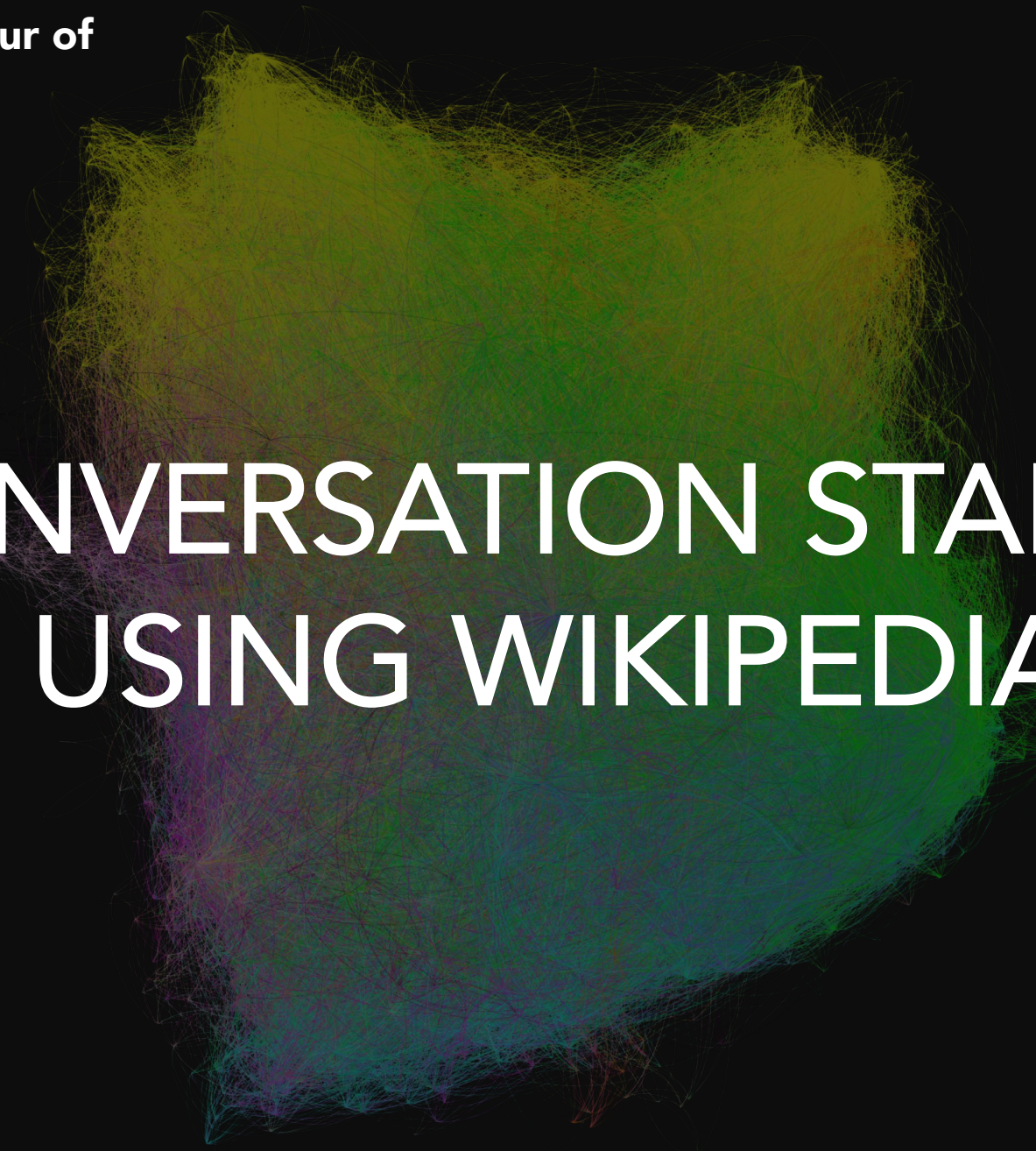
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

