Wikipedia Networks

Big Data Course – Project Presentation

Nina Varchavsky-Bergin

18.12.2019

Wikipedia is community-driven

Wikipedia is community-driven

Is there an impact of language and culture on wikipedia article networks structure ?

The dataset:

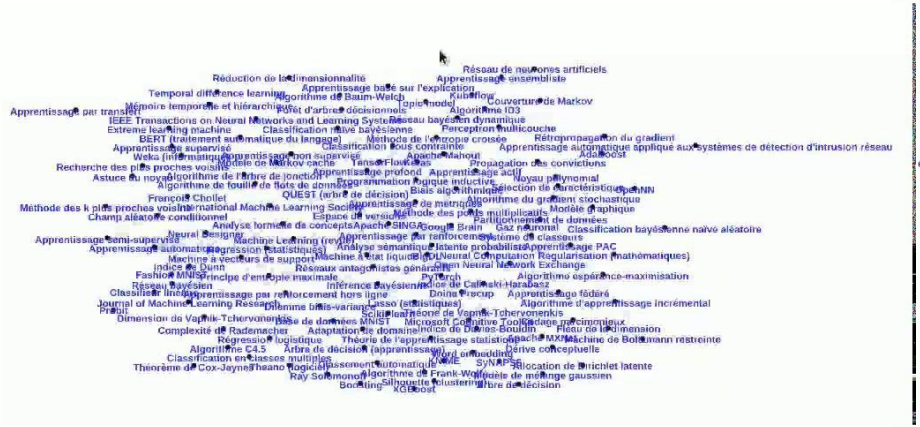
- French
  Apprentissage automatique
- English
  Machine Learning

Id
Title
Content of the page
URL
Length of the article
Links to other articles
Last modification date
Wikibase number
Wikidata URL
Aliases

