Raffael Vogler

raffael.vogler.de@gmail.com

https://github.com/joyofdata/graph-word-distance

Final project for SNA course on coursera.org

Description

The graph I am going to have a look at is representing the Levenshtein distances of English words. To simplify the examination and interpretation I restricted the graph edges to Levenshtein distances of 1. The words, represented by vertices, are extracted from three novels I obtained from gutenberg.org. Those three novels are „Moby Dick‟ (by Melville), „Great Expectations‟ and „David Copperfield‟ (both by Dickens). A word is simply any uninterrupted sequence of lower letters a to z. Only words that occured at least 5 times and are at least of length 5 are taken into account. Additionally nodes of degree 0 are discared – that means if for a word A exists no word B with Levenshtein distance of 1 then it is not taken into account.

Technicalities

For peforming this analysis I make use of R and Gephi. Script `setup.r` contains the transformation of the three underlying texts into an edge list of word pairs and calculates the mutual Levenshtein distance. The main output is a GraphML file containing a restricted set of edges – those which are fulfilling the conditions described above. But actually Levenshtein distances are calculated for all word combinations - leading to a CSV sized 800 MB and containing more than 36 million distances.

`evaluation_part1.r` to `evaluation_part4.r` hold all calculations performed in R which will be used in this document in the respective sections.

Gephi I am going to use exclusively for visual illustrations and especially for generating a useful layout.

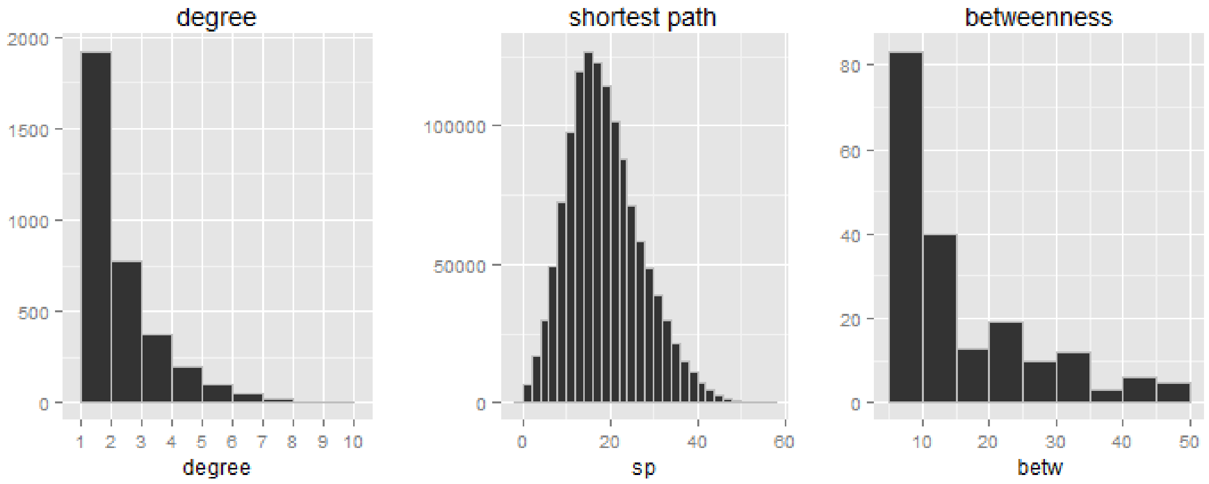
Conclusions

Unfortunately there aren't going to be any important conclusions beyond the documented measurements. In that sense this project serves primarily the purpose for me to practice usage of graph analysis with R and Gephi of course plus conveying an idea about how such a graph looks like - which is interesting in its own right as I think.

Content

1. Basic measurements and differentiation from Erös-Renyi-Graph

The graph consists of 3'273 edges and 3'459 vertices. Its diameter is 56.



