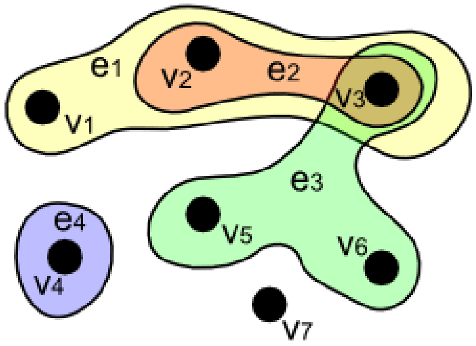
A számítógépes szemantika alapjai

A szemantika fő megközelítései

- 3 nagy iskola
  - A szemantika logikai formulákkal történő leírása

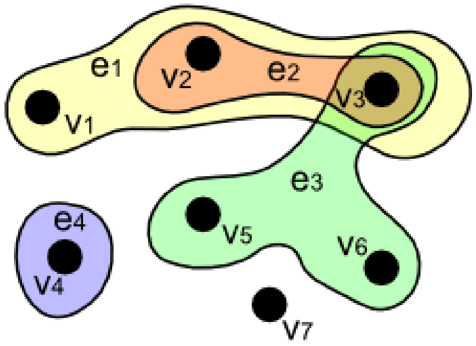
A szemantika fő megközelítései

- 3 nagy iskola
  - A szemantika logikai formulákkal történő leírása
  - Tudásbázisok létrehozása (hiper)gráfok segítségével
    - Pl. Wordnet, ConceptNet, Microsoft Concept Graph, Babelnet, (Open)Cyc, ...



A szemantika fő megközelítései

- 3 nagy iskola
  - A szemantika logikai formulákkal történő leírása
  - Tudásbázisok létrehozása (hiper)gráfok segítségével
    - Pl. Wordnet, ConceptNet, Microsoft Concept Graph, Babelnet, (Open)Cyc, ...



  - **Szavak vektorokkal történő jellemzése**

Montague nyelvtan

- “There is in my opinion no important theoretical difference between natural languages and the artificial languages of logicians” (Universal Grammar 1970)
- A jelentést logikai formulákkal ragadják meg
- Kategóriákra alapozó nyelvtan
- Felteszi a jeletés rekurzív és kompozicionális voltát
- “Colorless green ideas sleep furiously.”
- Lásd még: Frege, Russell, Tarski munkássága

A disztribúciós hipotézis (Firth)

- “You shall know a word by the company it keeps” (Firth, 1957)
  - Az ötlet már az 1935-ös *The technique of semantics* című munkájában is fellelhető

> Secondly, the complete meaning of a word is always contextual, and no study of meaning apart from a complete context can be taken seriously.

- Lásd még: Zellig Harris, Charles Osgood (szemantikus differenciál)

