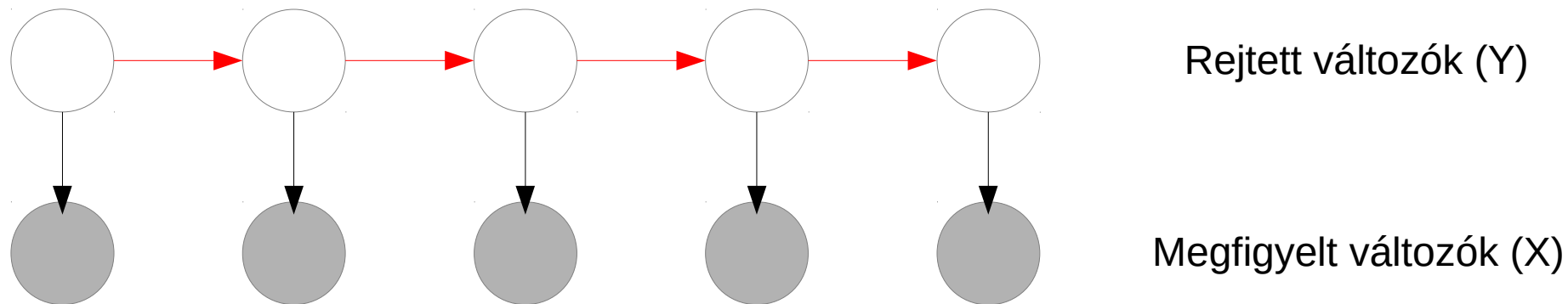




Brown klaszterek

Rejtett Markov Modell (HMM)



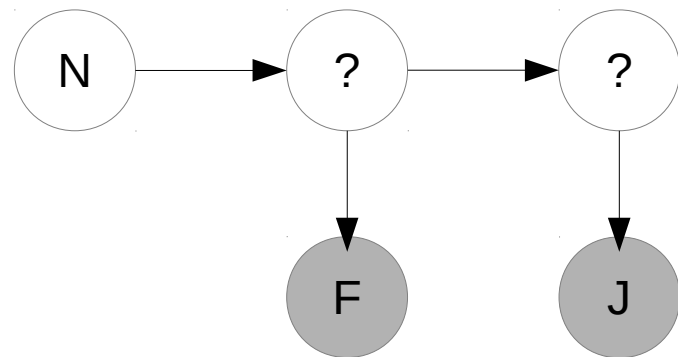
- Valószínűségi modell, amely az **állapotátmenet**, - és az emissziós valószínűségek~~ből~~ áll
 - Markov feltevés: az adott pillanat rejtett állapota csak a megelőzőtől függ
 - Segítségével a megfigyelt változókat leginkább megmagyarázni képes rejtett változókat határozhatjuk meg

HMM példa

- Tudjuk, hogy 2 napja napos idő volt, és a következő két napban forró, majd jeges teát fogyasztottunk. Milyen időjárás volt a legvalószínűbb az elmúlt 2 nap folyamán?

	Napos	Esős
Napos	0,8	0,2
Esős	0,3	0,7

	Forró	Jeges
Napos	0,1	0,9
Esős	0,6	0,4

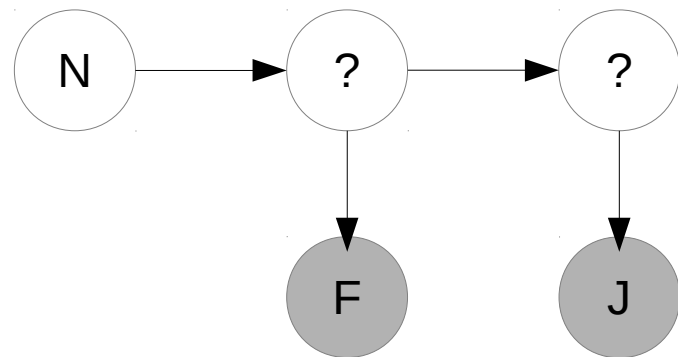


HMM példa

- Tudjuk, hogy 2 napja napos idő volt, és a következő két napban forró, majd jeges teát fogyasztottunk. Milyen időjárás volt a legvalószínűbb az elmúlt 2 nap folyamán?
 - $NN \rightarrow (0,8 * 0,1) * (0,8 * 0,9) = 0,0576$

	Napos	Esős
Napos	0,8	0,2
Esős	0,3	0,7

	Forró	Jeges
Napos	0,1	0,9
Esős	0,6	0,4



HMM példa

- Tudjuk, hogy 2 napja napos idő volt, és a következő két napban forró, majd jeges teát fogyasztottunk. Milyen időjárás volt a legvalószínűbb az elmúlt 2 nap folyamán?

