Project Progress Review #3

(Interim Exploration Progress)

Project Title : Linguistic Analysis of Indo-European Languages
Project ID : PW19SMP003
Project Guide : Prof. Shreekanth M Prabhu
Project Team : Roshan U[01FB15ECS246],
Sanath Bhimsen[01FB15ECS260],
Mukesh M Karanth[01FB15ECS361].

Project Abstract and Scope

This project is a research oriented project which deals with linguistic analysis of Indo-European Languages using Social Network analysis.

We use data set that contains words from multiple languages to perform similarity measures and centrality measures between the words of different languages to find hidden links between languages.

The scope of the project is subject to the project being a minor project with time constraints, hence we are making use of transliterated words, the number of words is limited to a max of 200 and we are using a select number of centrality and similarity measures.

Data Set

English	German	Hindi	Italian	Latin	Spanish	French	Russian	Sanskrit
the	das		ila	quod	la	la		
of	fon	ka	di	autem	de	de	iz	
to	zu	seva mere	aa	ut	a	à	vey	
and	unt	tatha	a	et	y	et	a takzhe	tu
a	ein	e	un	autem	uno	une		
in	im	mein	nela	apud	en	dans	vey	
is	is	hai	eh	is	es	est	yavlyayetsya	
it	est	yaha	isso	it	eso	il	Eto	
you	zie	aap	tu	vos	thu	vous	vy	bhavantham
that	das	us	quello	quod	ese	cette	tot	yath
he	er	vah	lui	quod	el	il	on	
was	waar	tha	era	was	estaba	était	Bylo	
for	zum	ke liye	pera	quia	para	pour	za	hi
on	auf	par	sopra	in	en	sur	na	
are	zind	kar rahe hain	siamo	sunt	son	sont	yavlyayutsya	
with	mit	saath mein	con	apud	con	avec	sa	sardham
as	wie	jaisa	come	quod	como	comme	kak	
I	ish	main	io	ego	yo	je	ya	

Research Design Approach

Design approach: **Top-Down Approach**

Benefits:

- **Decreased Risk:** Since the approach is planned well in advance.
- **Good Organization:** Tasks are determined and filtered down without any confusion because project goals are set and will not be affected by outside opinions.
- **Minimized Cost:** Members are free to complete their own tasks unique to their role in the project and aren't saddled with the responsibility of setting project-wide goals.

Drawbacks:

- Limited Creativity: Members are engrossed in their responsibilities and are unable to contribute innovations/ideas to the project — sometimes leading to frustration and a lack of motivation to perform.
- Slow Response to Challenges: When a challenge arises as a result of a decision, it can take time for the members to establish a solution because there are limited minds contributing to decisions.

Design Constraints, Assumptions & Dependencies

DEPENDENCIES:

-> The usage is strictly limited only to study the effects and results of social network analysis on Linguistics and its representations.

-> The project requires packages which are good for visualization purposes and have good functionality for Network Analysis.

ASSUMPTIONS & CONSTRAINTS:

->A single language was taken from each of the lineages of the Indo-European Language, Proto-Indo-European.

->The number of words chosen to represent each language were all transliterated in English.

-> The maximum number of words in a single language is restricted to 200.

Interim Exploration

Basic Approach:

- We formulated the project idea and established the end-user requirements
- Then, we designed the project's features and functionality for the user requirements .
- Next, we decided upon the technologies that would be necessary to envision our project.
- Finally, we divided the work amongst the members and started working on our parts to get the project work done.

Is there a need for changing the approach?

No, The development of the project is going forward smoothly as all the members are aware of their responsibilities and the tasks are completed on time.

Technologies Used

TECHNOLOGIES:

-> Applications:

-Google Translate API: Used to retrieve transliterated words quickly and in bulk.

-Jupyter Notebook: For its GUI and ease of use.

-Rstudio: consolidated representation of results, command prompt and single shot execution of programs.

-> Languages:

- Python: Used because of its ease of programing and amazing libraries that provide wide range of functionality in analytics.

- R : Used for visualisation because of its packages that aid good visualisation and its simplicity.

Project Progress So far

Percentage completion of Project: ~60%

Progress from last time:-

- ✓ Added Russian Language instead of Persian.
- ✓ Included some of the suggestions from the panel, which include replacing few transliterated words with their actual pronunciation in the case of German, Spanish and Italian.
- ✓ Performed similarity measures and computed similarities between various languages and within same languages.