Nodes: articles
Edges: hyperlinks

Scraped thanks to wptools module

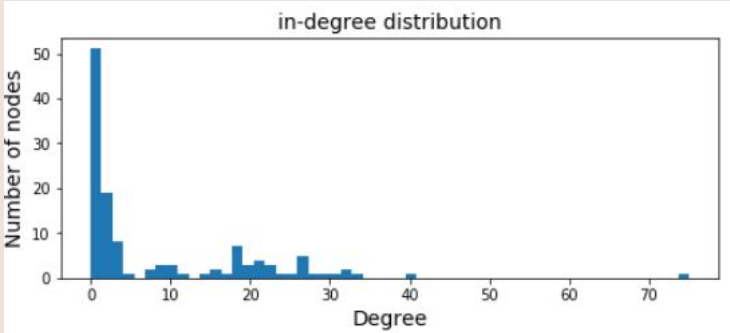
French

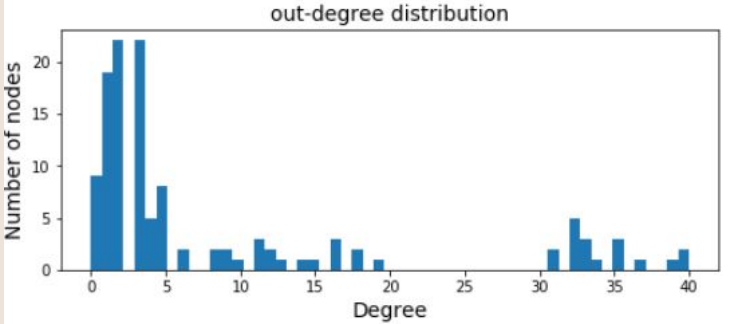


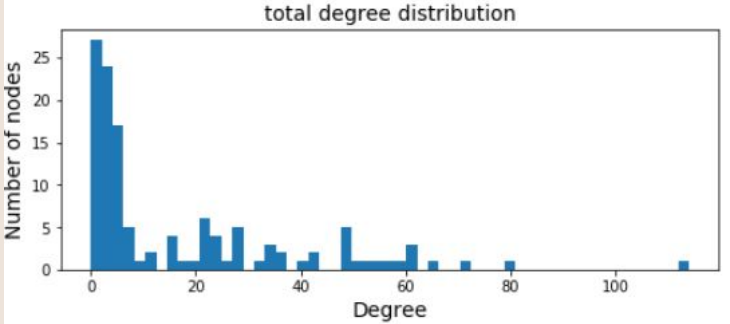
English

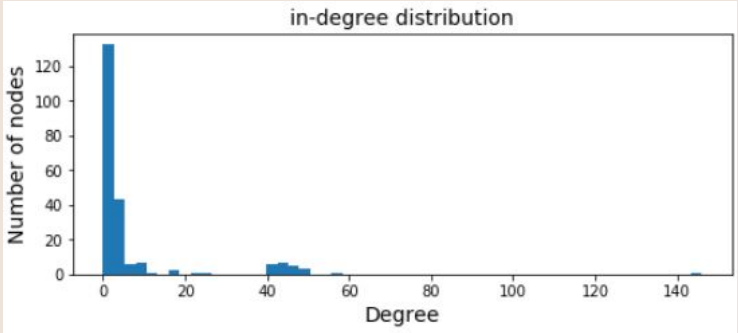
Basic characteristics

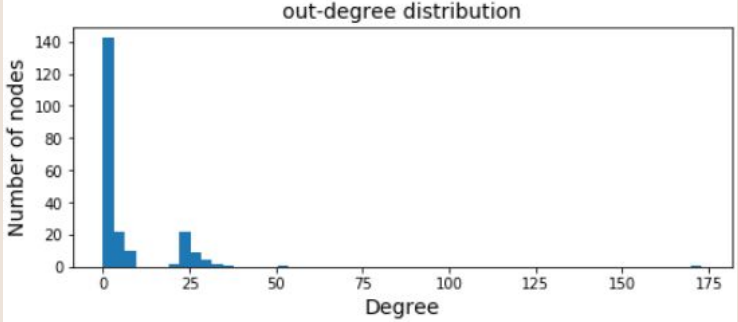
	French	English
Nodes	124	216
Edges	1063	1633
Ratio nodes/edges	0.12	0.13
In degree	8.57	7.56
Out degree	8.57	7.56
Total degree	17.14	15.12