A Network Tour of Data Science

22 janvier 2019

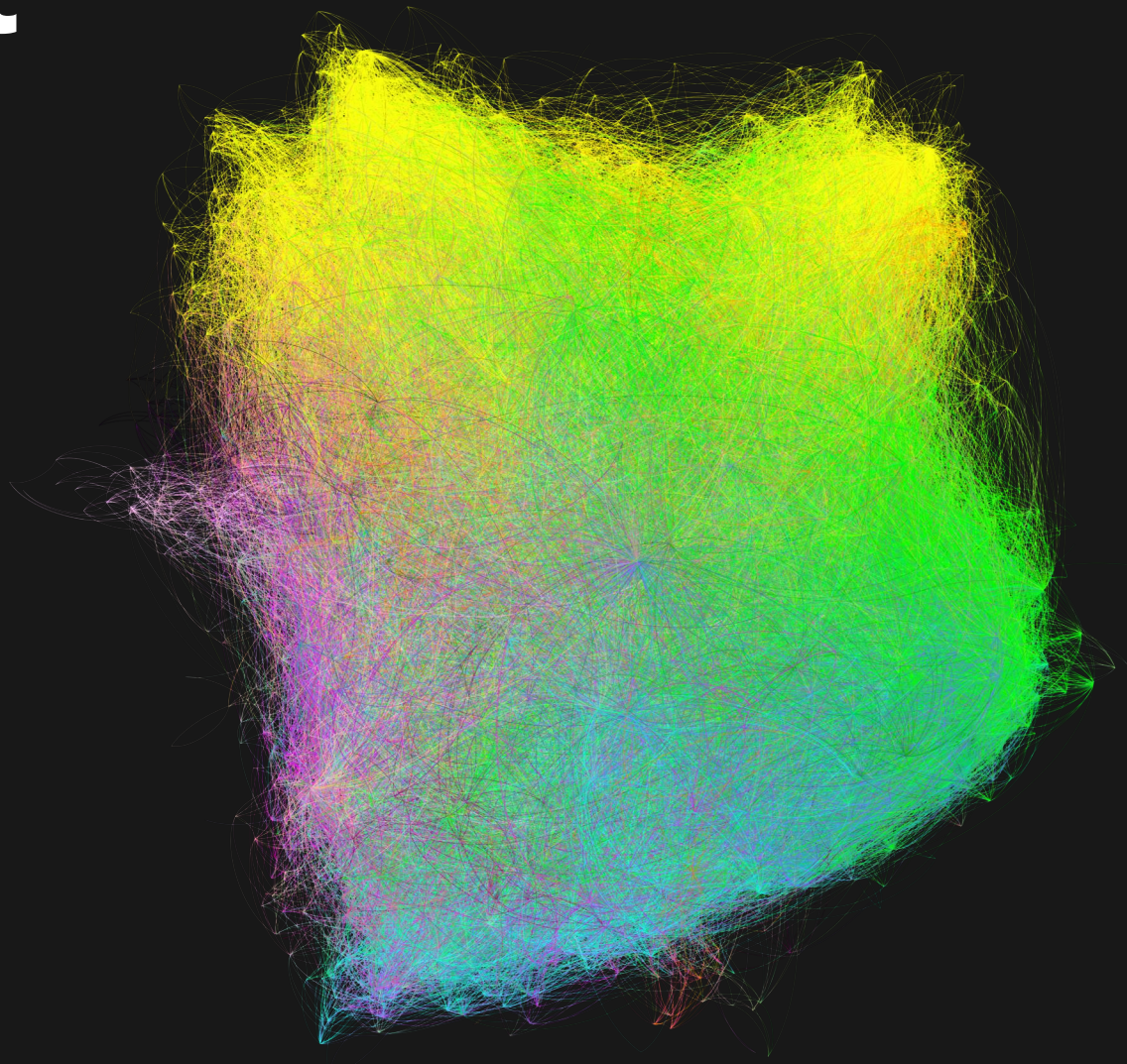


CONVERSATION STARTER USING WIKIPEDIA

