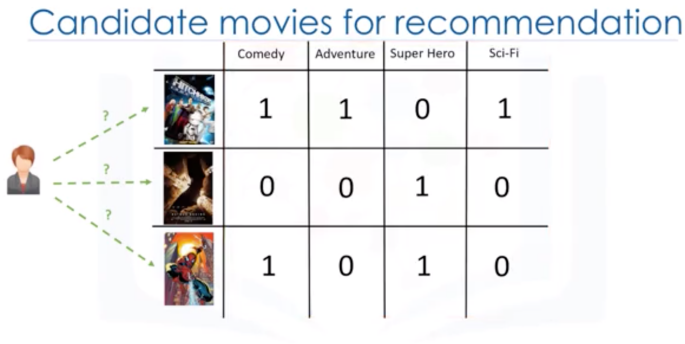
Ví dụ: dữ liệu có cấu trúc



28

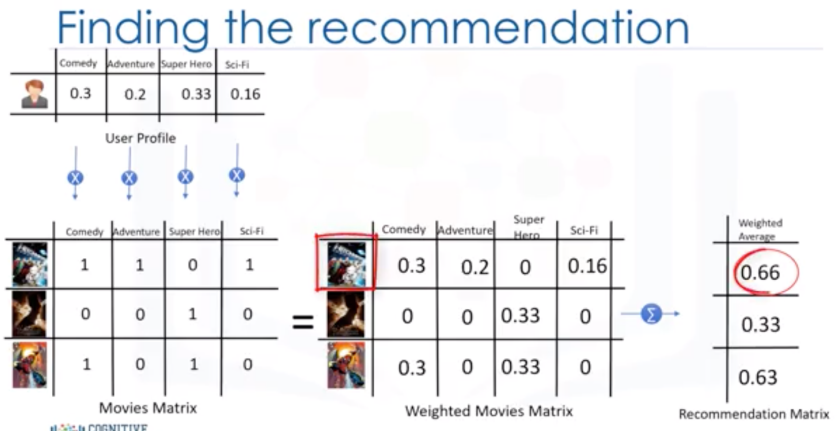
28

Ví dụ: dữ liệu có cấu trúc

29

29

Ví dụ: dữ liệu có cấu trúc



30

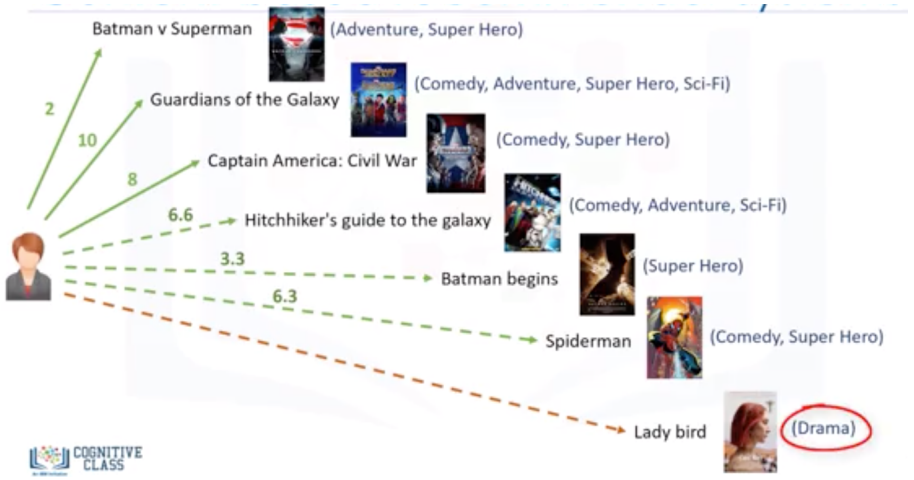
30

Ví dụ: dữ liệu có cấu trúc

31

31

Ví dụ: dữ liệu có cấu trúc



32

32

Ví dụ: dữ liệu không có cấu trúc

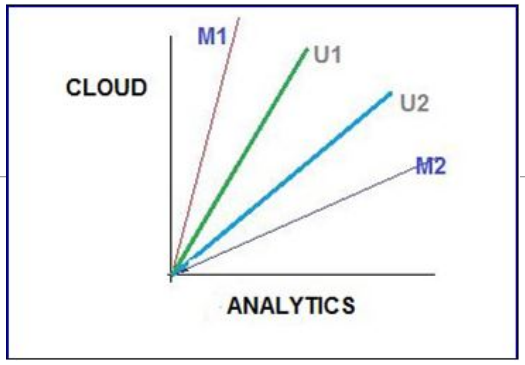
Ví dụ, chúng ta cần tìm tài liệu "***IoT and analytics***" trên Google và danh sách các tài liệu đầu tiền trả về chứa các từ khoá như bảng bên dưới