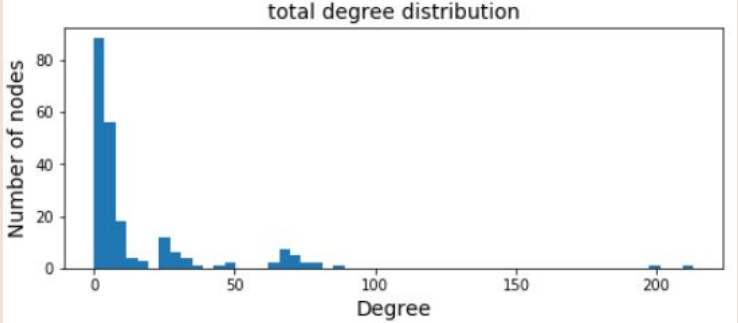

in-degree distribution
Number of nodes
Degree

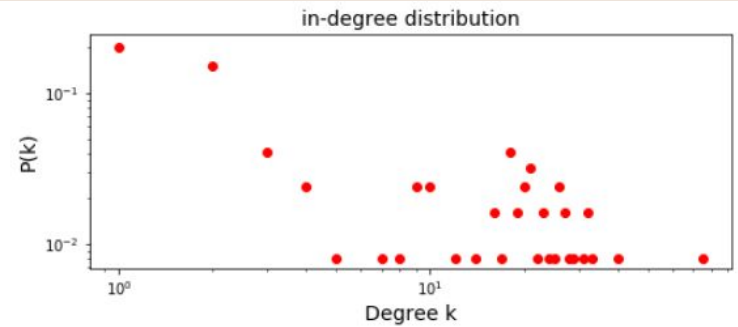

out-degree distribution
Number of nodes
Degree

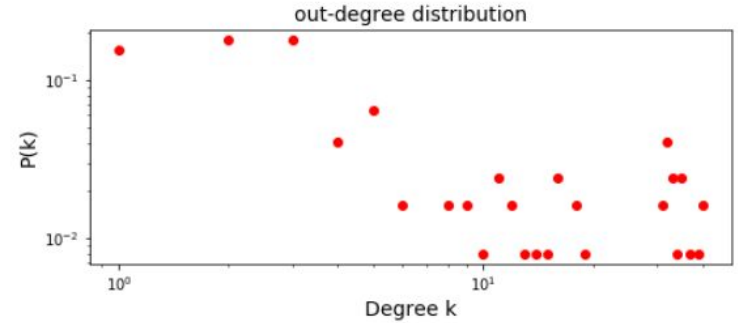

total degree distribution
Number of nodes
Degree

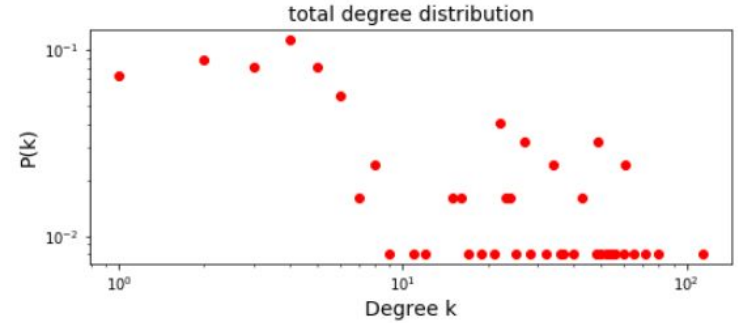

in-degree distribution
Number of nodes
Degree

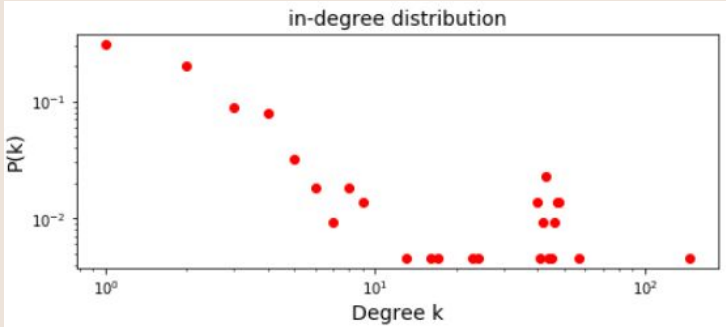

out-degree distribution
Number of nodes
Degree

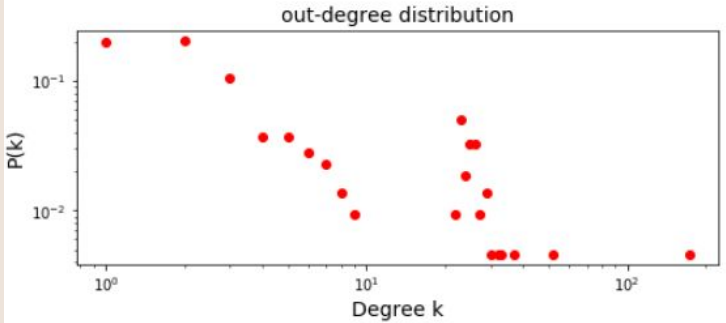

total degree distribution
