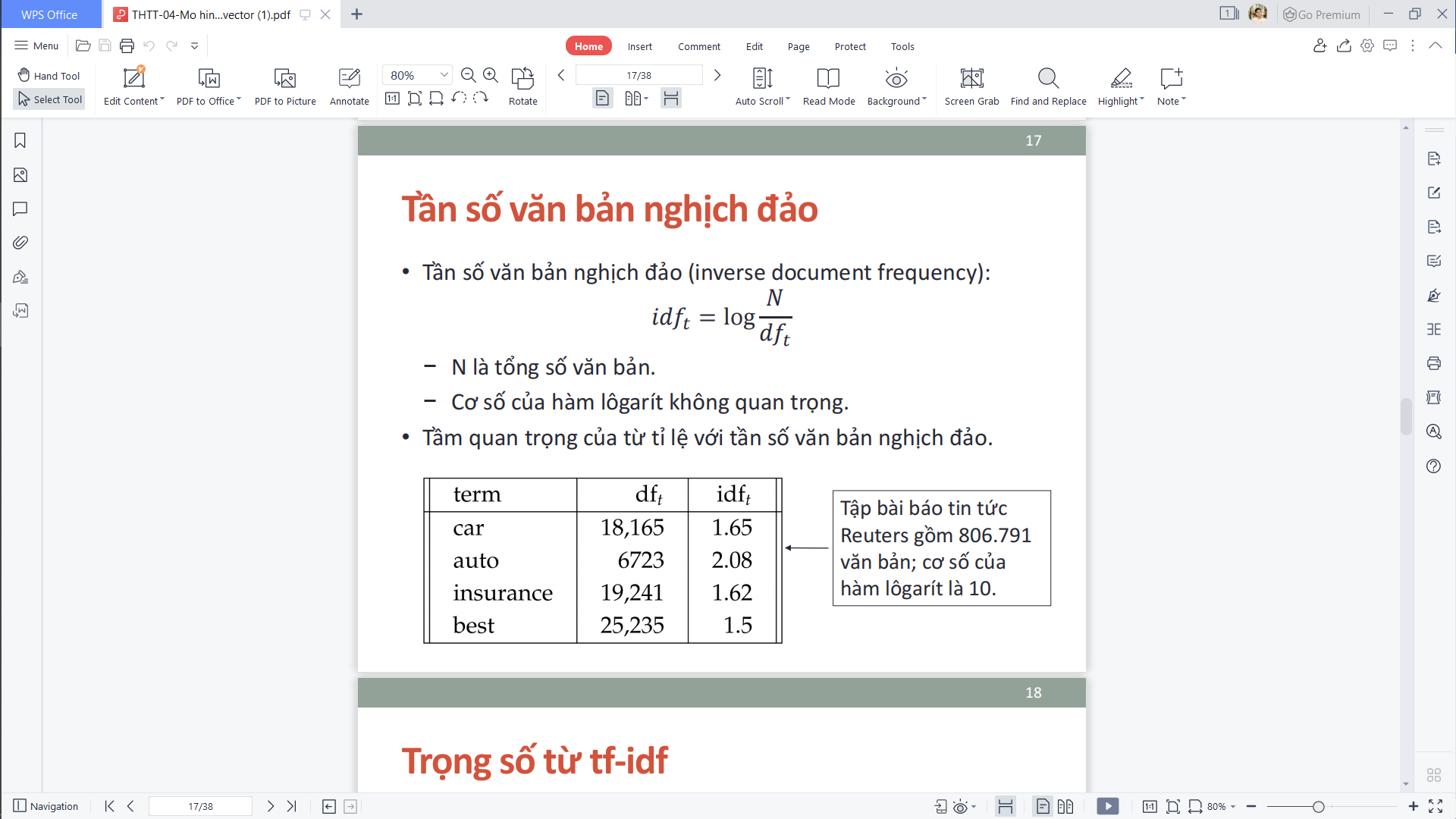
truy vấn gốc?

(a) month

(b) moonish

* Truy vấn gốc: mon\*h
* Truy vấn boole $m And Mo AND on And h$
* Cá a và b thoải mãn boole nhưng b không thỏa mãn câu truy vấn gốc