Project Demo

Project Demo:

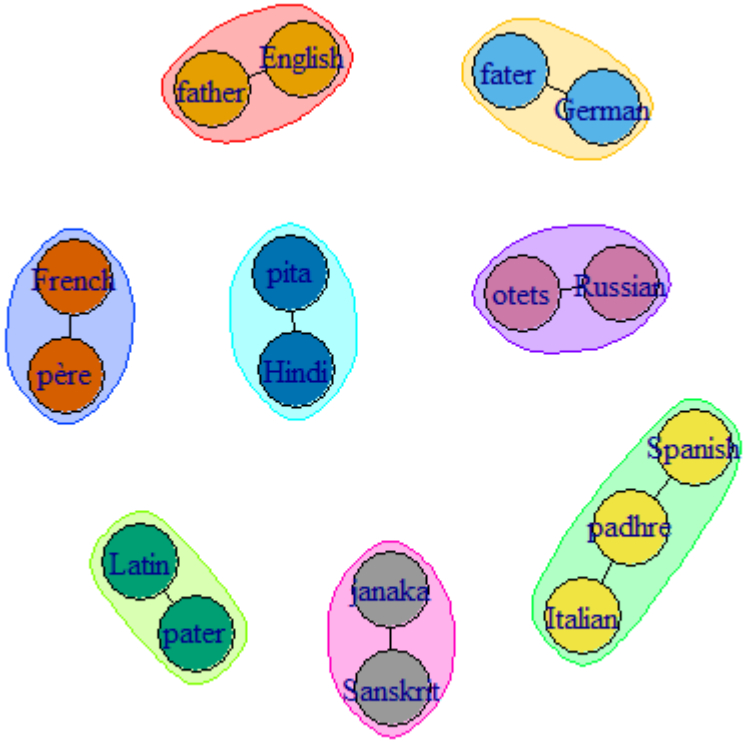
- DataSet Creation
- Python Similarity Calculations
- R Visualisation

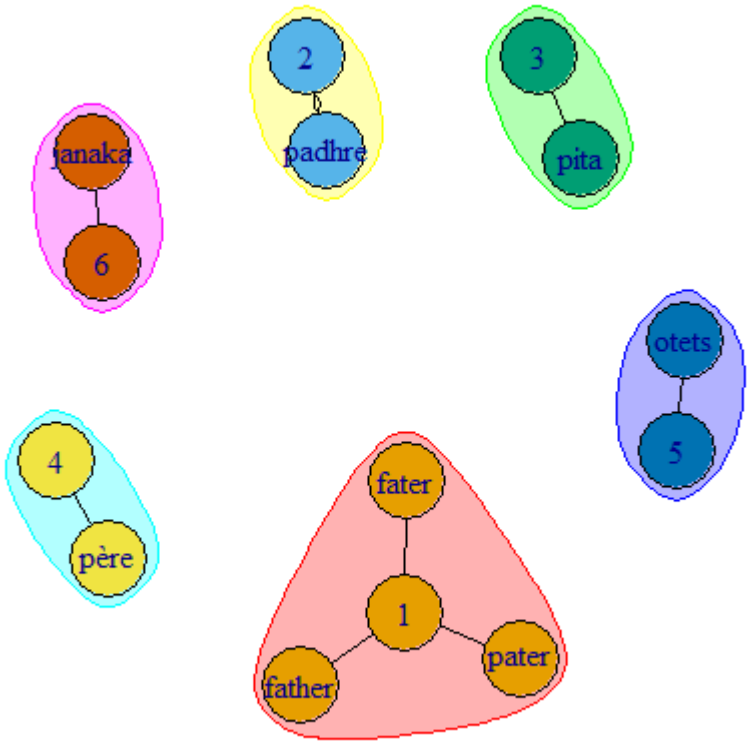
Output

```
{1: ['father', 'fater', 'pater'],
 2: ['père', 'padhre', 'padhre'],
 3: ['pita'],
 4: ['otets'],
 5: ['janaka']}
```

