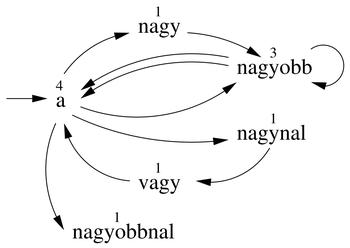
A szöveg mint skálafüggetlen hálózat

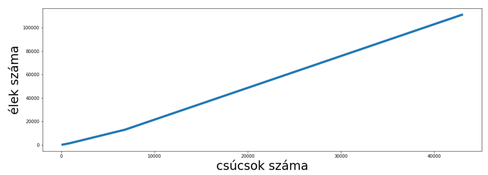
Makrai Márton és Sass Bálint
MTA Nyelvtudományi Intézet
{makrai.marton,sass.balint}@nytud.mta.hu

Hatványeloszlás, szavak, élek

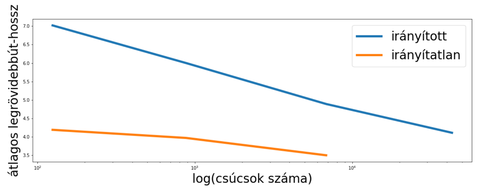
- szógyakoriságok (Zipf, 1935)
- skálafüggetlen gráf (Barabási and Albert, 1999)
- most: irányított gráf súlyozott élekkel bigramgyakorliságokból



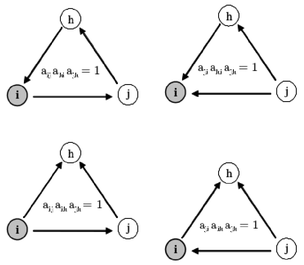
1. ábra „A nagyobb nagyobb a nagynál vagy a nagy nagyobb a nagyobbnál." példamondat ábrázolása. A dupla nyilat ábrázolhatjuk egy 2-es súllyal bíró szimpla nyíllal is.



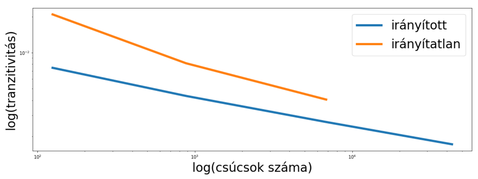
Irányított kisvilág globálisan és. . .



. . . lokálisan



2. ábra Irányított klaszterezési együttható (Fagiolo, 2007)



Különcség (*eccentricity*)

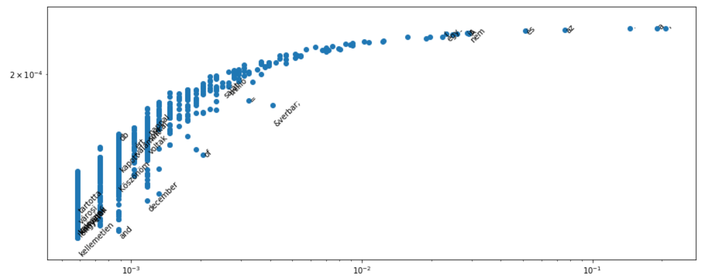
- egy v csúcs e_v *különcsége* a v-ból az összes többi csúcsba vezető legrövidebb utak hosszának maximuma

# mondat	sugár, r	átmérő, d	center	periféria
	$\min e_v$	$\max_v e_v$	$\{v \mid e_v = r\}$	$\{v \mid e_v = d\}$
	$\rightarrow$ –	$\rightarrow$ –		
100	11 7	23 13	{., !, ?}	{nádcukorból}
1k	9	19	{,}	{Megadható, two}

Kapcsolódó irodalom

- TextRank (Mihalcea and Tarau, 2004), kulcsszókinyerés
  results [...] are worse than results obtained with undirected graphs, which suggests that [...] there is no natural "direction"
- trigram (Ferrer i Cancho and Solé, 2001)
- szemantikus hálók (Steyvers and Tenenbaum, 2005)
- a skálafüggetlen-hipe kritikája (Willinger et al., 2009)

Közelségi központiság (*closeness centrality*), irányítatlan



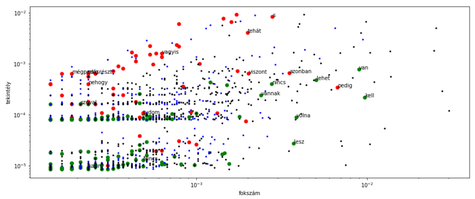
3. ábra A szavak egységes eloszlásban helyezkednek el. Az eloszlásból néhány olyan elem lóg ki, amely „nem illeszkedik a magyar szövegbe": ilyen az egyenlőségjel és egy HTML entitás (`&verbar;`), illetve két angol szó (a *the* és az *of*), melyek előfordulnak a korpuszban. Ezeknek a tokeneknek tehát kisebb a közelségi központiság értékük annál, mint amit gyakoriságuk alapján várnánk. A kilógó elemek pontos karakterizálásához további vizsgálat szükséges.