– $NN \rightarrow (0,8 * 0,1) * (0,8 * 0,9) = 0,0576$

– $NE \rightarrow (0,8 * 0,1) * (0,2 * 0,4) = 0,0064$

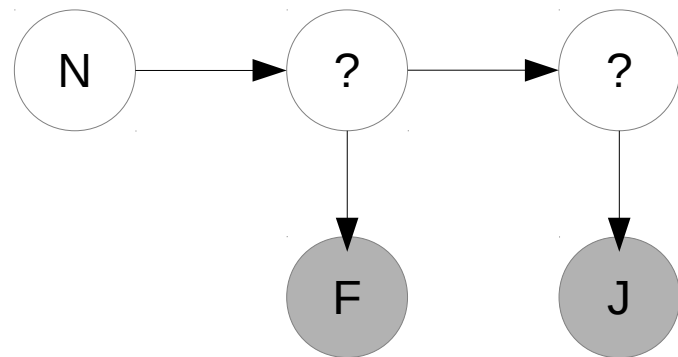
– $EN \rightarrow (0,2 * 0,6) * (0,7 * 0,9) = 0,0324$

– $EE \rightarrow (0,2 * 0,6) * (0,7 * 0,4) = 0,0336$

$\Sigma = 0,18$

	Napos	Esős
Napos	0,8	0,2
Esős	0,3	0,7

	Forró	Jeges
Napos	0,1	0,9
Esős	0,6	0,4



HMM feladatok – Tanítás

- Cél: a tanítószekvencia megfigyelését legvalószínűbbé tevő paraméterek meghatározása
 - Ha a rejtett változók ismertek (lennének), akkor egyszerű maximum likelihood módon elvégezhető
 - A rejtett változók azonban nem (legfeljebb részlegesen) ismertek
 - A szekvencia hosszában (l) és a lehetséges rejtett állapotok számában (H) exponenciálisan sok (H^l) lehetséges rejtett állapot szekvencia

HMM feladatok – Tanítás

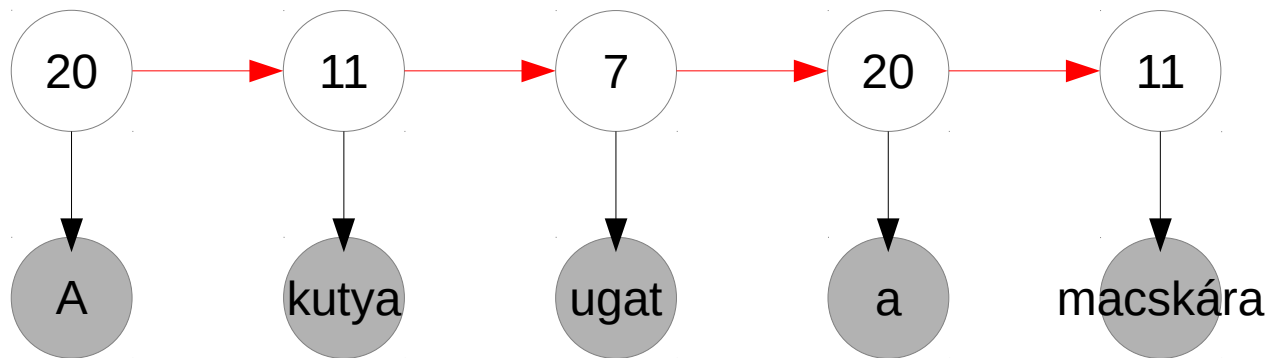
- Cél: a tanítószekvencia megfigyelését legvalószínűbbé tevő paraméterek meghatározása
 - Ha a rejtett változók ismertek (lennének), akkor egyszerű maximum likelihood módon elvégezhető
 - A rejtett változók azonban nem (legfeljebb részlegesen) ismertek
 - Expectation Maximization (EM) algoritmus
 - A szekvencia hosszában (l) és a lehetséges rejtett állapotok számában (H) exponenciálisan sok (H^l) lehetséges rejtett állapot szekvencia
 - Dinamikus programozással kiküszöbölhető a H^l szekvencia explicit kiszámítása