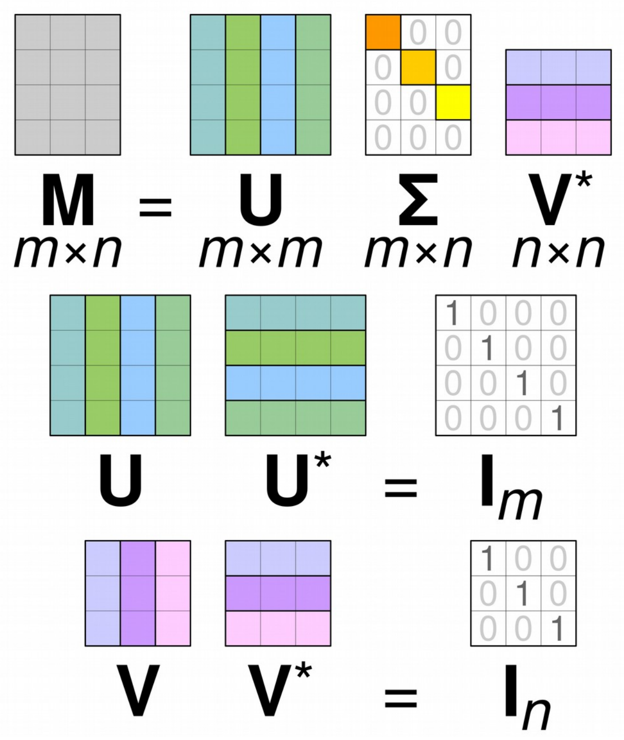
Szavak mint vektorok

- Az ún. term-dokumentum mátrix maga fölfogható a szavak egy reprezentációjaként
  - A mátrix ij eleme megmondja, hogy az i szó a j dokumentumban hányszor fordult elő

	d1	d2	...	dn
w1=korong	3	4	...	
w2=ütő	2	0	...	1
w3=hazafutás	0	0	...	5
...			...	
wm	1	2	...	1

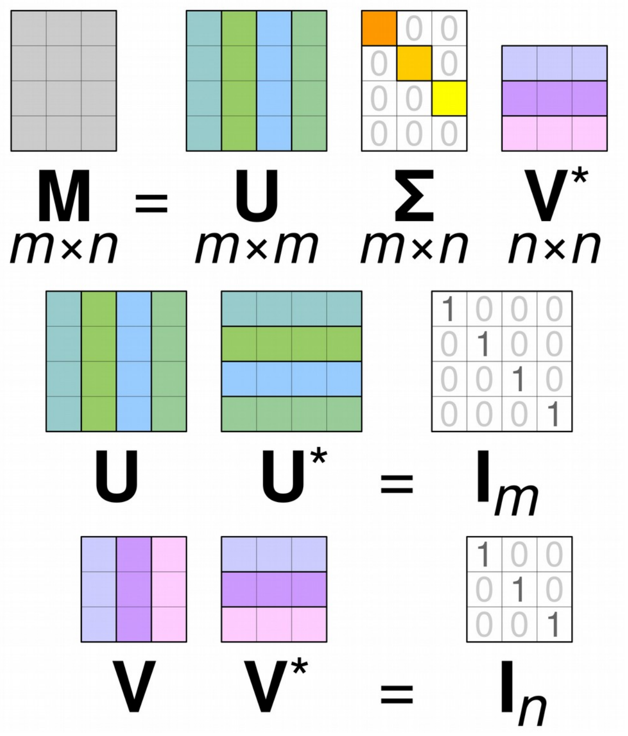
Szinguláris érték felbontás (SVD)

- Bármely mátrix fölírható U*Σ*V' szorzat alakban
  - Ahol U, V' ortogonális, Σ diagonális



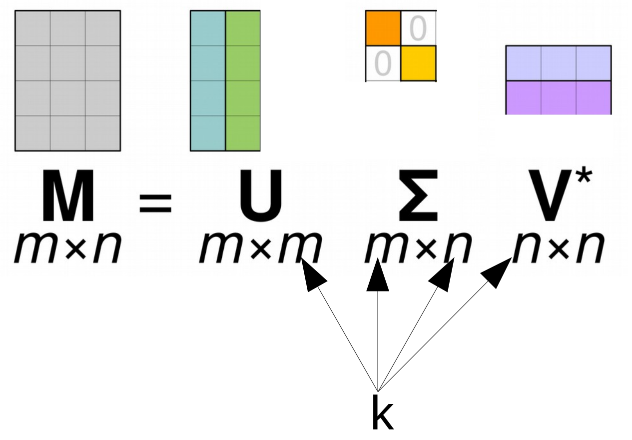
Szinguláriserték felbontás (SVD)

- Bármely mátrix fölírható U*Σ*V' szorzat alakban
  - Ahol U, V' ortogonális, Σ diagonális
- "Csonkolt"-SVD (truncated-SVD)
  - U, V mátrixokból $k<\min(m,n)$ oszlopot tartsunk csupán meg
  - Σ-nek hagyjuk el a k*k-n "felüli" részét



