Developer Documentation

Fully Support UTF-8 Only For LaTeX Files

Context and Problem Statement

The feature search for citations displays the content of LaTeX files. The LaTeX files are text files and might be encoded arbitrarily.

Considered Options

- Support UTF-8 encoding only
- Support ASCII encoding only
- Support (nearly) all encodings

Decision Outcome

Chosen option: "Support UTF-8 encoding only", because comes out best (see below).

Positive Consequences

- All content of LaTeX files are displayed in JabRef

Negative Consequences

- When a LaTeX files is encoded in another encoding, the user might see strange characters in JabRef

Pros and Cons of the Options

Support UTF-8 encoding only

- Good, because covers most tex file encodings
- Good, because easy to implement
- Bad, because does not support encodings used before around 2010

Support ASCII encoding only

- Good, because easy to implement
- Bad, because does not support any encoding at all

Support (nearly) all encodings

- Good, because easy to implement

- Bad, because it relies on Apache Tika's `CharsetDetector`, which resides in `tika-parsers`.

  This causes issues during compilation (see
  https://github.com/JabRef/jabref/pull/3421#issuecomment-524532832).

  Example: `error: module java.xml.bind reads package javax.activation from both java.activation and jakarta.activation`.