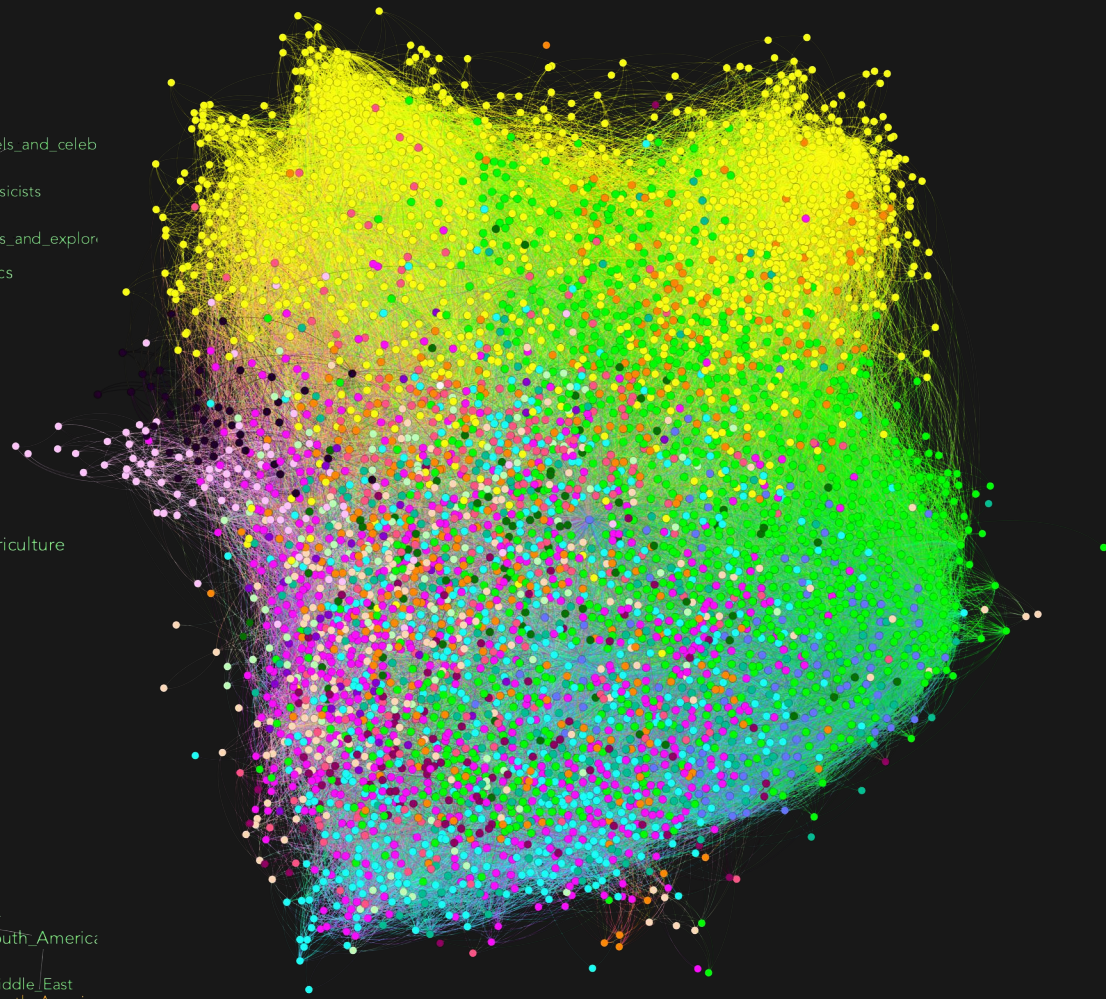
RÉMI CLERC, PATRICK LEY, JORDAN METZ, DAVID SANCHEZ DEL RIO