Szinguláriséérték felbontás (SVD)

- Bármely mátrix fölírható U*Σ*V’ szorzat alakban
  - Ahol U, V’ ortogonális, Σ diagonális
- “Csonkolt”-SVD (truncated-SVD)
  - U, V mátrixokból k<min(m,n) oszlopot tartsunk csupán meg
  - Σ-nek hagyjuk el a k*k-n “felüli” részét
  - Egyfajta tömörítés: M legjobb k rangú közelítését kapjuk így meg



Látens Szemantikus Indexelés (Deerwester, S., et al, 1988)

- Az M mátrix legyen a term-dokumentum mátrix
  - U mint term–látens téma mátrix
  - V’ mint látens téma–dokumentum mátrix
  - A látens téma tekinthető a jelentéscsoportoknak

Az input mátrix súlyozása

- Gyakori, de érdektelen szavak → nyers gyakoriságok helyett alkalmazzunk súlyozást (pl. tf–idf)
  - tf: (dokumentumon belüli) term gyakoriság
  - Idf: invertált dokumentum frekvencia log(N/df(t))
    - N a korpusz dokumentumainak száma
    - df(t): azon dokumentumok száma, melyben t term előfordul

Az input mátrix súlyozása

- Gyakori, de érdektelen szavak → nyers gyakoriságok helyett alkalmazzunk súlyozást (pl. tf–idf)
  - tf: (dokumentumon belüli) term gyakoriság
  - Idf: invertált dokumentum frekvencia log(N/df(t))
    - N a korpusz dokumentumainak száma
    - df(t): azon dokumentumok száma, melyben t term előfordul
  - Pl. ha t term az i dokumentumban 3szor szerepel, egyébként pedig a korpusz minden negyedik dokumentumában található meg, akkor

    tf–idf(t,i) = 3 * log(4) = 6

SVD kookkurencia mátrixon

- Ugyanaz, mint eddig, csak a term-dokumentum mátrix helyett a kookkurencia (együttelőfordulási) mátrixon dolgozzunk
  - A mátrix egy ij eleme magadja, hogy az i szó környezetében j szó hányszor fordul elő
    - A term–dokumentum mátrixszal ellentétben itt egy négyzetes termek száma*termek száma mátrixszal van dolgunk
  - Fontos hiperparaméter w a figyelembe vett környezet mérete
    - Kicsi w: inkább szintaktikus kapcsolatok
    - Nagy w: jobban szemantika

Pointwise Mutual Information

- A nyers gyakoriságokat itt is szokás transzformálni (pl. PMI)
  - PMI(x,y) = log(P(x,y)/(P(x)*P(y)))
- Két esemény együttes valószínűsége hogy viszonyul marginálisaik szorzatához
  - Marginálisuk szorzat = együttes valószínűségük, **amennyiben függetlenek**

Pointwise Mutual Information

- A nyers gyakoriságokat itt is szokás transzformálni (pl. PMI)
  - PMI(x,y) = log(P(x,y)/(P(x)*P(y)))
- Két esemény együttes valószínűsége hogy viszonyul marginálisaik szorzatához
  - Marginálisuk szorzat = együttes valószínűségük, **amennyiben függetlenek**

	kutya	olvas	ugat
kutya	0	1	8
olvas	1	2	0
ugat	8	0	0

Pointwise Mutual Information

- A nyers gyakoriságokat itt is szokás transzformálni (pl. PMI)
  - PMI(x,y) = log(P(x,y)/(P(x)*P(y)))
- Két esemény együttes valószínűsége hogy viszonyul marginálisaik szorzatához
  - Marginálisuk szorzat = együttes valószínűségük, **amennyiben függetlenek**
    - P(kutya,ugat)=8/20
    - P(kutya)=9/20, P(ugat)=8/20
    - PMI(kutya, ugat)=log(20/9) ≈ 1.15

	kutya	olvas	ugat
kutya	0	1	8
olvas	1	2	0
ugat	8	0	0

Pointwise Mutual Information variánsok

- Pozitív PMI (PPMI)
  - Motiváció: a negatív értékek nem igazán érdekesek
  - PPMI(x,y) = max(0, PMI(x,y))