5. **Tần số văn bản nghịch đảo**

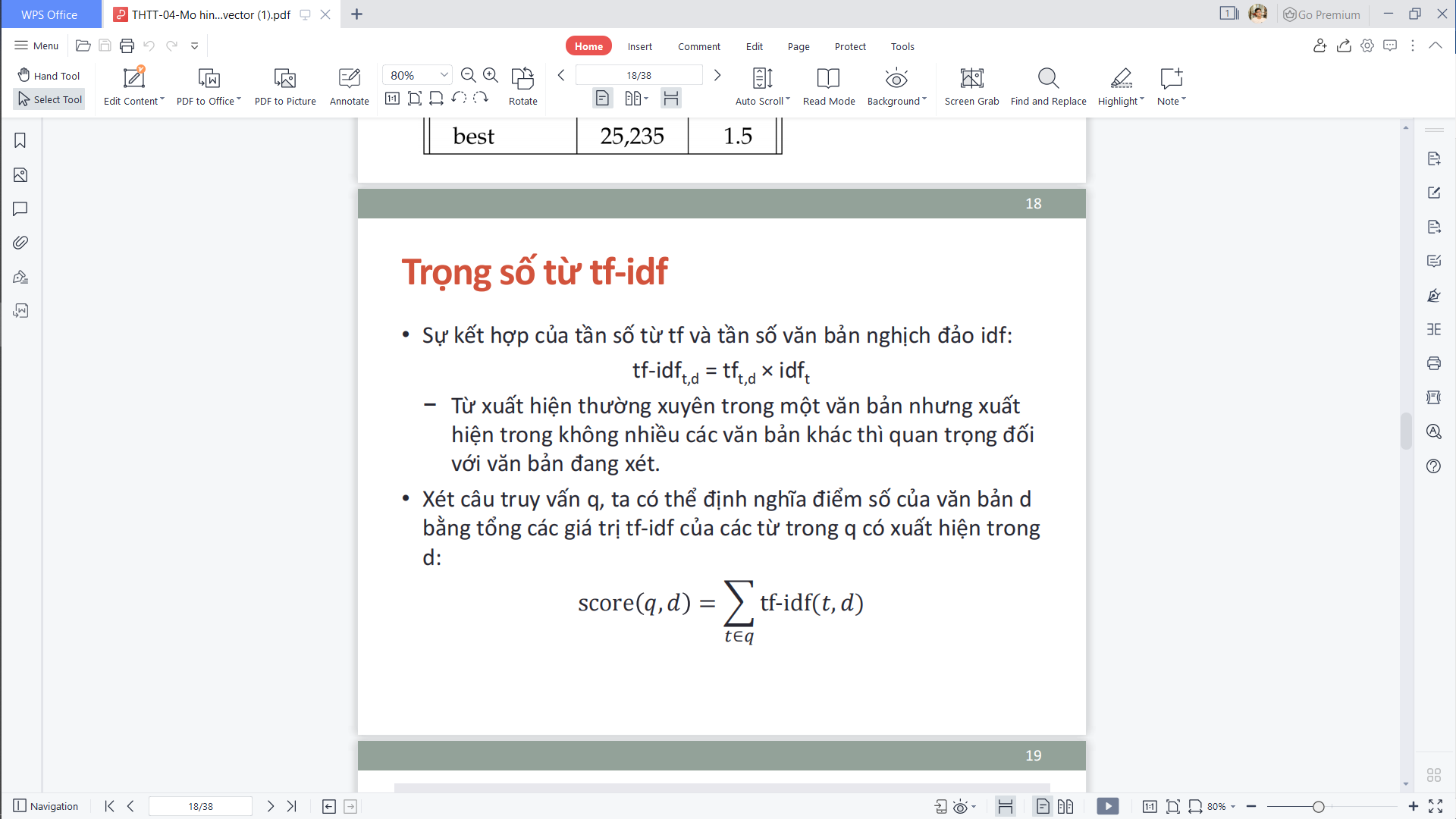


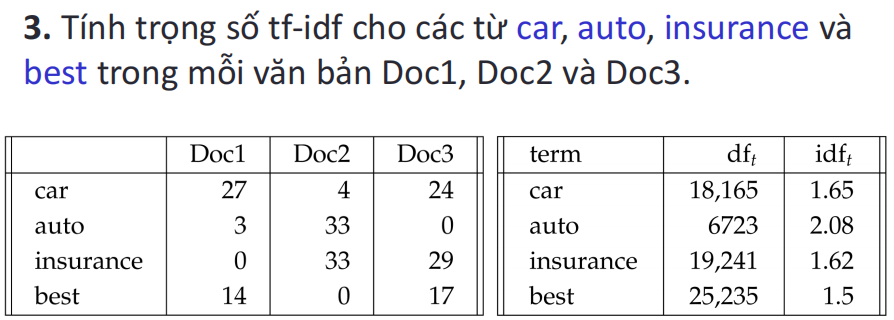
Ký hiệu là tft,d: Số lần từ t xuất hiện trong văn bản d.

Tần số văn bản dft (document frequency) là số văn

bản chứa từ t đang xét.

6. **Trọng số từ tf-idf**





tf-idft,d = tft,d × idft

|  |  |  |  |
| --- | --- | --- | --- |
|  | Doc1 | Doc2 | Doc3 |
| Car | 27 \* 1.65 = 44.55 | 4\*1.65=2.6 | 24\*1,65 = 15.6 |
| Auto |  |  |  |
| Insurance |  |  |  |
| best |  |  |  |