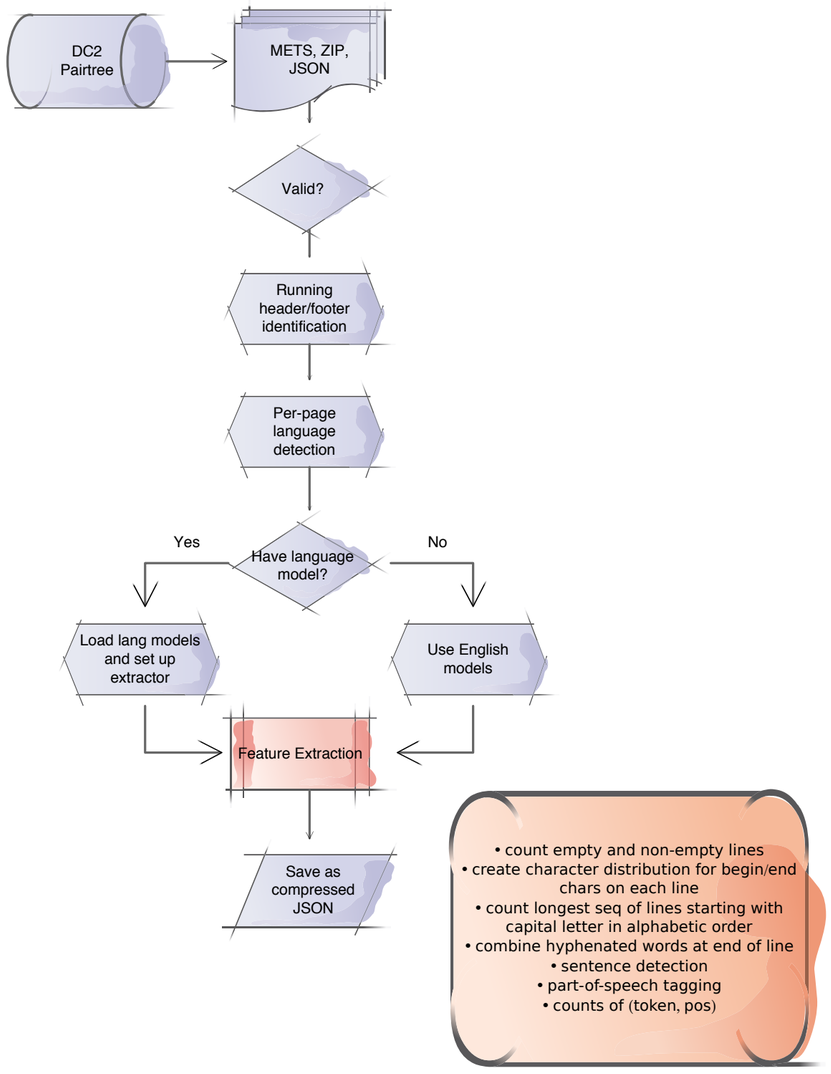
DC2
Pairtree
METS, ZIP,
JSON
Valid?
Running
header/footer
identification
Per-page
language
detection
Have language
model?
Yes
No
Load lang models
and set up
extractor
Use English
models
Feature Extraction
Save as
compressed
JSON
• count empty and non-empty lines
• create character distribution for begin/end chars on each line
• count longest seq of lines starting with capital letter in alphabetic order
• combine hyphenated words at end of line
• sentence detection
• part-of-speech tagging
• counts of (token, pos)