I. Consider the following documents **D1-D4** using 8 different terms:

D1 = {breakthrough drug schizophrenia}

D2 = {new schizophrenia drug}

D3 = {new approach treatment schizophrenia}

D4 = {new hope schizophrenia patient}

Fill in the term-document incidence matrix for this document collection.

schizophrenia AND drug

1111 AND 1100 = 1100

↳ so   D1 & D2

new AND NOT (drug OR approach)

0111 AND $\sim(1100$ OR $0010)$ =

= 0111 AND $\sim(1110)$ =

= 0111 AND 0001 =

= 0001

↳ so D4

	D1	D2	D3	D4
approach	0	0	1	0
breakthrough	1	0	0	0
drug	1	1	0	0
hope	0	0	0	1
new	0	1	1	1
patient	0	0	0	1
schizophrenia	1	1	1	1
treatment	0	0	1	0

What are the results returned for the below Boolean queries:

- schizophrenia AND drug   Answer: D1, D2
- new AND NOT(drug OR approach)   Answer: D4

II. Given the following four sessions: {D1 D2 D3}, {D1 D3}, {D1 D3}, {D2 D3}, answer the questions related to using the Markov chain for mining navigational patterns.
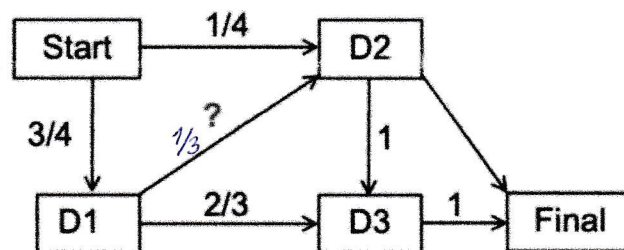
- Sessions:

D1 D2 D3

D1 D3

D1 D3

D2 D3

What is P(D1→D2)?   Answer: $1/3$

What is the probability of P(Start→D1→D3)?   Answer: $3/4 \cdot 2/3 = 1/2$

What is the probability of P(D3|D1)?   Answer:

D1 → D3 + D1 → D2 → D3

$\frac{2}{3} + \frac{1}{3} \cdot 1 = 1$