Pointwise Mutual Information variánsok

- Pozitív PMI (PPMI)
  - Motiváció: a negatív értékek nem igazán érdekesek
  - PPMI(x,y) = max(0, PMI(x,y))
- Normalizált (P)PMI
  - Motiváció: a (P)PMI mutatónak kedveznek a kevés előfordulással rendelkező események
    - Ha x és y csak egymással fordul elő, akkor PMI(x,y)=–log(P(x,y)), ami ritka (x,y) eseménypárosra nagyon magas értéket jelent!
  - Ezt kompenzálandó, a kapott értéket osszuk el –log(P(x,y))-nal
    - -1 (ha $P(x,y) \rightarrow 0$) és 1 (ha P(x)=P(x,y)=P(y)) közé szorítjuk ezzel

PMI és az alacsony előfordulások

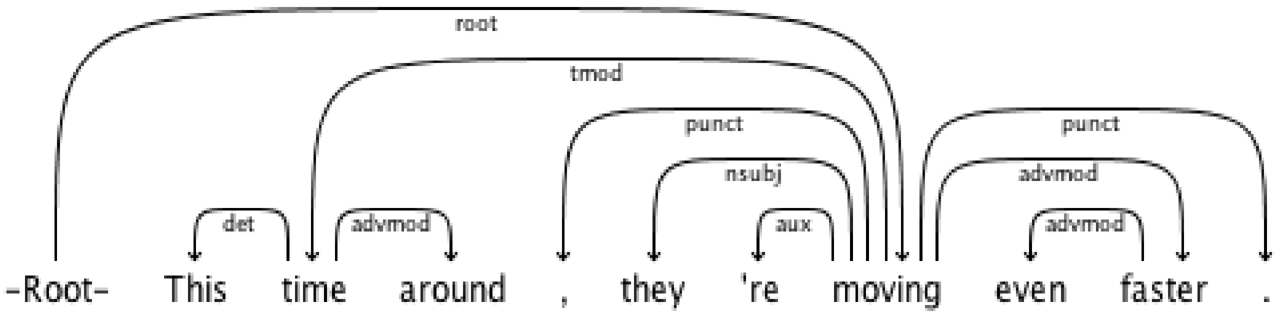
- A PMI értékek “javítására” hallucináljunk megfigyeléseket
  - Laplace-simítás: a tényleges megfigyelések értékeit növeljük 1-el
    - Ad hoc megoldásnak tűnik, de nem is annyira az (lásd: multinomiális eloszlás Dirichlet priorral történő Maximum A Posteriori becslése)

PMI és az alacsony előfordulások

- A PMI értékek “javítására” hallucináljunk megfigyeléseket
  - Laplace-simítás: a tényleges megfigyelések értékeit növeljük 1-el
    - Ad hoc megoldásnak tűnik, de nem is annyira az (lásd: multinomiális eloszlás Dirichlet priorral történő Maximum A Posteriori becslése)
- A tényleges megfigyeléseket emeljük valamijen $x<1$ hatványra, és így normalizáljunk
  - A magas értékek ezt jobban megsínylik
    - Pl. $x=.75$ választása esetén $[0.01, 0.05, 0.94] \rightarrow [0.029, 0.097, 0.874]$