	Degree	Shortes path	Betweenness
N	3’459	1’261’146	3’459
Mean	1.89	18.27	3148.91
SD	1.38	8.43	13’928.04
Median	1	17	0
Min	1	1	0
Max	13	56	178’480.5
Range	12	55	7.44
Skew	2.34	0.55	66.17
Kurtosis	7.99	0.13	236.82

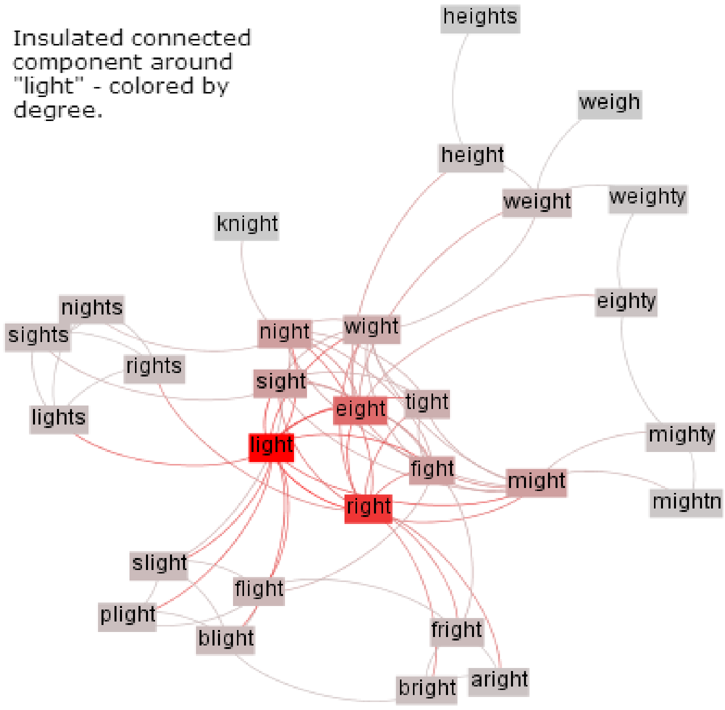
2. TOP 10 words

	Degree
light	13
right	12
eight	11
share	11
stare	11
hears	11
might	10
night	10
sight	10
fight	10

	Betw.
swing	178K
swings	177K
sings	168K
fling	162K
sling	161K
stones	155K
tones	155K
tongs	151K
songs	151K
flying	142K

	Cl. [10^{-7}]
takes	1.23405
stakes	1.23405
stare	1.23404
shares	1.23402
state	1.23402
wakes	1.23402
shakes	1.23402
cakes	1.23401
states	1.23401
scared	1.23400

	EVC
light	1
right	0.95
fight	0.93
eight	0.91
sight	0.91
night	0.89
might	0.89
wight	0.89
tight	0.86
flight	0.35

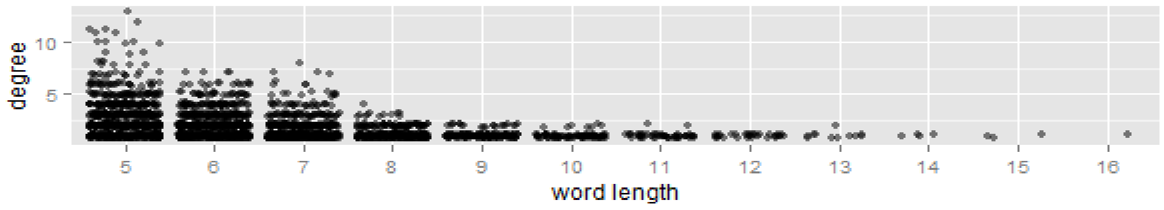


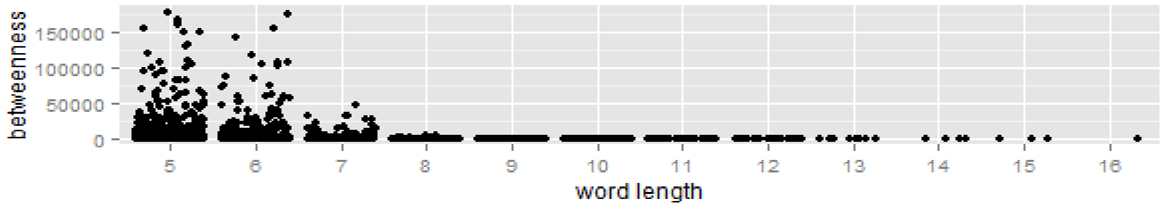
Insulated connected component around "light" - colored by degree.

