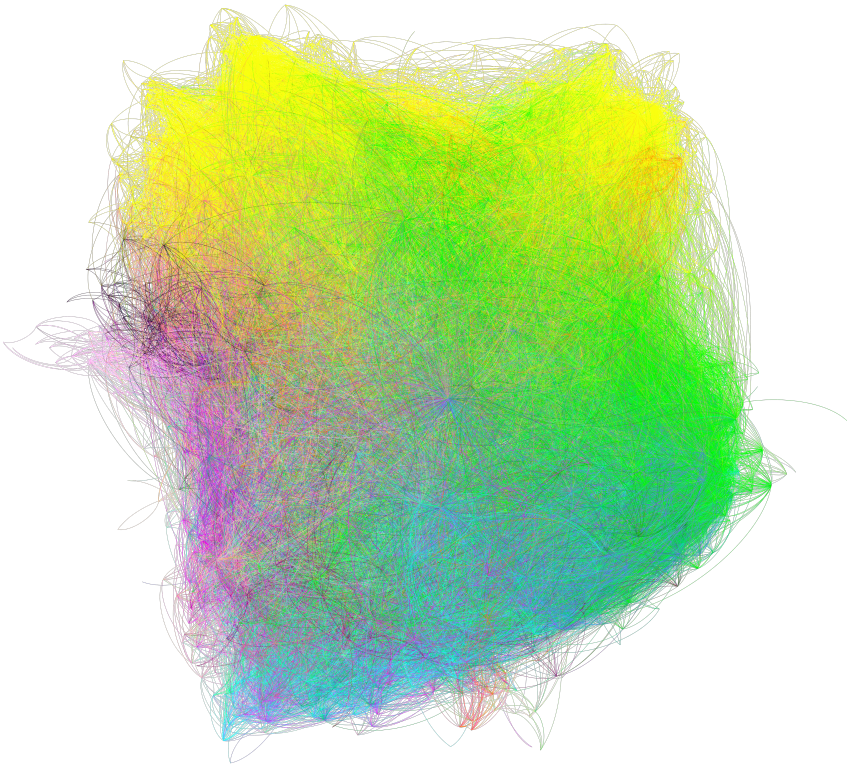
A Network Tour Of Data Science

Conversation starter using Wikipedia



Rémi Clerc

David Sanchez del Rio

Patrick Ley

Jordan Metz

November 5, 2020

1 Introduction

Wikipedia is an online encyclopedia boasting over 5 million articles in English and many more in multiple different languages. As such the dataset that is used to represent this data is enormous. Yet still we interact with it each day, shaping it through every new information that we deem noteworthy. In itself it represents a part of the human experience captured in a form that is easy to analyze using Data Science.
During the semester we were able to study our graph and its properties and one thing that stood out was the strong connection between each and every node. This connectivity was captured in the noteworthy "Wikipedia Game", where one has to connect two articles using only the links in each web page, and is described by the "six-degrees of separation" hypothesis, that states that any two articles can be linked through six steps or less.
In this project we wanted to go further by looking more precisely at the emergence of this connectivity. In a first part, we designed an algorithm that finds the shortest paths between any two nodes in the graph to exploit the six-degrees of separation claim. Secondly, we wanted to use this inter-connectivity to be able to interpret the relations between different Wikipedia categories and provide a service through which two people from different backgrounds could get a couple of recommended topics that both have in common.
To illustrate the results of the project, let us follow a story:

> *"Josh is an Electrical Engineer. While he was in the train going to the EPFL, a young and beautiful woman, Jessica, sat next to him. Josh, obviously interested, saw that she started reading a book: "Alice's Adventure in Wonderland". As he was curious about her, he asked her what she was studying. She answered that she was doing a Bachelor of Arts in Literature in Geneva, and this was one of her favorite books. Josh panicked, he did not know enough about literature to find a topic of conversation that is both interesting for him and for her."*

In this project, we will help Josh find interesting topics to have an engaging discussion with her.

2 Exploration of the graph

The information at our disposal were:

1. the name of the pages; referred to as **nodes**
2. which page has a link to which other; referred to as **edges**
3. the category and subcategory for each page; referred to as **features**

Which gives us the graph in Figure 1.

As the only feature at our disposal were the categories, the graph of categories has been drawn for a better visualization of categories distribution. In Figure 2, the size of the titles is related to the number of nodes inside this category.