Number of nodes
Degree

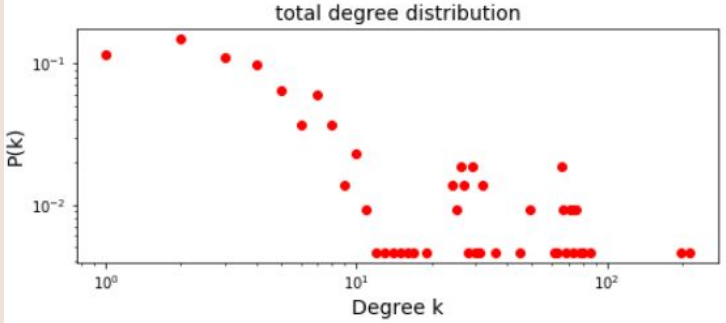

in-degree distribution
P(k)
Degree k


out-degree distribution
P(k)
Degree k


total degree distribution
P(k)
Degree k


in-degree distribution
P(k)
Degree k

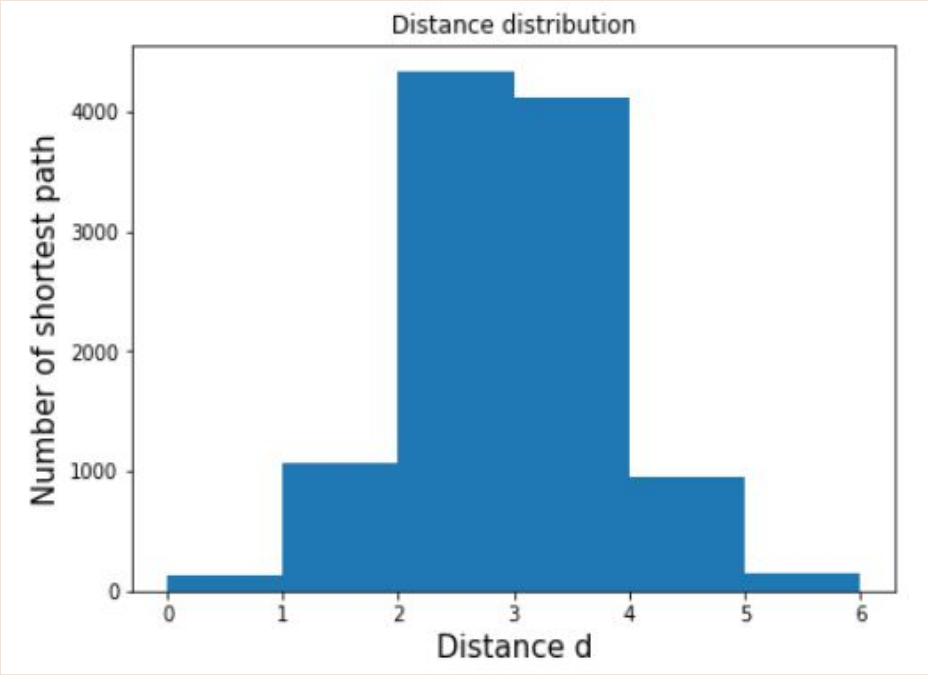

out-degree distribution
P(k)
Degree k


total degree distribution
P(k)
Degree k

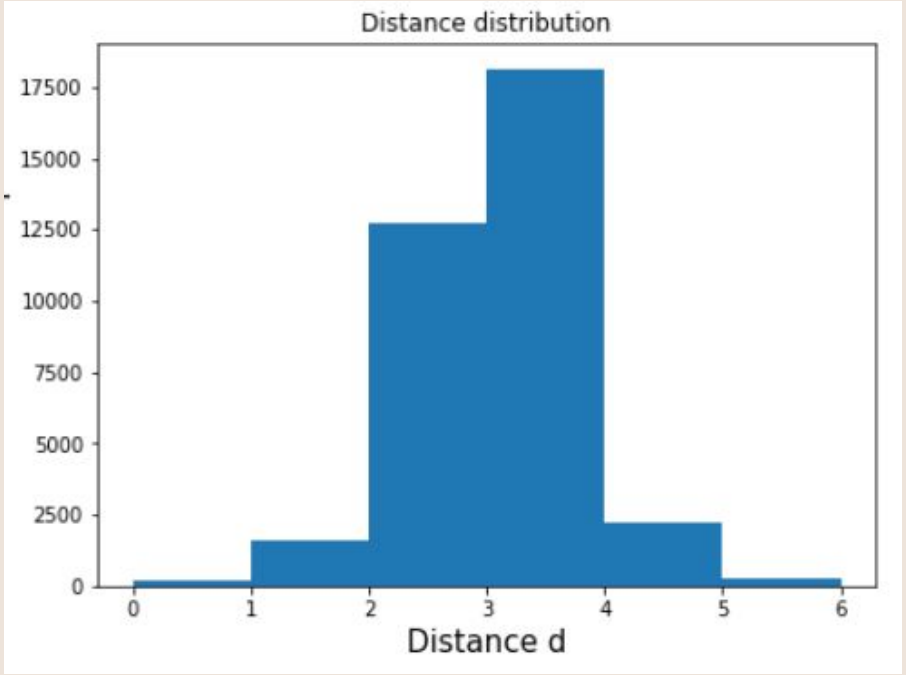
Top 10 total degree nodes

	title	degree
0	Apprentissage automatique	75
1	Réseau de neurones artificiels	40
2	TensorFlow	33
3	Keras	32
4	Apprentissage supervisé	32
5	Theano (logiciel)	31
6	Méthode des k plus proches voisins	29
7	Apprentissage non supervisé	28
8	Microsoft Cognitive Toolkit	27
9	Scikit-learn	27

	title	degree
0	Machine learning	146
1	Statistical classification	57
2	Convolutional neural network	48
3	Statistical learning theory	48
4	Computational learning theory	48
5	Machine Learning (journal)	47
6	Empirical risk minimization	47
7	Unsupervised learning	47
8	Semi-supervised learning	46
9	Dimensionality reduction	46