Goals of the project

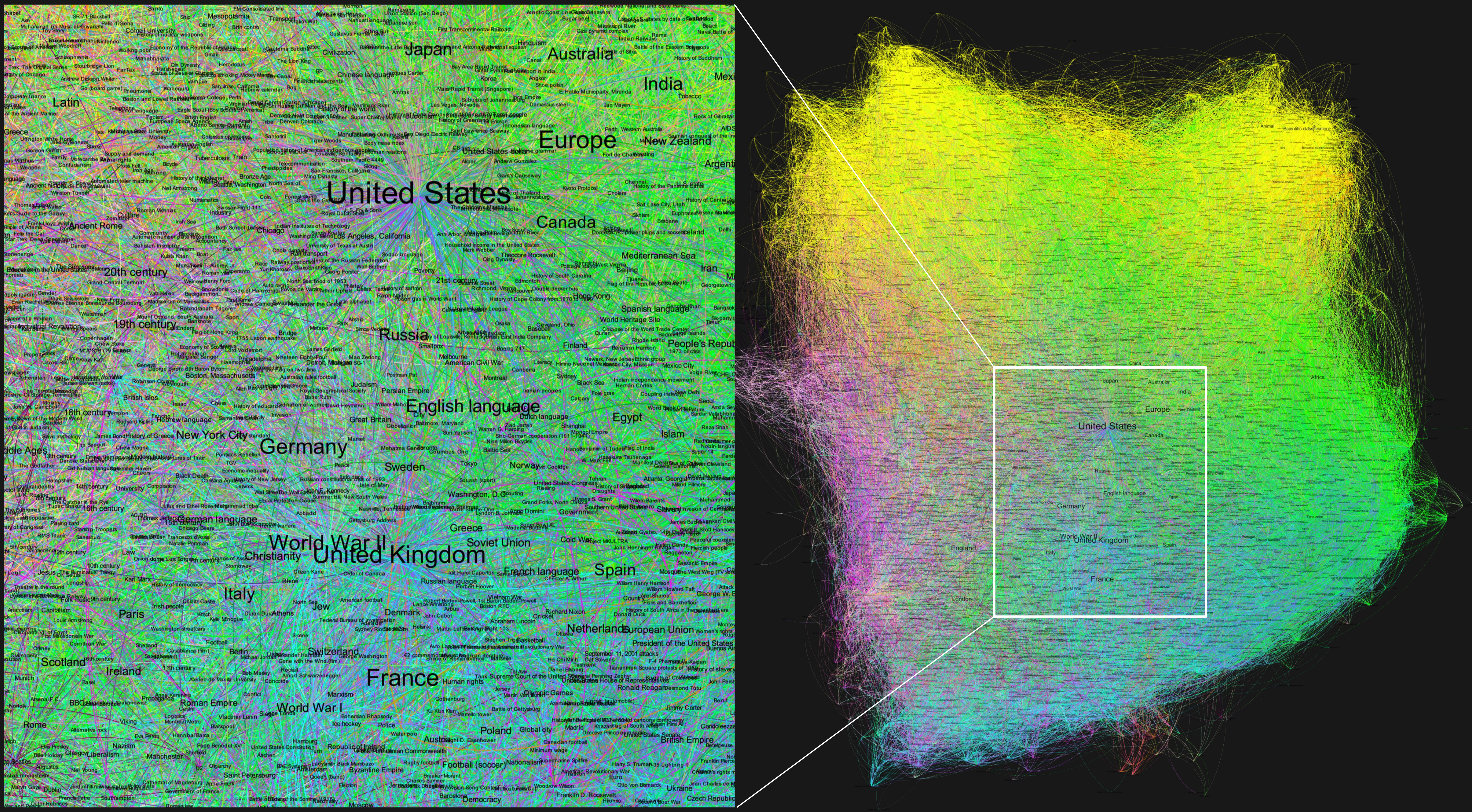
- Connectivity of the graph : **six degrees of separation**
- Study how categories influence the connectivity of the graph
- Find a way to recommend discussion topics