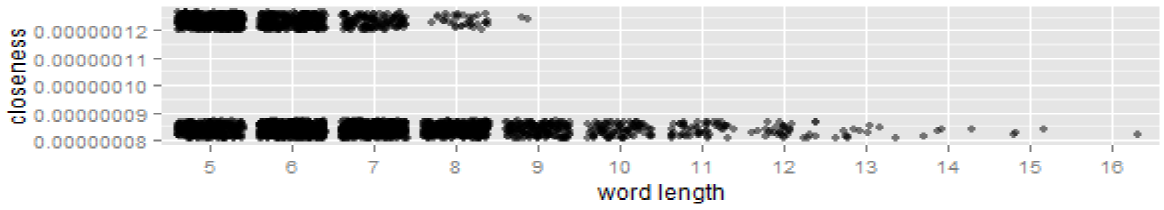
3. Kendall correlation between word length and:

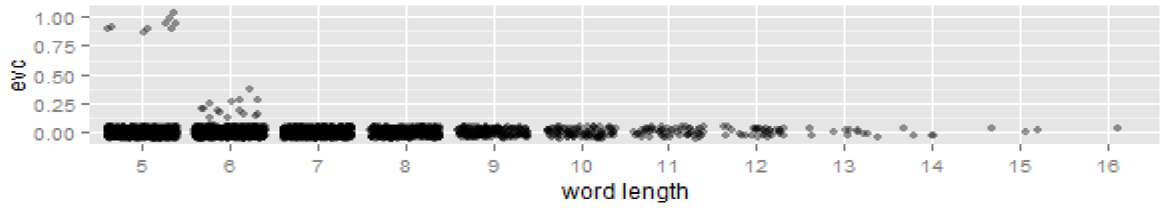
	correlation	p-value
Degree	-0.37	$< 22 \times 10^{-17}$
Betweenness	-0.35	$< 22 \times 10^{-17}$
Closeness	-0.45	$< 22 \times 10^{-17}$
Eigen vector centrality	-0.17	$< 22 \times 10^{-17}$

Looking at the corresponding (jittered) scatter plot renders me sceptical whether those coefficients are truly meaningful.





