Network Tour of Data Science

Project proposal

1. Story (20 points)

The concept of a global language system was developed by Dutch sociologist Abram de Swaan in 2001 in his book ***Words of the World: The Global Language System.*** According to him,"the multilingual connections between language groups do not occur haphazardly, but, on the contrary, **they constitute a surprisingly strong and efficient network** that ties together – directly or indirectly – the six billion inhabitants of the earth."In his book, the world's languages are divided into a hierarchy consisting of ***four levels***, namely the *peripheral, central, supercentral and hypercentral* languages.

Our research study is trying to analyze the global languages system from the network sciences viewpoint and see if there are indeed some strong support for the global language system theory and eventually improve or build a new model of the global language system.

The data that we choose to use:

3 datasets from Global Language Network project by the MIT Media Lab:

1. Books published and translated from each language,
2. Tweets Published in each language(trans-language is taken into account)
3. wikipedia pages in each language

Some extra datasets which might be useful:
1. Famous Individuals in Wikipedia by Language
2. Famous Individuals in Wikipedia by Country
3. Language Speakers and Families

The tools that we choose to use:

- ***clustering (spectral clustering, k-means)***
- ***dimensionality reduction (PCA, MDS, LLE, ISOMAP, Laplacian eigenmaps, t-SNE)***

- relevance of chosen data and tools
  - Can the data answer the question or support the product?
  - Are the chosen tools relevant?

2. Acquisition (10 points)

In our case, the data has a clear graph structure, where each node represent a language and edges should be links between two languages. There are many different ways to define the link of two languages thus many different ways to construct a language network such as to build links between languages based on the similarity between languages, or build links based on the concurrence of languages.(spoken in a same area). We plan to build **several networks** based on different algorithms and **each network** may represent a aspect of the world languages system. And a comparison among the proprieties of these networks could be done.

3. Exploration (20 points)

In order to have a better graph proprieties , we may first only look at the language with number of users larger than a predefined threshold(like 100 000). A set of classical network proprieties will be analyzed:

- some properties of the graph (e.g., connected components, sparsity, diameter, clusters, degree distribution, spectrum)
- identify the type of graph (e.g., power law, small world, regular, sampled manifold)
- some properties of the nodes (e.g., clustering coefficient, modularity, centrality)
- some analysis of the attributes (e.g., their distribution, smoothness, graph Fourier transform)
- a vizualization of the network
- a reflection on the insights

4. Exploitation (30 points)
- at least one of the following tools, seen during the lectures, is used:
  - ***clustering (spectral clustering, k-means)***
  - graph Fourier transform
  - regularization (graph Tikhonov, graph total variation)
  - ***dimensionality reduction (PCA, MDS, LLE, ISOMAP, Laplacian eigenmaps, t-SNE)***
  - graph filters (Chebyshev, ARMA)
  - graph neural networks
- critical evaluation of the results
  - subjective or objective (baseline, existing work)
  - state the limitations: to what extent did you answer the question or provide a good product
- example bonus: use multiple relevant tools, or tools beyond what was seen in class

Annexes:

	global language system by Abram de Swaan
Hypercentral languages	English
Supercentral languages	Arabic, Chinese, English, French, German, Hindi, Japanese, Malay, Portuguese, Russian, Spanish, Swahili and Turkish.
Central languages	Bulgarian, Dutch, Persian, Italian etc.
Peripheral languages	Javanese, Kurdish, Irish etc

Datasets:

For this project, we will use the dataset from Global Language Network project by the MIT Media Lab

The Global Language Network project provides with 3 datasets to characterize the link among languages:Books published and translated from each language, Tweets Published in each language(trans-language is

taken into account) and wikipedia pages in each language. In our network, nodes will be the languages and links will be the connection between two languages.

4. Different ways to define similarity between two languages.
5. How to characterize the influence of a language from the viewpoint of network sciences.
6. Wikipedia and Tweets are online language sources, are their topological proprieties similar to that of book publishing?
7. Are the conclusions driven from these datasets aberrant to some actual facts? Will these because of the bias from the data?