CSC 4792 DATA MINING PROJECT REPORT

IDENTIFYING MISSING SECTIONS IN ZAMBIAN WIKIPEDIA PAGES

AUTHORS: MATTHEWS BWALYA MUMBA, PAUL MBAMBO, LUPANDU MASUMBA, OCEAN MBIZA, FIDELIA LUBINDA

DATE: 09-14-2025

Abstract

This project implements a CRISP-DM based modeling and evaluation pipeline to detect missing or underdeveloped sections on Zambian related English Wikipedia pages. Using a processed corpus of Zambia related articles, the team built a rule-based completeness scoring system and ran supervised classification experiments to predict the presence/adequacy of sections. This report summarizes data, methods, results, limitations and deployment recommendations.

Introduction

Wikipedia is widely used as a knowledge base and reference for contextual information about countries, institutions and public figures. Preliminary exploration suggested gaps in the English-language coverage of Zambia, i.e, several important topical sections (e.g., Infrastructure, Transportation, Environment, Health care, Education, Tourism, Military) are missing or underdeveloped in many Zambia-related pages. The aim of this modeling work is to systematically detect such gaps so that editors and volunteers can prioritize improvements. The technical goals for the Modeling phase were:

● Produce an automated process to detect missing sections in Zambian Wikipedia pages

●Implement a reliable method to compute a **completeness score** for each Zambia-related page.

●Train and evaluate classification models that predict missing or underdeveloped sections from pages/articles

This work maps directly into the Modeling and Evaluation phases of CRISP-DM.

Data Preprocessing And Implementation

Data Sources, the notebook loads a combined Wikipedia dataset and performs cleaning steps. According to the notebook outputs, the pipeline identified **27 duplicate articles** and produced a cleaned dataset with shape **627 rows** (articles) across relevant fields (title, wikitext, url, category). The dataset was saved to a combined CSV in the experimental runtime.

Data Cleaning, Duplicate articles were detected and removed. 27 duplicate articles were found and after cleaning approximately 627 articles remained, shape was (627,4). Some notebook cells raised minor warnings (pandas **SettingWithCopyWarning**) and at least one **NameError** was observed in a recommendation cell related to **top_features** articles.

Preprocessing And Normalization

●The notebook extracts Wikipedia section headings from each page and normalizes them using typical steps that is; lowercasing, punctuation removal, stemming/lemmatization, tokenization and stop word removal.

●Non informative headers such as **References, External links, and See also** were filtered out

Feature Engineering

●Binary presence features: for a curated list of *expected sections* (History, Geography, Demographics, Economy, Culture, Politics, Infrastructure, Transportation, Education, Health care, Environment, Tourism, Military), a 0/1 flag was created per article indicating presence/absence of the heading.

●Text features: TF–IDF vectors were computed on article wikitext/section text to support supervised models that infer implicit coverage even when headings differ.

Modelling Approach

Rule-based Completeness Scoring

An expected sections schema (e.g., History, Geography, Demographics, Economy, Culture, Politics, Infrastructure, Transportation, Education, Health care, Environment, Tourism, Military) is used.

For each page, presence/absence of expected sections is computed to produce a **Completeness Score** = (Number of present expected sections) / (Number of expected sections).

This score provides a simple, interpretable ranking of pages that need attention.

Supervised classification

Models explored (the notebook contains training code and evaluation cells):Decision Tree, Random Forest, Logistic Regression, and the (Notebook includes code cells that compute confusion matrices and classification metrics precision, recall, F1, accuracy.)

These classifiers are intended to complement the rule-based approach: they can detect semantic coverage when content exists but is not labeled under canonical headings.**Evaluation Strategy And Metrics**

For supervised models, standard metrics were used; Accuracy, Precision, Recall, F1-score. These metrics are appropriate for evaluating binary presence/absence classification of sections.

For completeness scoring: simple descriptive statistics and ranking (e.g., distribution of completeness scores across pages), and targeted manual inspection of top gap pages.

Completeness scoring: Pages are ranked by completeness score and visualized (bar charts/boxplots) to show the distribution of coverage across Zambia related pages.

Test Design

Train/test split on labeled data, Cross-validation for hyperparameter tuning when applicable (the notebook contains code scaffolding to do this). And Confusion matrices were produced for model inspection.

Data Summary

The dataset prepared in the notebook contains several hundred Zambia-related articles (approximately 627).

Many Zambian pages are partial/stubby; repeated issues encountered include missing sections and short section lengths (underdeveloped content).

Commonly missing or underdeveloped areas flagged during the project include: **Infrastructure, Transportation, Environment, Health care, Education, Tourism, and Military.** Which is consistent with earlier exploratory work.

The rule-based completeness score is an effective, interpretable first pass indicator of content gaps. Supervised classifiers can support this by detecting implicit content gaps when sections exist but lack substantive content.

Supervised Model Performance

The notebook shows that ensemble models (Random Forest) generally perform better for this problem; Decision Trees provided interpretable rules but tended to perform less strongly than ensembles. The classification experiments produced confusion matrices and metrics.

Recommendation Function

The **fxn_recommend_sections** function returns prioritized section suggestions for a given page's cleaned headings. It works on the rule-based logic and is designed to optionally incorporate top feature signals from classifiers (once **top_features** is available).

Discussion And Interpretation

Interpretability vs Complexity: The rule-based completeness score is straightforward and directly actionable for editors (a checklist). Supervised models add value by detecting semantic coverage even when headings differ, however they require labeled data and careful error analysis.

Common content gaps: The repeated absence of infrastructure and sectoral sections (health/education) suggests underrepresentation of civic/institutional content for Zambia on English Wikipedia.

Actionability: The output is most useful as a prioritized checklist: editors can focus on high-impact pages with large coverage gaps and add canonical sections with sourced content.

Limitations

●**Subjectivity of expected schema:** The canonical list of expected sections is partly subjective; Zambian Wikipedia community norms should be consulted before enforcing a schema.

●Several pages are stubs (very short, lacking major sections).

●Inconsistent heading naming and structural heterogeneity (e.g., Transport vs. Transportation, Health vs. Health care).

●**Data freshness:** Some pages may be outdated; last-updated timestamps should be included when prioritizing edits.

Model generalizability: Supervised models depend on labeled data, small or imbalanced datasets can reduce performance.

Conclusion

This project demonstrates a CRISP-DM-based approach to detect missing sections on Zambia-related Wikipedia pages using a combination of rule-based completeness scoring and supervised classification.

References

1. B. Stvilia, V. Twidale, L. Gasser, A. Smith. *Assessing information quality of a community-based encyclopedia.* ICIQ, 2008.
2. P. Chapman et al. *CRISP-DM 1.0: Step-by-step data mining guide*, 2000.
3. F. Pedregosa et al. *Scikit-learn: Machine Learning in Python.* JMLR, 2011.
4. B. Luyt, K. Tan. *Representing the historical: Wikipedia and the past.* JASIST, 2013.