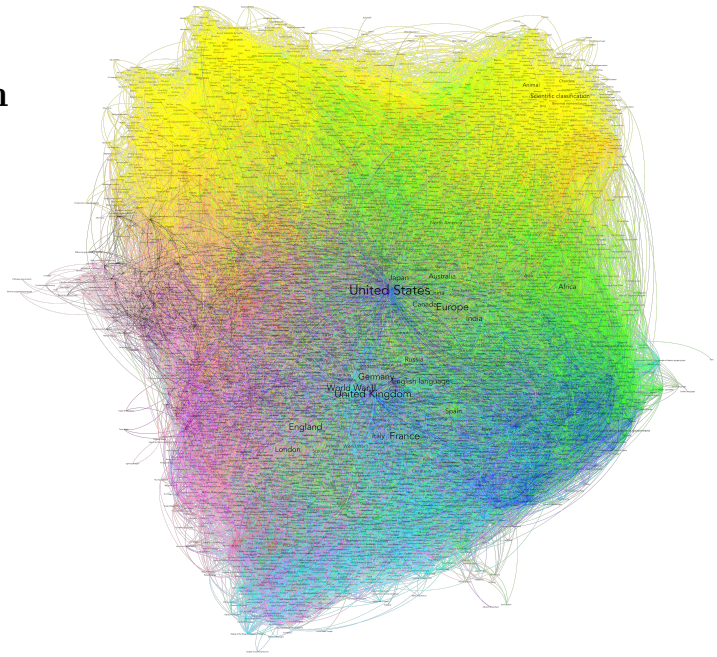


Figure 1: Wikipedia graph (Yellow: "Science", green: "Geography", Pink: "People", Blue: "Country", Cyan: "History")