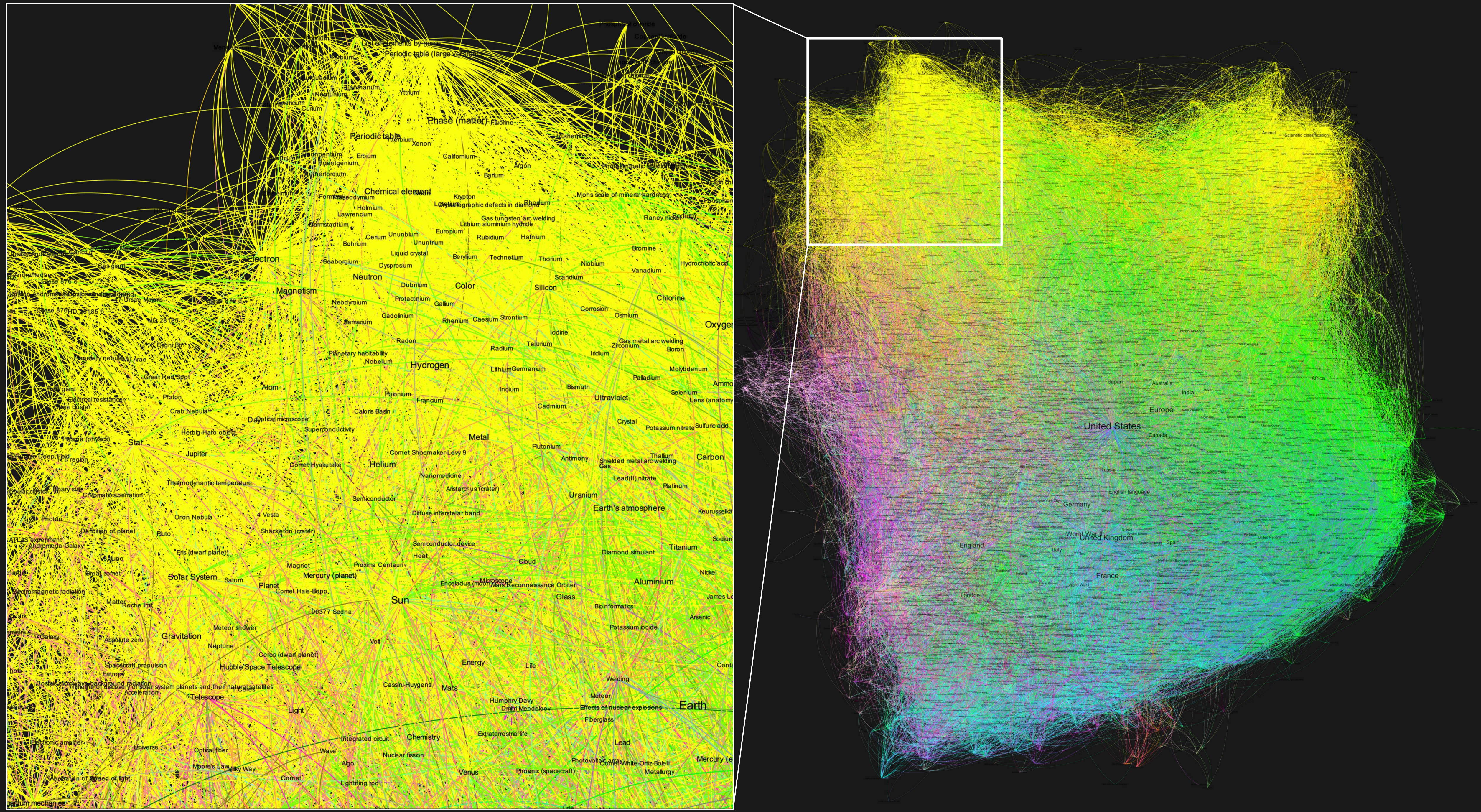
Wikipedia dataset



Science
People
Country
Geography
IT
Everyday Life
History
Mathematic
Music



United States
Japan
Australia
India
Europe
New Zealand
Latin
Canada
Ancient Rome
20th century
19th century
Russia
English language
Egypt
Germany
New York City
Sweden
Islam
World War II
United Kingdom
Christianity
Greece
Soviet Union
Spain
Italy
Paris
French language
Netherlands
European Union
Scotland
Ireland
Switzerland
France
Roman Empire
World War I
Rome
Poland
Austria
Football (soccer)
British Empire
England
London



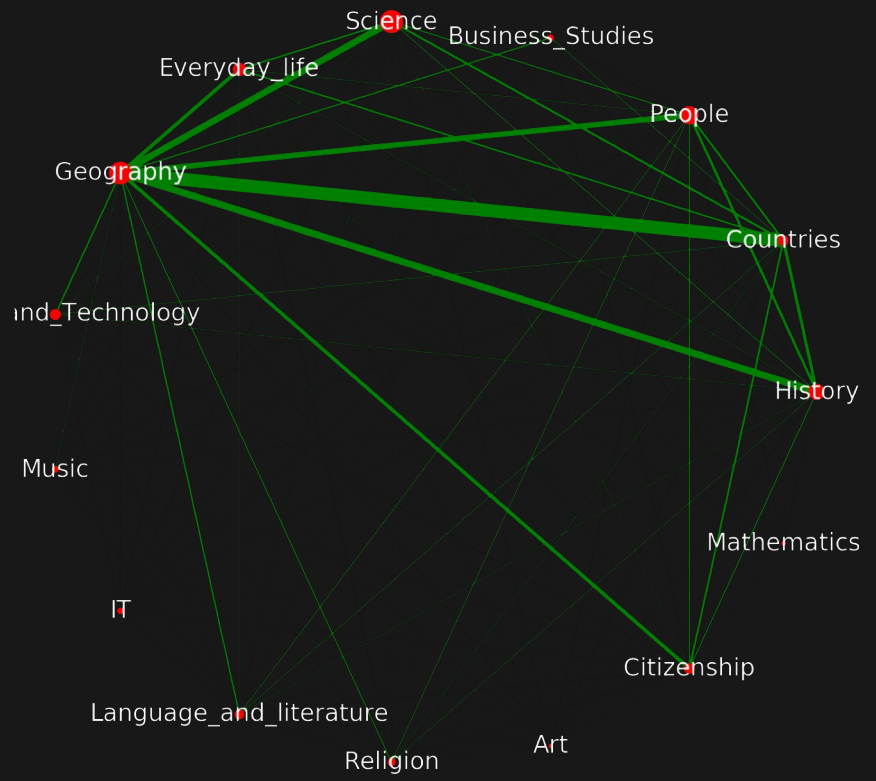
Periodic table
Phase (matter)
Chemical element
Electron
Magnetism
Neutron
Color
Silicon
Chlorine
Hydrogen
Atom
Star
Jupiter
Helium
Metal
Ultraviolet
Carbon
Uranium
Earth's atmosphere
Titanium
Solar System
Planet
Mercury (planet)
Sun
Aluminium
Gravitation
Hubble Space Telescope
Energy
Mars
Telescope
Light
Chemistry
Venus
Lead
Earth
United States
Europe
Canada
Germany
United Kingdom
England
France
London
English language
World War II
Japan
China
India
Australia
Russia
Italy
Spain
Africa