Tổng các tài liệu tìm kiếm là 1 triệu tài liệu, trong đó có 5.000 tài liệu chưa từ khoá "Analytics", 50.000 tài liệu chứa từ khoá "Data",..

Articles	Analytics	Data	Cloud	Smart	Insight
Article 1	21	24	***0***	2	2
Article 2	24	59	2	1	***0***
Article 3	40	115	8	10	19
Article 4	4	28	5	***0***	1
Article 5	8	48	4	3	4
Article 6	17	49	8	***0***	5
DF	5,000	50,000	10,000	5,00,000	7000

33

33



Mối tương quan giữa 2 từ khoá “cloud” và “analytics”. M1 và M1 là 2 tài lieu, U1 và U2 là 2 user.

M1 liên quan gần với từ khoá “cloud” hơn M2, ngược lại, M2 gần với từ khoá “Analytics” hơn M1.

User U1 thích tài liệu có chứ từ khoá ‘cloud’ hơn ‘analytics’ và ngược lại cho user U2 dựa trên việc tính cosin góc giữa ”user profile” và tài liệu

34

Hệ thống gợi ý dựa trên nội dung

Các bước chính:

- Biểu diễn mỗi “item” dưới dạng **một vector thuộc tính.**
- Gợi ý các “item” **tương tự** các “item” trong quá khứ **của chính người dùng.**

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

35

TF.IDF (Tần số từ khoá/ nghịch đảo tần số văn bản)

Phương pháp dựa trên tần số từ khoá (TF-Term Frequency)

Giá trị của một từ khoá w_{ij} được tính dựa trên tần số xuất hiện của từ khoá trong văn bản. Gọi tf_{ij} là số lần xuất hiện của từ khóa t_i trong văn bản d_j, khi đó có thể chọn cách tính w_{ij} theo một trong ba công thức dưới đây:

$$w_{ij} = tf_{ij}$$

$$w_{ij} = \sqrt{tf_{ij}}$$

$$w_{ij} = 1 + \log(tf_{ij})$$

TF = 3 so với TF = 4 khác rất nhiều so với TF = 10 so với TF = 1000

=> mức độ liên quan của một từ trong tài liệu thường không thể được tính bằng cách đếm đơn giản số lần xuất hiện

36

TF.IDF (Tần số từ khoá/ nghịch đảo tần số văn bản)

Phương pháp dựa trên nghịch đảo tần số văn bản (IDF-Inverse Document Frequency)

Gọi df_i là số lượng văn bản có từ khoá t_i trong tập m văn bản đang xét, thì giá trị trọng số từ w_{ij} được tính bởi công thức:

$$w_{ij} = log\frac{m}{df_j} = log(m) - log(df_i)$$

TF.IDF score : $w_{ij} = TF_{ij} * IDF_i$

Doc profile = set of words with highest TF.IDF scores, together with their scores

37

Ví dụ: dữ liệu không có cấu trúc

Ví dụ, chúng ta cần tìm tài liệu "***IoT and analytics***" trên Google và danh sách các tài liệu đầu tiền trả về chứa các từ khoá như bảng bên dưới

Tổng các tài liệu tìm kiếm là 1 triệu tài liệu, trong đó có 5.000 tài liệu chưa từ khoá "Analytics", 50.000 tài liệu chứa từ khoá "Data",..