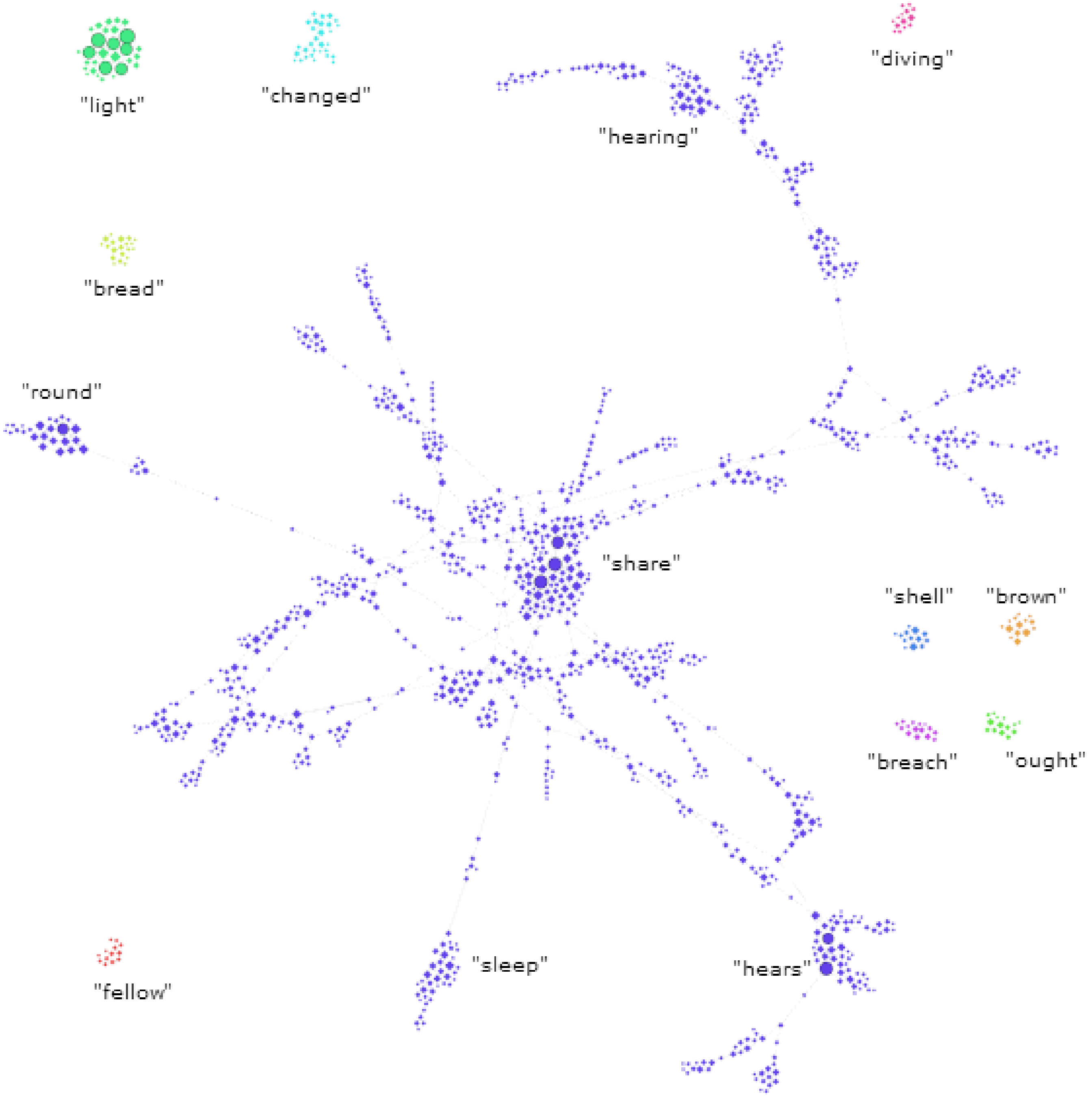




4. Cluster Analysis

a. Connected Components

Comp. size	Frequency	Nodes	Share of nodes
2	618	1236	35.73%
3	127	381	11.01%
4	56	224	6.48%
5	21	105	3.04%
6	9	54	1.56%
7	10	70	2.02%
8	3	24	0.69%
9	2	18	0.52%
10	3	30	0.87%
11	1	11	0.32%
12	2	24	0.69%
13	2	26	0.75%
14	3	42	1.21%
15	1	15	0.43%
20	1	20	0.58%
29	1	29	0.84%
30	1	30	0.87%
1120	1	1120	32.38%



Ten largest connected components

b. The 5 largest cliques

rounds (7), wounds (6), pounds (6), sounds (6), bounds (6), mounds (6)

tearing (5), hearing (8), bearing (7), wearing (7), fearing (5), rearing (7)

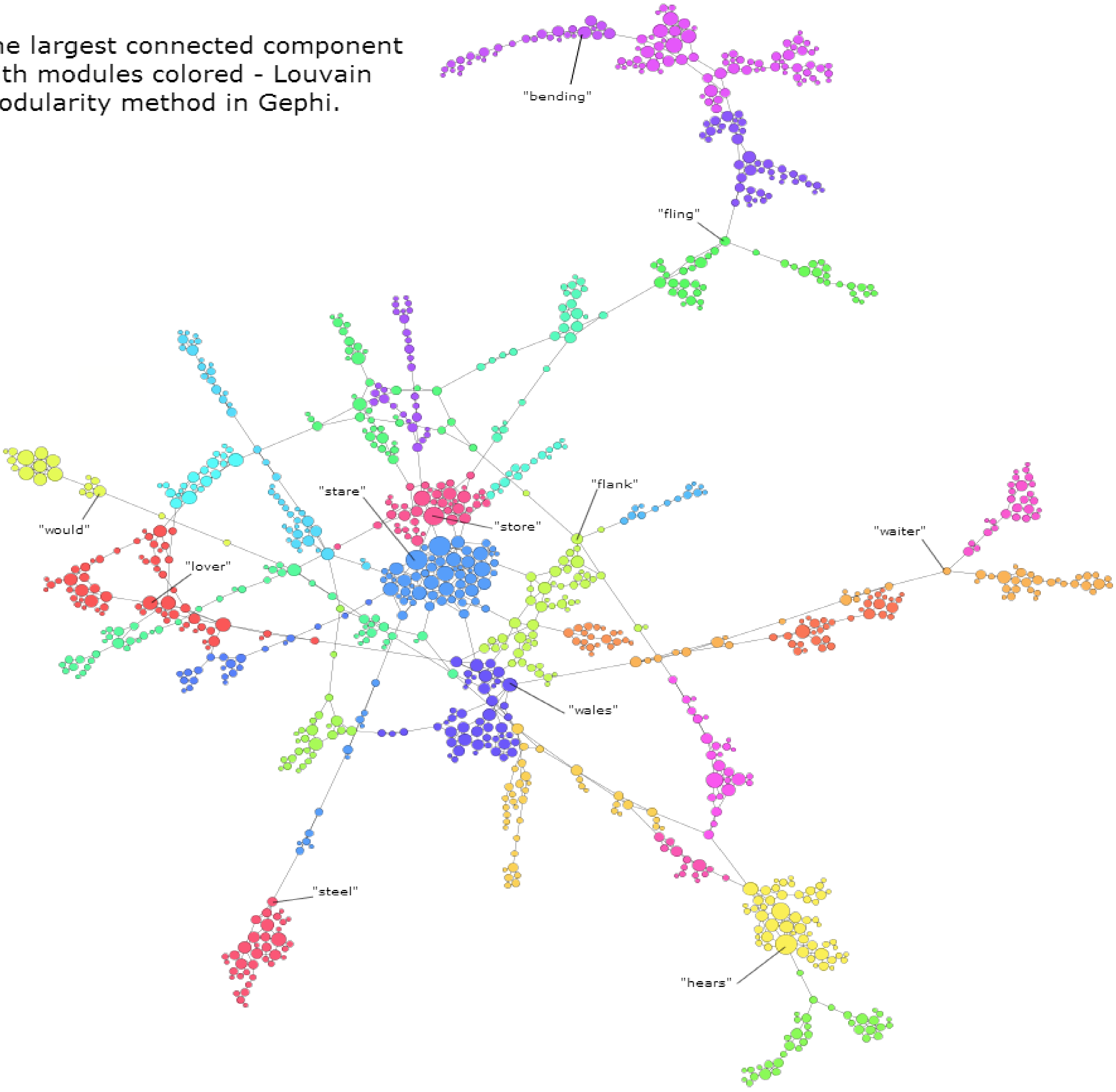
hound (6), round (9), found (6), sound (7), bound (8), pound (7), wound (8)