Analysis of the categories

Catégorie	nodes	edges (catégorie uniquement)	Density [edge/node]	edges (total)	edges (to/from outside)	Outside connection density	Outside/inside difference	Nodes with multiple categories	Density of multinodes [%]	Nodes with DIFFERENT categories	Density of Difnodes [%]
History	545	3874	7,108256881	25725	21851	40,09357798	5,640423335	145	26,60550459	80	14,67889908
Countries	229	5523	24,11790393	36201	30678	133,9650655	5,554589897	229	100	97	42,3580786
People	689	3073	4,460087083	23819	20746	30,11030479	6,751057598	111	16,11030479	37	5,370101597
Business Studies	88	358	4,068181818	4532	4174	47,43181818	11,65921788	14	15,90909091	9	10,22727273
Science	1120	13978	12,48035714	41026	27048	24,15	1,935040778	98	8,75	22	1,964285714
Everyday_life	372	1715	4,610215054	13964	12249	32,92741935	7,142274052	78	20,96774194	48	12,90322581
Geography	1084	22570	20,82103321	80707	58137	53,63191882	2,575852902	287	26,47601476	154	14,20664207
Design and Technology	253	530	2,09486166	6467	5937	23,46640316	11,20188679	49	19,36758893	29	11,46245059
Music	96	503	5,239583333	2588	2085	21,71875	4,145129225	1	1,041666667	0	0
IT	83	442	5,325301205	1810	1368	16,48192771	3,095022624	10	12,04819277	3	3,614457831
Language and literature	194	903	4,654639175	8035	7132	36,7628866	7,898117386	22	11,34020619	8	4,12371134
Religion	134	762	5,686567164	6655	5893	43,97761194	7,733595801	20	14,92537313	2	1,492537313
Art	38	126	3,315789474	1624	1498	39,42105263	11,88888889	5	13,15789474	3	7,894736842
Citizenship	224	1225	5,46875	13291	12066	53,86607143	9,849795918	34	15,17857143	15	6,696428571
Mathematics	45	222	4,933333333	1173	951	21,13333333	4,283783784	2	4,444444444	0	0
			Médiane = 5.24			Médiane = 36.76	Médiane = 6.75	Médiane = 34		Médiane = 15	
			Moyenne = 7.63			Moyenne = 41.28	Moyenne = 6.76	Moyenne = 73.6		Moyenne = 33.18	
			Nodes with a higher score have more eges per node inside the category itself			Categories with a higher score have more connections to other categories	Categories with a high score have a lot more connections to the outside than in the category. Categories with a low score are more connected to themsleves than outside				