Articles	Analytics	Data	Cloud	Smart	Insight
Article 1	21	24	***0***	2	2
Article 2	24	59	2	1	***0***
Article 3	40	115	8	10	19
Article 4	4	28	5	***0***	1
Article 5	8	48	4	3	4
Article 6	17	49	8	***0***	5
DF	5,000	50,000	10,000	5,00,000	7000

38

38

Ví dụ: dữ liệu không có cấu trúc

Tính giá trị TF:

$$w_{ij} = 1 + \log(tf_{ij})$$

Để tính TF cho article 1: :$w_{Article1\_Analytics}$ **= 1 + $\log_{10}$21 = 2.322**.

Articles	Analy
Article 1	
Article 2	
Article 3	
Article 4	
Article 5	
Article 6	
DF	

Articles	Analytics	D
Article 1	2.322219295	
Article 2	2.380211242	
Article 3	2.602059991	
Article 4	1.602059991	
Article 5	1.903089987	
Article 6	2.230448921	

tf_{ij} là số lần xuất hiện của từ khóa **t_i** trong văn bản **d_j**

39

39

Ví dụ: dữ liệu không có cấu trúc

Tính giá trị TF:

$$w_{ij} = 1 + \log(tf_{ij})$$

Để tính TF cho article 1: :$w_{Article1\_Analytics}$ **= 1 + $\log_{10}$21 = 2.322**.

Articles	Analytics	Data	Cloud	Smart	Insight
Article 1	2.322219295	2.380211242	0	1.301029996	1.301029996
Article 2	2.380211242	2.770852012	1.301029996	1	0
Article 3	2.602059991	3.06069784	1.903089987	2	2.278753601
Article 4	1.602059991	2.447158031	1.698970004	0	1
Article 5	1.903089987	2.681241237	1.602059991	1.477121255	1.602059991
Article 6	2.230448921	2.69019608	1.903089987	0	1.698970004

40

40

Ví dụ: dữ liệu không có cấu trúc

Độ dài các vector đại diện cho từng article: Căn bậc hai của tổng các giá trị bình phương của từng thuộc tính

lengthVector $_{Article1}$ =$\sqrt{2.32^2 + 2.38^2 + 0^2 + 1.3^2 + 1.3^2}$

- = 3.8

Articles	Analytics	Data	Cloud	Smart	Insight	Length of Vector
Article 1	2.322219295	2.380211242	0	1.301029996	1.301029996	**3.800456039**
Article 2	2.380211242	2.770852012	1.301029996	1	0	**4.004460697**
Article 3	2.602059991	3.06069784	1.903089987	2	2.278753601	**5.380804488**
Article 4	1.602059991	2.447158031	1.698970004	0	1	**3.527276247**
Article 5	1.903089987	2.681241237	1.602059991	1.477121255	1.602059991	**4.257450611**
Article 6	2.230448921	2.69019608	1.903089987	0	1.698970004	**4.326697114**

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

41

41

Ví dụ: dữ liệu không có cấu trúc

Chuẩn hoá độ dài các vector đại diện cho từng article về 1

Giá trị TF = Giá trị TF / length of Vector,

Article 1 $_{\text{Analytics}}$ = 2.32/3.8 = 0.61

Articles	Analytics	Data	Cloud	Smart	Insight	Length of Vector
Article 1	2.322219295	2.380211242	0	1.301029996	1.301029996	**3.800456039**
Article 2	2.380211242	2.770852012	1.301029996	1	0	**4.004460697**
Article 3	2.602059991	3.06069784	1.903089987	2	2.278753601	**5.380804488**

Articles	Analytics	Data	Cloud	Smart	Insight	Sum of Normalized Lengths
Article 1	0.61103701	0.626296217	0	0.342335231	0.342335231	**1**
Article 2	0.594389962	0.691941368	0.324895184	0.249721517	0	**1**
Article 3	0.483581962	0.568817887	0.353681311	0.371691632	0.423496822	**1**
Article 4	0.454191812	0.693781224	0.481666273	0	0.283504872	**1**
Article 5	0.447002246	0.62977624	0.376295614	0.34694971	0.376295614	**1**
Article 6	0.51550845	0.621766675	0.439848211	0	0.392671352	**1**