HMM feladatok – Inferencia

- Cél: a modell paramétereinek alapján egy megfigyelési sorozatot legjobban magyarázó rejtett állapotsorozat meghatározása
 - A szekvencia hosszában (l) és a lehetséges rejtett állapotok számában (H) ugyancsak exponenciálisan sok (H^l) lehetséges magyarázó rejtett állapot szekvencia
 - A tanítás során használthoz hasonló dinamikus programozási megoldás

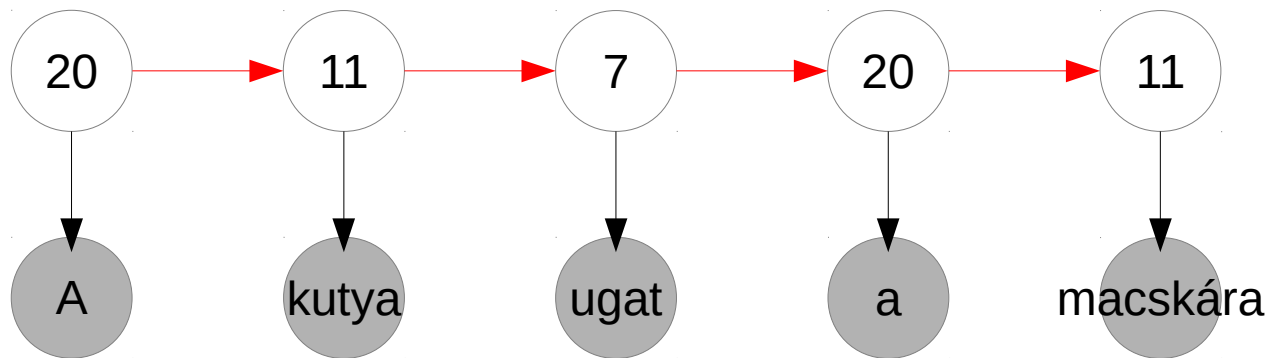
Brown klaszterezés

- Tegyük fel, hogy minden egyes korpuszban megfigyelt szót egy **rejtett** szóosztályok “generálják”
 - Pl. a {macska, kutya, egér, ...} szavakat egy adott (állatokhoz kötődő dolgokat összefogó) klaszter generálja



Brown klaszterezés

- Tegyük fel, hogy minden egyes korpuszban megfigyelt szót egy **rejtett** szóosztályok “generálják”
 - Pl. a {macska, kutya, egér, ...} szavakat egy adott (állatokhoz kötődő dolgokat összefogó) klaszter generálja



Lényegében egy HMM-el van dolgunk!

Példa Brown klaszterek

- Magyar Twitterről Percy Liang implementációjával kinyerve
cluster path 0110110010 **cluster path 110010111010**

490 words, 14,612 tokens [freq](#) [alpha](#) [suffix](#)

Words in frequency order

1	ülföld	1,196
2	megnövekedett	930
3	áció	461
4	ép	424
5	ülföldifoci	337
6	ácsony	328
7	ékelyföld	327
8	átokközt	314
9	ánsok	307

37 words, 13,219 tokens [freq](#) [alpha](#) [suffix](#)

Words in frequency order

1	akit	3,231
2	amire	1,889
3	amiért	1,650
4	amiket	1,622
5	akivel	1,030
6	akiket	932
7	amiről	799
8	akikkel	309
9	akire	289

Példa Brown klaszterek

- Magyar Twitterről Percy Liang implementációjával kinyerve
cluster path 0110110010 **cluster path 110010111010**

490 words, 14,612 tokens [freq](#) [alpha](#) [suffix](#)

Words in frequency order

1	ülföld	1,196
2	megnövekedett	930
3	áció	461
4	ép	424
5	ülföldifoci	337
6	ácsony	328
7	ékelyföld	327

^[01101011110](#) (522)

^[011010111110](#) (400)

^[011010111111](#) (400)

37 words, 13,219 tokens [freq](#) [alpha](#) [suffix](#)

Words in frequency order

1	akit	3,231
2	amire	1,889
3	amiért	1,650
4	amiket	1,622
5	akivel	1,030
6	akiket	932
7	amiről	799