Mean: 2.48
Variance: 0.78



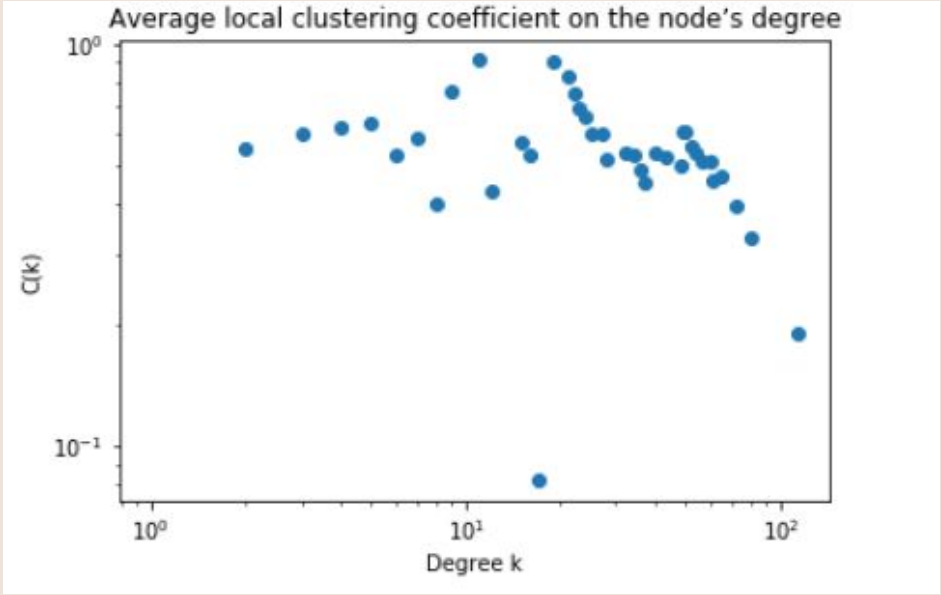
Mean: 2.61
Variance: 0.55

Connected components

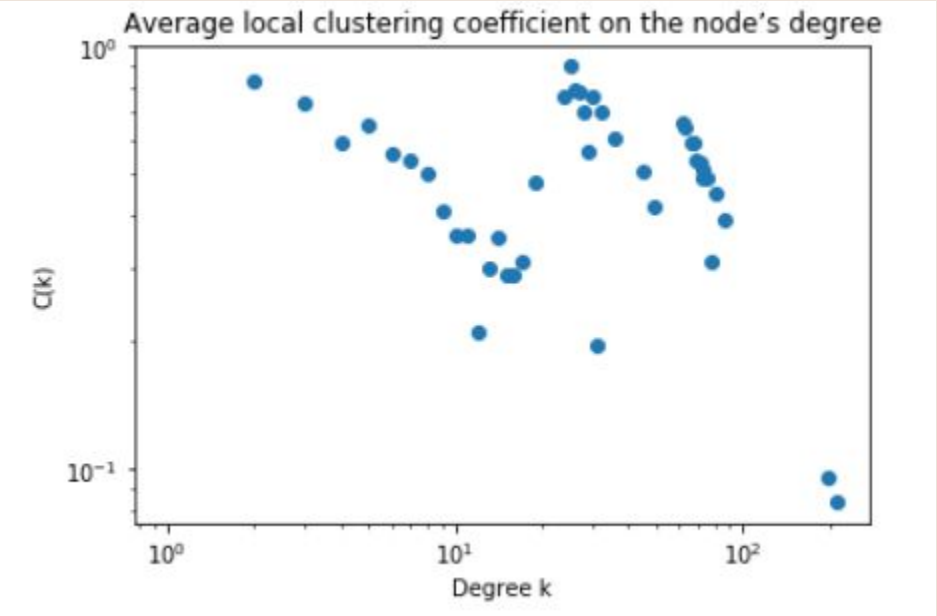
	French	English
Weak	8	8
Multi-node, weak	1 (117 nodes)	1 (209 nodes)
Strong	31	51
Multi-node, strong	3 (2-3-91 nodes)	1 (166 nodes)

Main connected components characteristics

		French	English
Average shortest path length	**Weak cc**	1.96	2.11
	Strong cc	2.46	2.58
Diameter	**Strong cc**	6	6



Average clustering coeff: 0.5



Average clustering coeff: 0.53

Conclusion

Similarities:

- Directed and disconnected
- Scale free networks
- Small-world structure

Differencies:

- More nodes in English than in French

Next steps

Metadata exploration

Community detection

Bipartite En-Fr Network analysis

Notebook and wpnetwork module on my GitHub

github.com/Ninanouchka/wikipedia-article-network

Bibliography

1. Barabási, Albert-László. *Network Science*. Consulté le 17 décembre 2019. http://networksciencebook.com/.
2. Arenas, A., L. Danon, A. Díaz-Guilera, P. M. Gleiser, et R. Guimerá. « Community Analysis in Social Networks ». *The European Physical Journal B* 38, nº 2 (1 mars 2004): 373-80. https://doi.org/10.1140/epjb/e2004-00130-1.
3. Fujiwara, Yuya, Yu Suzuki, Yukio Konishi, et Akiyo Nadamoto. « Extracting Difference Information from Multilingual Wikipedia ». In Web Technologies and Applications, édité par Quan Z. Sheng, Guoren Wang, Christian S. Jensen, et Guandong Xu, 496-503. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2012. https://doi.org/10.1007/978-3-642-29253-8_42.
4. Massa, Paolo, et Federico Scrinzi. « Manypedia:ComparingLanguagePointsof Viewof Wikipedia Communities », s. d., 9. https://www.opensym.org/ws2012/p13wikisym2012.pdf
5. Pfeil, Ulrike, Panayiotis Zaphiris, et Chee Siang Ang. « Cultural Differences in Collaborative Authoring of Wikipedia ». Journal of Computer-Mediated Communication 12, nº 1 (1 octobre 2006): 88-113. https://doi.org/10.1111/j.1083-6101.2006.00316.x.
6. Nemoto, Keiichi, et Peter A. Gloor. « Analyzing Cultural Differences in Collaborative Innovation Networks by Analyzing Editing Behavior in Different-Language Wikipedias ». Procedia - Social and Behavioral Sciences, The 2nd Collaborative Innovation Networks Conference - COINs2010, 26 (1 janvier 2011): 180-90. https://doi.org/10.1016/j.sbspro.2011.10.574.