42

42

Ví dụ: dữ liệu không có cấu trúc

Tính độ tương tự giữa 2 article dựa vào công thức cosin

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

Articles	Analytics	Data	Cloud	Smart	Insight	Sum of Normalized Lengths
Article 1	0.61103701	0.626296217	0	0.342335231	0.342335231	**1**
Article 2	0.594389962	0.691941368	0.324895184	0.249721517	0	**1**
Article 3	0.483581962	0.568817887	0.353681311	0.371691632	0.423496822	**1**
Article 4	0.454191812	0.693781224	0.481666273	0	0.283504872	**1**
Article 5	0.447002246	0.62977624	0.376295614	0.34694971	0.376295614	**1**
Article 6	0.51550845	0.621766675	0.439848211	0	0.392671352	**1**

Cos(Article1-Article2) = 0.61*0.59+0.62*0.69+0*0.32+0.34*0.25+0.34*0 = 0.87

Cos(Article1-Article3) = 0.61*0.48+0.62*0.56+0*0.35+0.34*0.37 + 0.34*0.42 = 0.91

43

43

Ví dụ: dữ liệu không có cấu trúc

Tính độ tương tự giữa 2 article dựa vào công thức cosin

Articles	Analytics	Data	Cloud	Smart	Insight	Sum of Normalized Lengths
Article 1	0.61103701	0.626296217	0	0.342335231	0.342335231	1
Article 2	0.594389962	0.691941368	0.324895184	0.249721517	0	1
Article 3	0.483581962	0.568817887	0.353681311	0.371691632	0.423496822	1
Article 4	0.454191812	0.693781224	0.481666273	0	0.283504872	1
Article 5	0.447002246	0.62977624	0.376295614	0.34694971	0.376295614	1
Article 6	0.51550845	0.621766675	0.439848211	0	0.392671352	1

Cos(Article1-Article2) = 0.61*0.59+0.62*0.69+0*0.32+0.34*0.25+0.34*0 = 0.87

Cos(Article1-Article3) = 0.61*0.48+0.62*0.56+0*0.35+0.34*0.37 + 0.34*0.42 = 0.91

Thực hiện tương tự với các article còn lại để tìm ra item tương đồng nhất với article 1

44

44

Thực hiện ví dụ trên sử dụng chỉ số TF.IDF

Tính giá trị IDF:

$$w_{ij} = log\frac{m}{df_j}$$

df_i là số lượng văn bản có từ khoá **t_i** trong tập **m** văn bản đang xét

Để tính IDF cho "smart": m= toàn bộ tập dữ liệu 1.000.000

$IDF_{smart} = log\frac{m}{df_j} = \frac{1.000.000}{500.000} = 0.30$

Articles	Analytics	Data	Cloud	Smart	Insight
DF	5,000	50,000	10,000	500.000	7000
IDF	2.301029996	1.301029996	2	0.301029996	2.15490196

45

45

Thực hiện ví dụ trên sử dụng chỉ số TF.IDF

Bảng giá trị TF

Articles	Analytics	Data	Cloud	Smart	Insight
Article 1	2.322219295	2.380211242	0	1.301029996	1.301029996
Article 2	2.380211242	2.770852012	1.301029996	1	0
Article 3	2.602059991	3.06069784	1.903089987	2	2.278753601
Article 4	1.602059991	2.447158031	1.698970004	0	1
Article 5	1.903089987	2.681241237	1.602059991	1.477121255	1.602059991
Article 6	2.230448921	2.69019608	1.903089987	0	1.698970004

Bảng giá trị IDF

Articles	Analytics	Data	Cloud	Smart	Insight
IDF	2.301029996	1.301029996	2	0.301029996	2.15490196

Articles	Analytics	Data	…
Article1	2.32 *2.30 = 5.34		
Article 2	2.38*2.3 = 5.47		

$w_{ij} = TF_{ij} * IDF_i$

46

46

Thực hiện ví dụ trên sử dụng chỉ số TF.IDF

Các em tiếp tục làm tương tự như ví dụ trên để tìm ra item tương tự article 1 khi sử dụng chỉ số TF.IDF

47

47

Phương pháp dựa trên nội dung

- Không cần dữ liệu của người dùng khác
- Có thể gợi ý được sản phẩm mới cũng như không phổ biến
- Có thể đưa ra gợi ý cho người dùng có thị hiếu độc đáo
- Có khả ngăn cung cấp sự giải thích cho việc đưa ra gợi ý (dựa trên content-features)

48

48