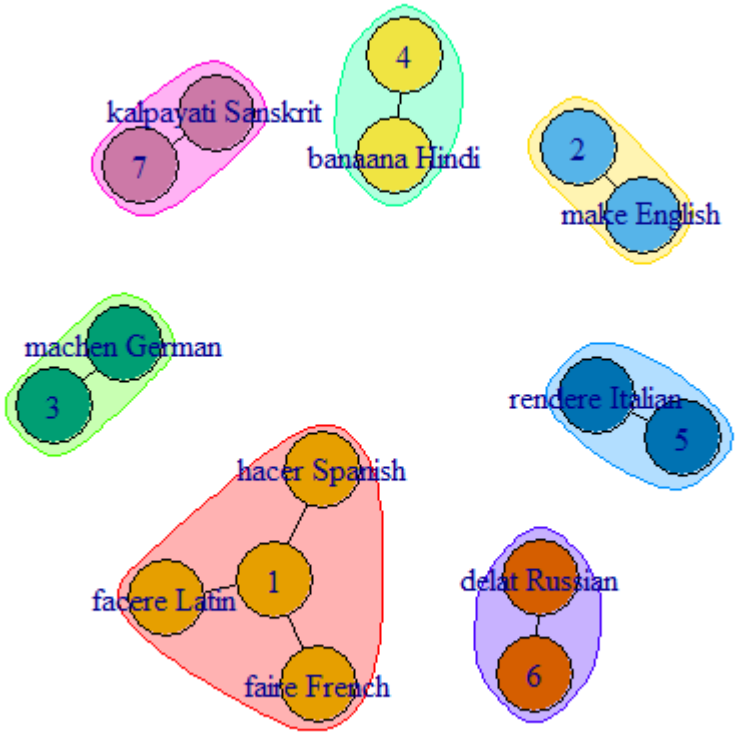
Cluster	Language	Word
1	English	father
1	German	fater
1	Latin	pater
2	Italian	padhre
2	Spanish	padhre
3	Hindi	pita
4	French	père
5	Russian	otets
6	Sanskrit	janaka


PES

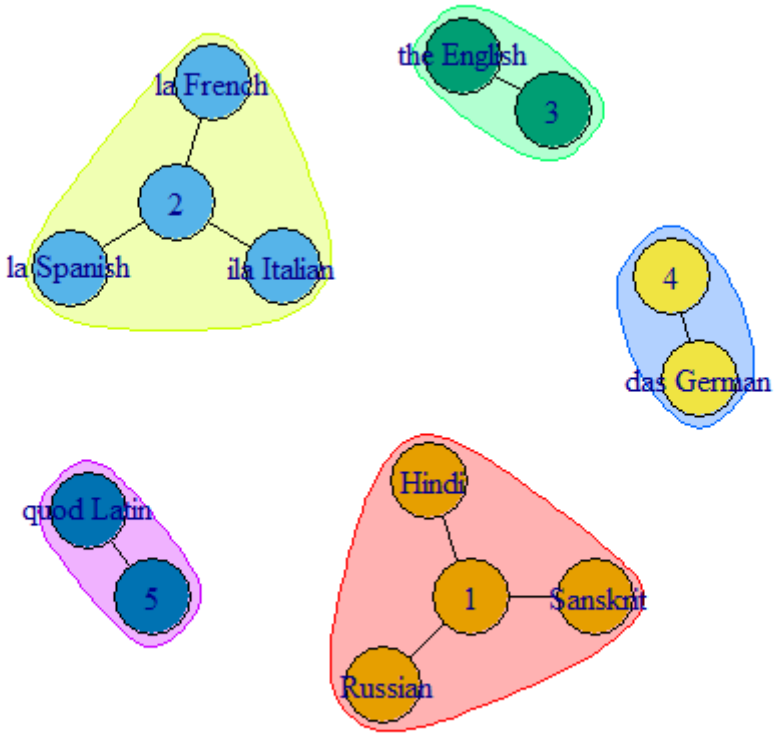

English
father
fater
German
French
père
pita
Hindi
otets
Russian
Spanish
padhre
Italian
Latin
pater
janaka
Sanskrit



Clusters for the word “Father”

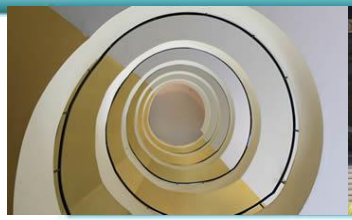


Clusters for the word “Make”



Clusters for the word “The”

PES UNIVERSITY






Thank You