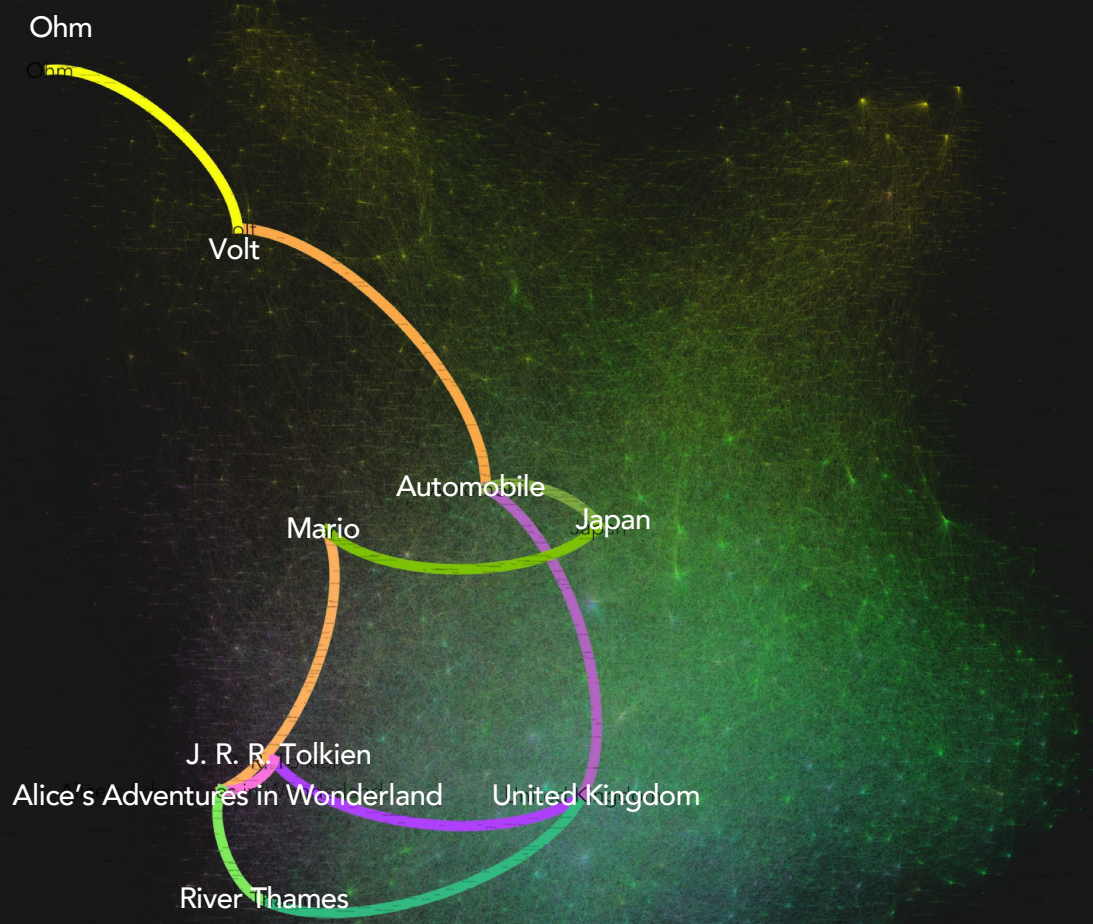
Connectivity between categories

	History	Countries	People	Business_Stu	Science	Everyday_life	Geography	Design and te	Music	IT	Language & li	Religion	Art	Citizenship	Mathematics	Total
History	0	4432	3864	351	1509	939	9228	900	145	46	879	1070	247	1790	100	25500
Countries	4432	0	3187	1331	3391	2716	16696	1104	484	139	1161	902	173	2960	45	38721
People	3864	3187	0	305	2033	1021	7523	641	354	113	1323	1352	328	1878	204	24126
Business Studies	351	1331	305	0	313	207	2240	161	13	99	80	40	14	399	24	5577
Science	1509	3391	2033	313	0	2802	8767	824	76	158	563	470	119	817	185	22027
Everyday_life	939	2716	1021	207	2802	0	5395	269	119	177	567	227	94	595	47	15175
Geography	9228	16696	7523	2240	8767	5395	0	2658	901	247	2485	1821	399	5188	92	63640
Design and Technology	900	1104	641	161	824	269	2658	0	39	90	134	63	78	271	37	7269
Music	145	484	354	13	76	119	901	39	0	19	116	35	19	78	4	2402
IT	46	139	113	99	158	177	247	90	19	0	102	7	4	119	56	1376
Language and literature	879	1161	1323	80	563	567	2485	134	116	102	0	367	92	397	42	8308
Religion	1070	902	1352	40	470	227	1821	63	35	7	367	0	78	435	44	6911
Art	247	173	328	14	119	94	399	78	19	4	92	78	0	60	8	1713
Citizenship	1790	2960	1878	399	817	595	5188	271	78	119	397	435	60	0	52	15039
Mathematics	100	45	204	24	185	47	92	37	4	56	42	44	8	52	0	940
Total	25500	38721	24126	5577	22027	15175	63640	7269	2402	1376	8308	6911	1713	15039	940	238724
Contribution	10,68179152	16,21998626	10,10623146	2,336170641	9,226973409	6,356713192	26,6584005	3,044938925	1,006182872	0,576397848	3,480169568	2,89497495	0,717565054	6,299743637	0,393760158	100



Shortest path algorithm

- Choose a start and a target node
- The algorithm will check all the node's neighbors and progressively the neighbor's neighbors etc.
- Doing this from both start and target nodes increases its performance.