wears (6), years (5), tears (5), fears (6), bears (9), hears (11)

light (13), sight (19), might (10), night (10), right (12), eight (11), fight (10), tight (8), wight (9)

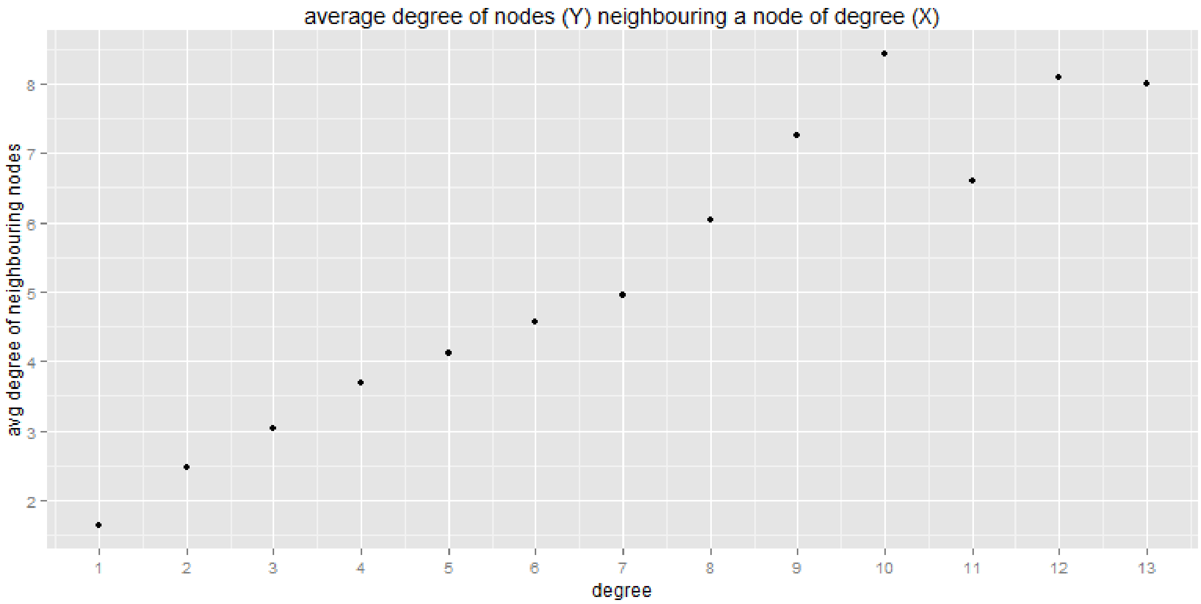
c. The largest connected component colored by modules

The largest connected component with modules colored - Louvain modularity method in Gephi.



5. Assortativity of degree

Kendall-correlation of degrees of directly connected vertices: **0.51** with p


average degree of nodes (Y) neighbouring a node of degree (X)

So, clearly there is a tendency observable of words being distant 1 (Levenshtein distance) having a similar number of neighbours (words as well being distant 1 Levenshtein-distance).