HITS, hyperlinkindukált témakeresés *(Hyperlink-Induced Topic Search)*

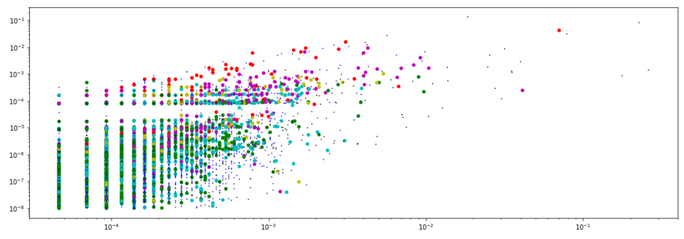
	pl.	miért fontos
hub	index.hu, vajdasag.lap.hu	linkek
tekintély *(authority)*	`http://www.nytud.hu/oszt/korpusz/`	tartalom

- kölcsönös definíció
- számítása iteratíve
  - tetszőleges inicializáció
  - majd minden iterációban $\rightarrow$

$$h(v_1) = \Sigma\{a(v_2) \mid \langle v_1, v_2 \rangle \in E\}$$
$$u(v_2) = \Sigma\{a(v_1) \mid \langle v_1, v_2 \rangle \in E\}$$
$$u \quad /= \sum_v u(v)^2$$
$$a \quad /= \sum_v a(v)^2$$

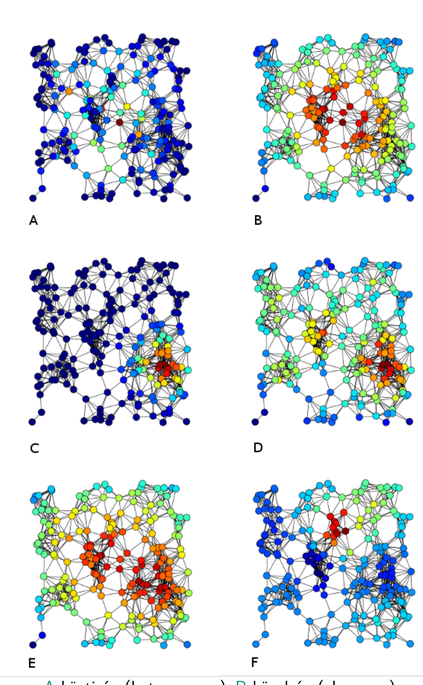


4. ábra A nagyobb, piros ponttal jelölt kötőszavak balra fent (magasabb authority), a nagyobb, zöld ponttal jelölt igék jobbra lent (alacsonyabb authority) helyezkednek el a fokszám (gyakoriság) vs authority grafikonon.



5. ábra Tekintély szófajok szerint: kötőszók, igék, határozók, melléknevek, és számnevek. 10 K mondat, csak a $> 10^{-8}$ tekintélyű szavakat ábrázoltuk.

Központiság



A köztiség (betweenness) B közelség (closeness)
C sajátvektor- D fok (itt gyakoriság)
E harmonikus F Katz

Adat, eszköz, kód

- MNSZ 2 (Oravecz et al., 2014)
- networkx (Hagberg et al., 2008)
- `https://github.com/makrai/textBetweenness/`

További kutatás

- klaszterek szófajok szerint?
- az élsúlyok skálázása távolságként
- irányított gráfok hatékony implementációja
- szemantika

Hivatkozások

A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *science*, 286(5439): 509–512, 1999.

G. Fagiolo. Clustering in complex directed networks. *Physical Review E*, 76(2):026107, 2007.

R. Ferrer i Cancho and R. Solé. The small world of human language. *Proceedings of The Royal Society of London. Series B, Biological Sciences*, 268:2261–2266, 2001.

A. A. Hagberg, D. A. Schult, and P. J. Swart. Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference (SciPy2008)*, pages 11–15, Pasadena, CA USA, Aug. 2008.

R. Mihalcea and P. Tarau. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, 2004.

Cs. Oravecz, T. Váradi, and B. Sass. The Hungarian Gigaword Corpus. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC2014)*, Reykjavík, 2014.

M. Steyvers and J. B. Tenenbaum. The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive science*, 29(1):41–78, 2005.

W. Willinger, D. Alderson, and J. C. Doyle. Mathematics and the internet: A source of enormous confusion and great potential. *Notices of the American Mathematical Society*, 56 (5):586–599, 2009.

G. K. Zipf. *The Psycho-Biology of Language; an Introduction to Dynamic Philology*. Houghton Mifflin, Boston, 1935.