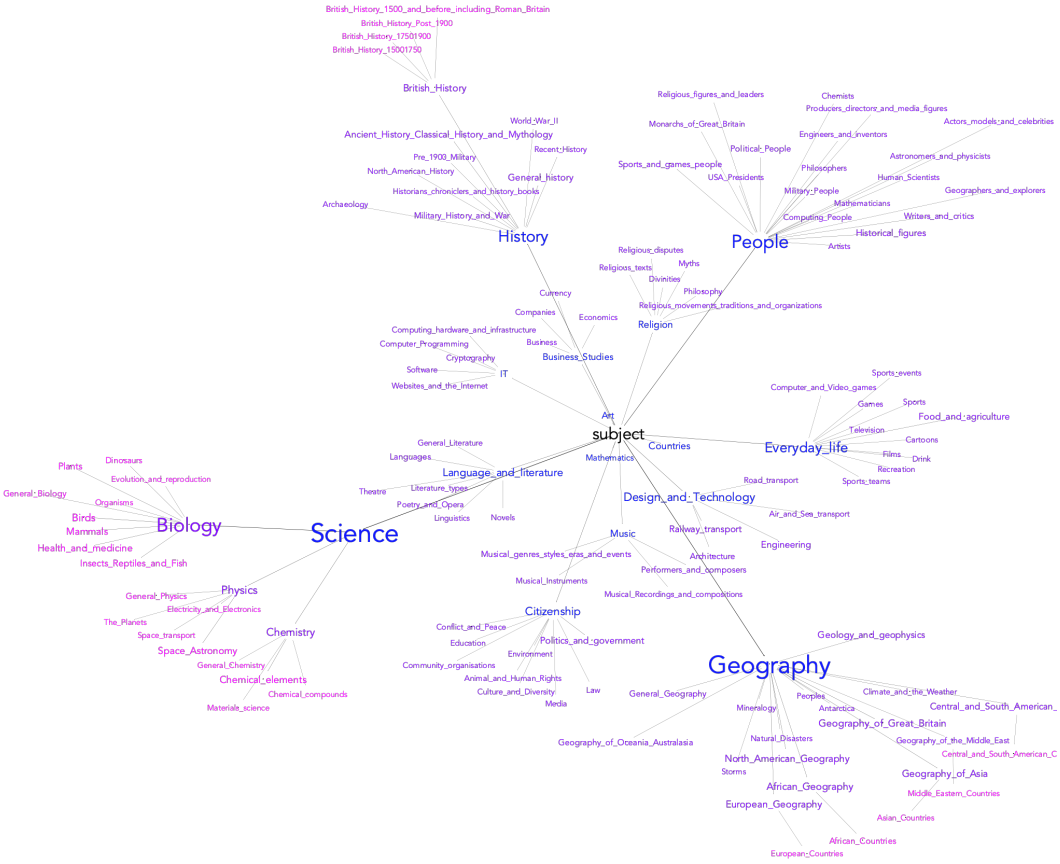


Figure 2: Categories and subcategories of Wikipedia

We can see that "Science" and "Geography" are the two biggest categories of our dataset, with a big proportion of "Science" being the "Biology" subcategory. "History" and "People" are a bit smaller categories, but still represent a large part of the dataset.
One of the more interesting parts of the Wikipedia dataset, in contrast to other datasets, is its connectivity; meaning the links in between pages. This gives us interesting information on how we group our knowledge and also gives us a very objective understanding of the relation in between different subjects, by just counting the number of connections in between. Obviously Wikipedia does not give a perfect analysis nor is it complete but the tendencies are still compelling.

In the following table we registered the number of links from each category to each other. It stands out that "Geography" has lot of links with other categories (about 26% of the overall links). Hence, it contributes the most to the connectivity of the graph.

	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

Figure 3: Connectivity of the categories of Wikipedia

The results can be seen in a more legible way in Figure 4, where the size of the edges is proportional to the number of connections between each category. It becomes increasingly clear how big a role "Geography" plays in the inter-connectivity in between categories. This makes sense since often articles will have general geographical information. Something remarkable to note is also the fact that even though "Science" is one of the biggest categories yet it boast comparatively little connections when compared to "Geography" or even "Countries". This shows us that most of its connections are internal and not with other categories.

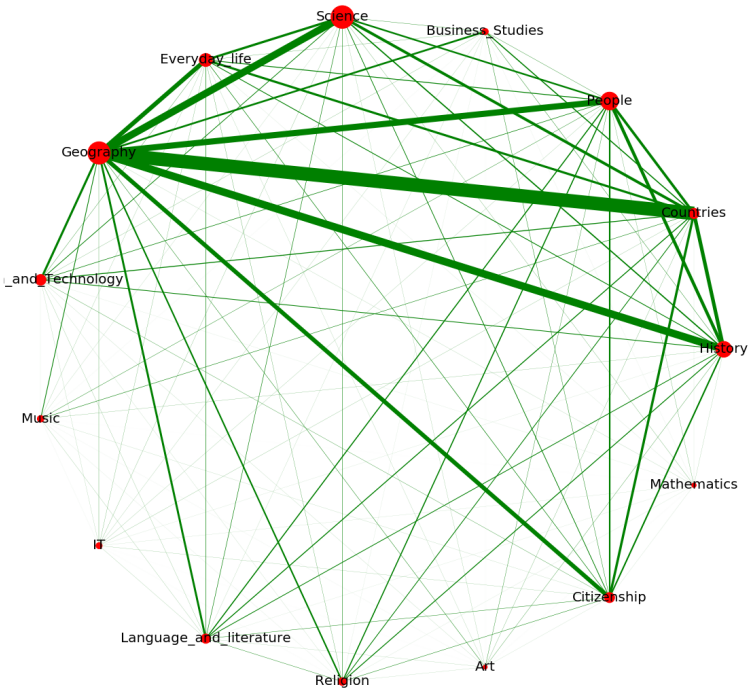


Figure 4: Connectivity of the categories of Wikipedia

3 Exploitation of the data

Six-degrees of separation is the idea that we can link any node with any other through at most six steps. This hypothesis applies to graphs that follow a "Small World" property, which is the case for Wikipedia. We tested this using our dataset by starting at an article (i.e a node in our graph) and trying to find the shortest path to an end node. At first we implemented an algorithm that brute forces through the neighbours of the start node and their neighbours until the end node is reached. Since our data set is sufficiently small this is not impossible computationally, but the complexity is exponential. We later improved this algorithm by propagating the paths from the end node as well, until the propagated paths from start node and end node reach each other. This significantly improved the performance.
To help Josh find a way to the heart of the women, we decided to start from "Ohm", trying to reach "Alice's Adventures in Wonderland". The graph on Figure 5 shows what paths are the shortest between the two.

Those paths may be quite interesting for Josh, however, he may also not know anything worthwhile about Automobiles or the Thames. We want to give him a better choice. Starting from a graph that only contains the starting category (e.g. Science) and the goal category (e.g. Language and Literature), we will add a third category from the dataset, the one that adds the largest number of new connections between the starting and end categories.

