Organization: OCLC

Primary mentor: Jenny Toves, Software Architect, OCLC Research

Project Title	**Detecting Missing Marks (diacritics) in Text**
Description	Analyze patterns of diacritic usage in MARC records. Using the facets language of item, language of cataloging, presence of 880 fields, age of record and cataloging rules, look for patterns to predict which records are likely to be missing diacritics.
Problems/ Research Questions	String data tagged by language is a foundational piece of linked data. Missing marks (diacritics) are a frequent problem in MARC records. Sometimes they are lost when transferred between systems. Sometimes they were never entered. Sometimes they are the wrong diacritic or in the wrong place. The ability to recognize and correct the problem is essential – there are too many to manually correct this problem, or address by curated lists.
Techniques	Extract title fields from Worldcat records. Examine Unicode categories of the characters to determine if marks or script data is present. Collect contextual information about the record (date of publication, date of record creation, language of item, language of cataloging, creator, holdings, etc.). Analyze collected data to see what patterns emerge.
Tools/ Languages used	Python, and/or R/RStudio (libraries to be identified), & Tableau for visualization
Data	Description: Worldcat: https://www.worldcat.org/ Data Type: MARC records Data Size: 8.5M records with item language Russian, Ukrainian or Bulgarian
Outcome	The results of this would ultimately lead to a process where records could be automatically corrected for missing diacritics. This analysis could tell us when/where

	that would likely be satisfactory. The likely next stage would be techniques to recognize expected vs unexpected diacritics.
Milestone Timeline	1) Week 1: - Establish environment - Transfer data. Extract title fields and contextual information. I can do the data transformation depending on the skill level of the fellow. 2) Month 1-2: - Analyze data to establish which is most likely to be correct, least likely to be correct and in the middle. Use the piles as training data. 3) Months 3-5: - Analyze the middle pile. 4) Month 6: - Prepare and present results.
References	- ALA-LC Romanization Tables: https://www.loc.gov/catdir/cpso/roman.html - Toves, J. and Whitacre, C. (2020): Adding Cyrillic Script to WorldCat: https://www.oclc.org/research/presentations/2020/113020-adding-cyr illic-script-to-worldcat.html