

Homework Three

Most of the top pagerank pages are on insects. At first, this is bewildering — I certainly never associated the Dutch with insects — but it actually makes sense after some thought. Pagerank measures, in a sense, the degree to which pages are linked to from other pages. Due to Linnaeus' binomial nomenclature, pages on life forms tend to link to other pages about life forms (kingdoms link to phyla, phyla to orders, and so on). This causes the broad area of wiki pages on life forms to have a very high density of links. It turns out that beetles form the largest order in the animal kingdom. This means that lots and lots of pages link to the beetles page, either from the lower levels of classification (species, genus, etc) or the upper levels (kingdom, phyla, etc). Thus, it is no surprise that the page for beetles is highly ranked. Similarly, many of the other top pages on insects refer to very large groups within the Linnean system.

Here are the top ranking pages for 2 iterations and 10 iterations, respectively:

318.71838	Engeland
321.00128	Midden-Java
321.13452	Chalcidoidea
322.5381	Beenvissen
323.1978	Brazilië
328.63797	Stekelhuidigen
329.58615	Lophotrochozoa
339.3521	Indonesisch
339.89572	Neopterygii
342.6916	Polen
348.07254	Engels
353.99695	Eenokkreeftjes
359.45734	Spanje
364.54678	Orderfamilie
366.28568	Hymenoptera
368.85892	Bladhaantjes
373.2342	Schietmotten
382.50626	Bladrollers
403.2085	Lamellae
405.04745	Grasmotten
440.1448	Parazoa
445.71768	Noctuoidea
463.0426	Beervlinders
503.3538	Soort
517.5293	Italië
544.11914	Mieren
579.1244	Malacostraca
581.9313	Kreeftachtigen
592.8008	Slakken
661.40436	Rechtvleugeligen
689.7467	België
709.7497	Schildwespen
822.194	Tweevleugeligen
831.2806	Spinners
846.13043	Spinnerellen
852.79987	Apocrita
925.52716	Diptera
975.40906	Nederland
1053.0732	Duitsland
1146.7549	Eumetazoa
1185.2098	Indonesië
1377.2195	Wekdieren
1402.296	Bakteriën
1452.8418	Loopkevers
1511.918	Frankrijk
2345.5686	Geleedpotigen
2948.6162	Vliesvleugeligen
3231.1846	Kevers
4294.235	Dierenrijk

206.72115	Ennominae
209.18346	Noord-Sumatra
210.39185	Tweekleppigen
213.04651	Maxillopoda
223.21669	Spanje
225.97327	Kniptorren
226.69673	Nymphalidae
228.82565	Engeland
230.53465	Polen
232.35258	Snuitmotten
233.76839	Lycaenidae
243.26741	Atjeh
253.7078	Sikkelmotten
254.6005	Trichoptera
258.01562	Roofvliegen
258.9948	Oost-Java
271.60266	Straalvinnigen
277.21338	Eulophidae
285.7205	Italië
290.6476	Midden-Java
335.02225	Bladhaantjes
339.72116	Eenogkreeftjes
343.2132	Bladrollers
350.59366	België
353.2658	Beervlinders
353.67157	Laminae
380.6047	Grasmotten
427.52814	Mieren
466.44998	Eumetazoa
473.95798	Slakken
501.53687	Soort
521.7227	Rechtvleugeligen
535.4946	Nederland
538.2799	Malacostraca
542.2365	Tweekvleugeligen
605.9817	Duitsland
703.4662	Schildwespen
786.86884	Spinnerullen
793.2409	Spanners
864.3758	Frankrijk
882.1034	Diptera
944.38153	Indonesië
1155.1004	Wekdieren
1265.2546	Loopkevers
1270.7402	Boktorren
1641.9781	Geleedpotigen
2612.3796	Vliesvleugeligen
2756.936	Dierenrijk
2823.4177	Kevers
hduser1@ip-172-31-47-143: /usr/local/hadoop-2.7.1\$	

This information can be used to determine “central hubs” on the Wikipedia network. From this, one can draw conclusions about the topic areas that have the best Wikipedia coverage, for example. Another example (contrived, as long as Wikipedia keeps its word) would be determining ad pricing based on the wiki page the ad appears on; the higher the page rank of a page, the more value its ad space is.