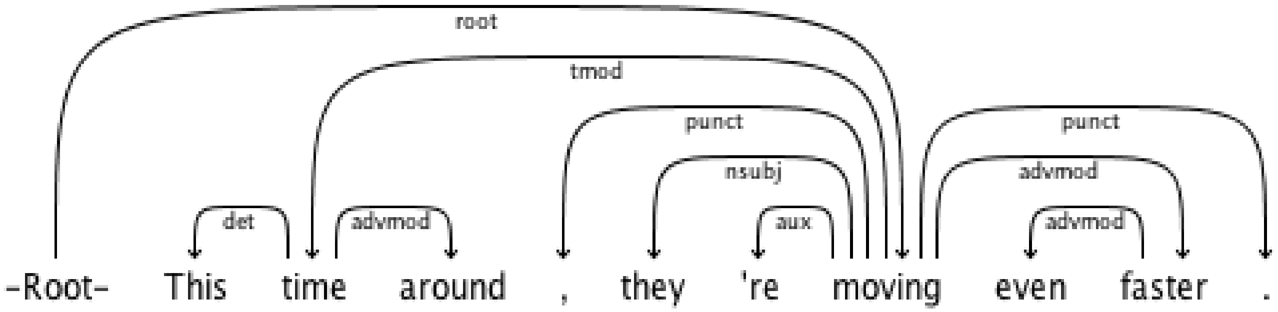
Kookkurenciamátrix kiterjesztése

- A környezetbe bevehetjük a nyelvtani stuktúrát is
  - A mátrix sorai továbbra is a szavak, a kontextusokat jelölő oszlopok azonban (szóalak–reláció) párosok lesznek
  - Egy lehetőség pl. ha dependenciaelemzést használunk



Kookkurenciamátrix kiterjesztése

- A környezetbe bevehetjük a nyelvtani stuktúrát is
  - A mátrix sorai továbbra is a szavak, a kontextusokat jelölő oszlopok azonban (szóalak–reláció) párosok lesznek
  - Egy lehetőség pl. ha dependenciaelemzést használunk



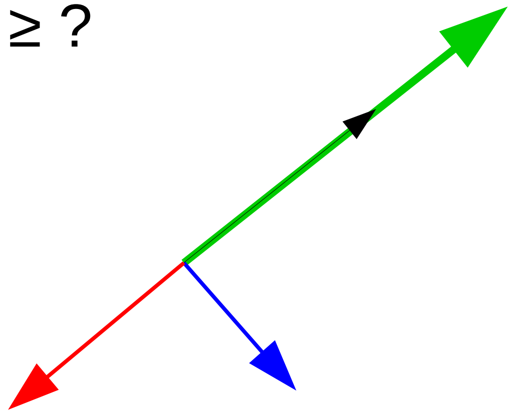
- Side note: Scientists discovered a new animal from space.
  - PP attachment problémaköre

Szóvektorok viselkedése

- Az előző módszerekkel szavakhoz vektorokat tudunk társítani
  - Hasonló jelentésű szópár → hasonlóan irányba mutató vektorok
  - Pontszorzat: **v**’***w** = Σ vi * wi
    - Pl. **v**’=[3, 1], **w**’= [5, 2] esetén **v**’***w** = ?

Szóvektorok viselkedése

- Az előző módszerekkel szavakhoz vektorokat tudunk társítani
  - Hasonló jelentésű szópár → hasonlóan irányba mutató vektorok
  - Pontszorzat: **v**’***w** = Σ vi * wi
    - Pl. **v**’=[3, 1], **w**’= [5, 2] esetén **v**’***w** = 17
  - Koszinusz hasonlóság: ? ≥ **v**’***w** / (||**v**|| * ||**w**||) ≥ ?



Szóvektorok viselkedése

- Az előző módszerekkel szavakhoz vektorokat tudunk társítani
  - Hasonló jelentésű szópár → hasonlóan irányba mutató vektorok
  - Pontszorzat: **v'*w** = Σ vi * wi
    - Pl. **v**'=[3, 1], **w**'= [5, 2] esetén **v'*w** = 17
  - Koszinusz hasonlóság: 1 ≥ **v'*w** / (||**v**|| * ||**w**||) ≥ -1
    - 1 esetén orientációjuk megegyezik
    - 0 esetén ortogonálisak
    - -1 esetén ellentétes irányba mutatnak