Connecting two categories

Best connecting category between two others

	History	Countries	People	Business	Science	Life	Geography	D&T	Music	IT	Language	Religion	Art	Citizenship	Mathematics
History		Geography	Geography	Geography	Geography	Geography	Countries	Geography	Geography	Geography	Geography	Geography	Geography	Geography	People
Countries			Geography	Geography	Geography	Geography	History	Geography	Geography	Geography	Geography	Geography	Geography	Geography	People
People				Geography	Geography	Geography	History	Geography	Geography	Geography	Geography	Geography	Geography	Geography	Science
Business					Geography	Geography	Countries	Geography	Geography	Geography	Geography	Geography	Geography	Geography	People
Science						Geography	Countries	Geography	Geography	Geography	Geography	Geography	Geography	Geography	People
Life							Science	Geography	Geography	Geography	Geography	Geography	Geography	Geography	People
Geography								Countries	Countries	Life	People	People	People	Countries	People
D&T									Geography	Geography	Geography	Geography	Geography	Geography	People
Music										Geography	Geography	Geography	Geography	Geography	People
IT											Geography	Geography	Geography	Geography	Science
Language												Geography	Geography	Geography	People
Religion													Geography	Geography	People
Art														Geography	People
Citizenship															People
Mathematics															

	History	Countries	People	Business	Science	Life	Geography	D&T	Music	IT	Language	Religion	Art	Citizenship	Mathematics
History		People	Countries	Countries	Countries	Countries	Countries	Countries	Countries	Life	People	People	People	Countries	People
Countries			History	Citizenship	Life	Science	History	History	People	Life	People	People	People	People	People
People				Countries	Countries	Countries	History	Countries	Countries	Life	Countries	History	History	Countries	Science
Business					Countries	Countries	Countries	Countries	Countries	Life	Countries	Countries	People	Countries	People
Science						Countries	Countries	Countries	Countries	Life	People	People	People	Countries	People
Life							Science	Countries	Countries	Science	Countries	People	People	Countries	People
Geography								Countries	Countries	Life	People	People	People	Countries	People
D&T									Countries	Life	Countries	Countries	People	Countries	People
Music										Countries	Countries	Countries	People	Countries	People
IT											Life	Life	Countries	Life	Science
Language												People	People	People	People
Religion													People	People	People
Art														People	People
Citizenship															People
Mathematics															

Connecting two categories

Used for recommendation:

	History	Countries	People	Business	Science	Life	Geography	D&T	Music	IT	Language	Religion	Art	Citizenship	Mathematics
History		People	Citizenship	Citizenship	People	Science	People	Science	People	Life	People	People	People	History	People
Countries			History	Citizenship	Life	Science	History	History	People	Life	People	People	People	People	People
People				Citizenship	History	Science	History	History	History	Life	History	History	History	History	Science
Business					Citizenship	Citizenship	Citizenship	History	People	Life	Citizenship	Citizenship	People	History	People
Science						People	Life	History	People	Life	People	People	People	People	People
Life							Science	History	People	Science	People	People	People	People	People
Geography								History	People	Life	People	People	People	People	People
D&T									People	Life	History	History	People	History	People
Music										Life	People	People	People	People	People
IT											Life	Life	Science	Life	Science
Language												People	People	People	People
Religion													People	People	People
Art														People	People
Citizenship															People
Mathematics															

Results of the conversation starter

- From two categories we've been able to create article/topic recommendations.
- Multiple paths between categories
- For Josh's little history, we got the following results.

Recommendation: Between Science and Literature: Should talk about **People**.
Recommended pages:

id	name	url
167	Albert Einstein	en.wikipedia.org/wiki/Albert_Einstein
2086	Immanuel Kant	en.wikipedia.org/wiki/ImmanuelKant
3881	Stephen Hawking	en.wikipedia.org/wiki/Stephen_Hawking
3260	Plato	en.wikipedia.org/wiki/Plato
737	C. S. Lewis	en.wikipedia.org/wiki/C._S._Lewis


Any questions ?


Thank you !