Project Progress Review #2

(Customer Requirement Specifications)

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 heavy 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.




Further Literature Survey

- **The Origins of Indo-European Languages** Colin Renfrew *Scientific American* Vol. 261, No. 4 (OCTOBER 1989), pp. 106-115
- Mapping the Origins and Expansion of the Indo-European Language Family **Remco Bouckaert, Philippe Lemey,** *Science* 24 Aug 2012: Vol. 337, Issue 6097, pp. 957-960 DOI: 10.1126/science.1219669

User Characteristics

User End Requirements:

-> Analysis of similarity and centrality measures on atleast 5-6 languages.

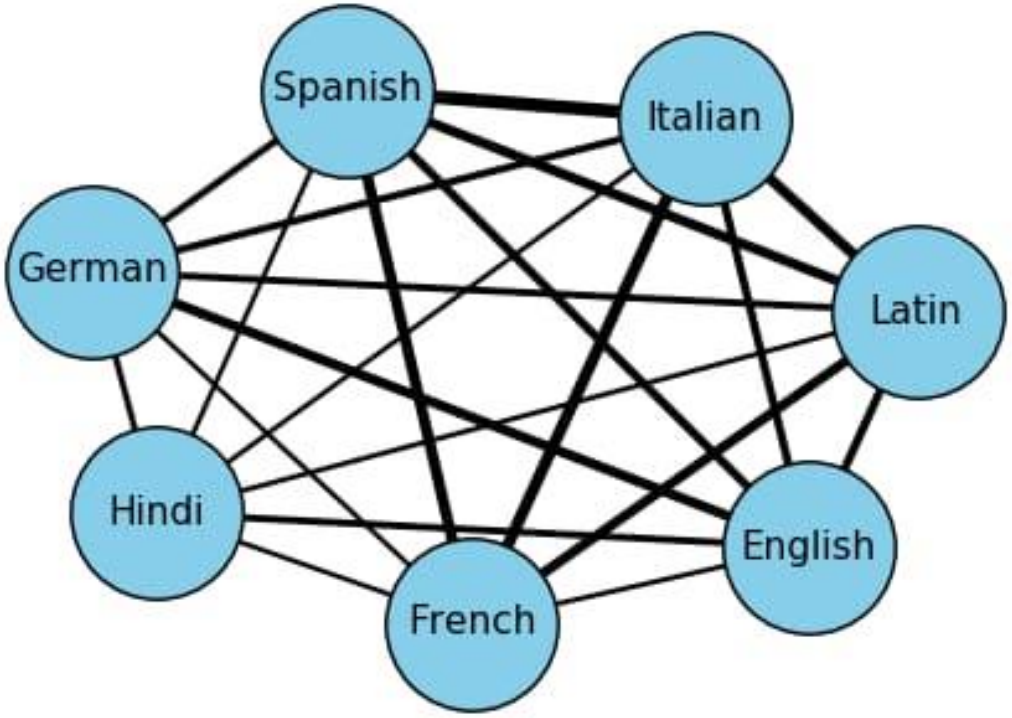
-> Visualisation of the developed network layout as a result of the analysis.

-> A User Interface where the user can actively contribute to addition of new words to improve quality of the vocabulary.

->A Better Understanding of the Indo-European Language Structure.

Data Set

Weighted Graph Showing Similarities between Languages



Dependencies / Assumptions / Risks

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:

->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.

RISKS:

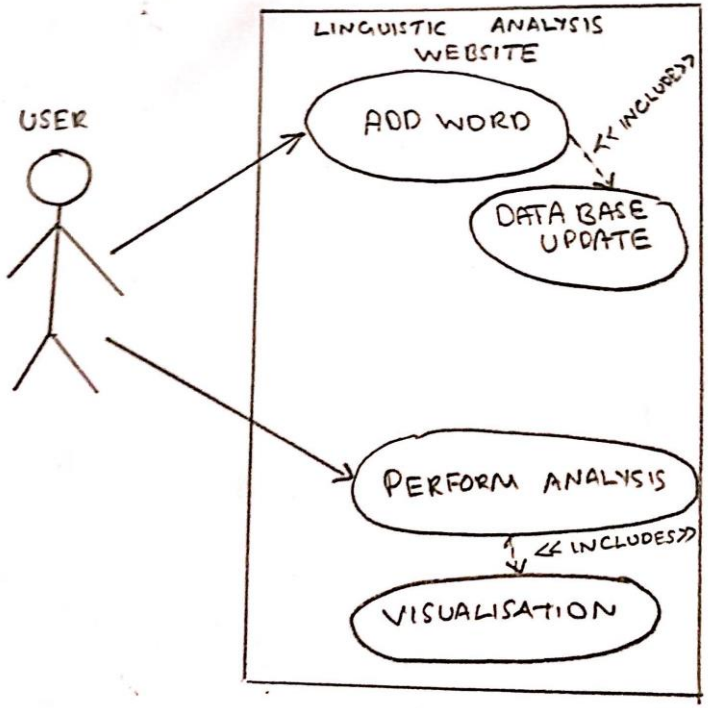
-> The limited number of centrality and similarity measures used for analysis could potentially affect the accuracy of the visualisation.

-> The limited pool of words chosen might not be sufficient enough to convey any good results.

-> The words chosen might not be the best choice to best depict all languages perfectly.

UI/ Use Case

USE CASE DIAGRAM



Modules

FEATURES:

-> Live Data set: Constantly updated by public and monitored by team.

-> Visualisations:

- Inter language representations
- Single Language visualisations
- Overall Network Structure

MODULES:

-> Centrality measure modules like closeness, between-ness, etc.

-> Similarity measure modules like Levenshtien edit distance, etc.

-> visualization modules.

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.

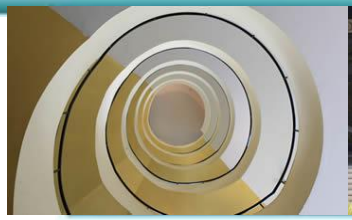
-Rstudio: consolidated representation of results, command prompt and single shot execution of programs.

-> 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.

PES UNIVERSITY






Thank You