College of Computer Studies

DEPARTMENT OF
SOFTWARE TECHNOLOGY

NLP1000

Introduction to Natural Language Processing

Project 2: Similarity Matrix and Language Clustering

Instructor : Nathaniel Oco
nathaniel.oco@dlsu.edu.ph

Course Site/Repository : AnimoSpace, Google Drive

Description

This project challenges students to explore relationships among languages using computational tools and techniques. By leveraging similarity measures, clustering algorithms, and any available data (e.g., transport networks), students will investigate which languages appear closely related and provide linguistic or sociocultural insights into the observed patterns. Students are free to use publicly available corpora and construct their own feature sets. Approaches may include, but are not limited to:

- Generating word lists across languages and computing lexical similarity (e.g., Levenshtein distance).
- Building n-gram profiles from raw corpora to measure character-level or token-level distributional similarity.
- Using dimensionality reduction and/or unsupervised clustering to visualize relationships.
- Exploring typological features or shared traits (historical, orthographic, phonological, geographical).

Deliverables

1. PDF file. Written report. Contents: Data sources, preprocessing, and feature engineering; Methodology for similarity and clustering; Similarity matrix and language clusters / dendrogram; Interpretation of matrix and tree; Evaluation; Conclusion and limitations; Declaration of AI usage (per person).
    a. Similarity Matrix. A computed matrix showing similarity or distance scores between Philippine languages. The number of languages assigned is equal to $8 + 2n$, where n represents the number of group members. The groups are allowed to include Spanish and English.
    b. Language clusters or Dendrogram. A visual representation of how the selected languages cluster.
2. Demo video. A narrated walkthrough with screen recording that summarizes the project, methodology, key results (matrix and tree), and insights.
3. Zip file. Source code, training data, feature set.

Provide a declaration of AI usage (per person)

1. AI tool(s) used
2. Describe briefly how you used the AI tool(s) in the completion of this assignment. Examples include: idea generation, grammar checking, paraphrasing, summarizing references, formatting help, etc. Provide sample prompts and outputs.
3. Extent of Use:
    a. Minimal – used only for grammar/spell check.
    b. Moderate – used for inspiration, then reworded significantly.

In 2–3 sentences, reflect on how the use of AI contributed to your learning. If no AI was used, reflect on why you chose not to.

Rubric

Criteria	Exemplary Full points	Satisfactory ¾ points	Developing ½ points	Beginning ¼ points
Data selection (10 points)	Total corpus size is 50,000 words per language and data sources are comparable	At least ¾ of the target corpus size	At least one half of the target corpus size	Less than 1,000 words
Computational Output & Methodology (60 points)	Output is reproducible, well-documented, and evaluated using standard metrics against a gold standard. Methodology is sound, justified, and effectively applied. Code and tools are appropriate, clearly explained, and properly referenced.	Output is mostly correct and reproducible. Methodology is appropriate but only partially justified.	Output is incomplete, inconsistent, or poorly explained. Methodology is vague or weakly justified.	Output is incorrect or missing; methodology absent or not credible.
Written report and demo video (30 points)	Clear, well-structured, and comprehensive. Strong introduction, methodology, results, and discussion with linguistic/sociocultural insights. Proper citations included.	Clear and organized with most required sections; some analysis but limited depth. Minor issues in clarity or citations.	Partial structure; limited analysis or weak discussion of results. Minimal sources.	Unclear, incomplete, or lacking analysis and discussion.

Bonus:

- The project includes the Yami language (romanized): +5 points
- The project incorporates features beyond orthographic data, such as geographical information (e.g., KWF language maps) or transport networks: +10 points

Deadline

October 25, 11:30 am.