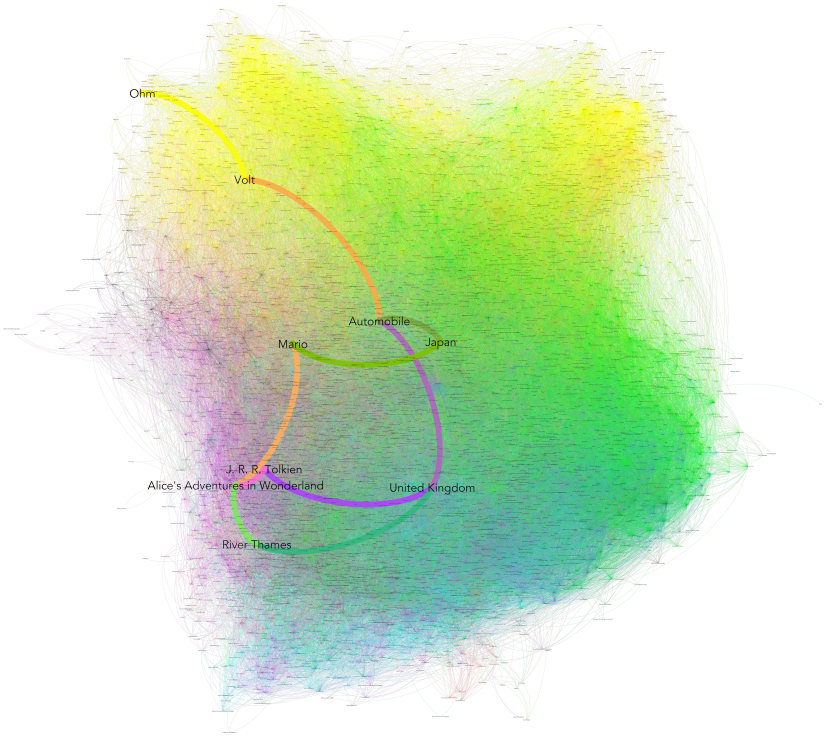


Figure 5: Shortest paths from "Ohm" to "Alice's Adventures in Wonderland"

We see in Figure A.1 that Geography is the category that adds the largest number of connections for almost all other pairs of categories (with the notable exception of Mathematics, which is the only category that is not well connected with Geography in the first place). This result was expected since Geography clearly dominated the category graph with about 26% of the connections of the graph. Since talking about where they are from would be "too easy", we decided to remove Geography from the possible choices to see what would happen. In Figure A.2, we see that now the most present categories are Countries and People. Since Countries could be a subcategory of Geography, we decided to remove it for the same reason.

	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															

Figure 6: Best linking category, excluding "Geography" and "Countries"

We see in the resulting spreadsheet (Figure 6) that the category that adds the most paths between "Science" and "Language and Literature" is "People", but to make it easier for Josh than just telling him to talk about people, we also decided to give more specific topics that linked the two categories. The algorithm we designed finds paths that have the starting node in the starting category (e.g. Science), a unique intermediary node in the linking category (e.g. People) and a goal node in the goal category (e.g. Language and Literature). We then switch the starting and goal category and find the same type of paths in the other direction. We keep only the intermediary nodes that were common in both directions, which would let both Josh and Jessica be familiar with the topic.
Let's see what Josh should talk about:

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`

The goal was to get articles that would be well connected to "Science" and "Language and literature", and in fact those recommended people all have a relationship to both fields. Josh could choose any of them to start a conversation with Jessica.
This example worked very well, but we have noticed that the results for categories that have a smaller contribution to the connectivity (e.g. Art, Music or IT) are sometimes less convincing. But the algorithm that finds the common topics has a lot of random factors (start node, end node, intermediary node), hence the results vary from execution to execution, as their effect may vary from person to person. So in case the results are not convincing, the user could run the algorithm another time until they get something they can talk about.

4 Conclusion

In the exploration phase, we found that our graph has a strong connectivity, and thus it is easy to connect two seemingly unrelated articles through a small path of articles. We also studied how the categories interact, and found that geographical subjects (Geography and Countries) dominate the interactions, and that we could obtain a more balanced graph by removing them. We exploited those properties to design a suggester of topics of conversation, whose results are convincing. We could easily see this product being adapted, for instance, as a mobile app, so it could be used to help out in difficult social situations.

Appendix A

Additional Figures

	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															

Figure A.1: Best linking category between two other

	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															

Figure A.2: Best linking category between two other, excluding "Geography"