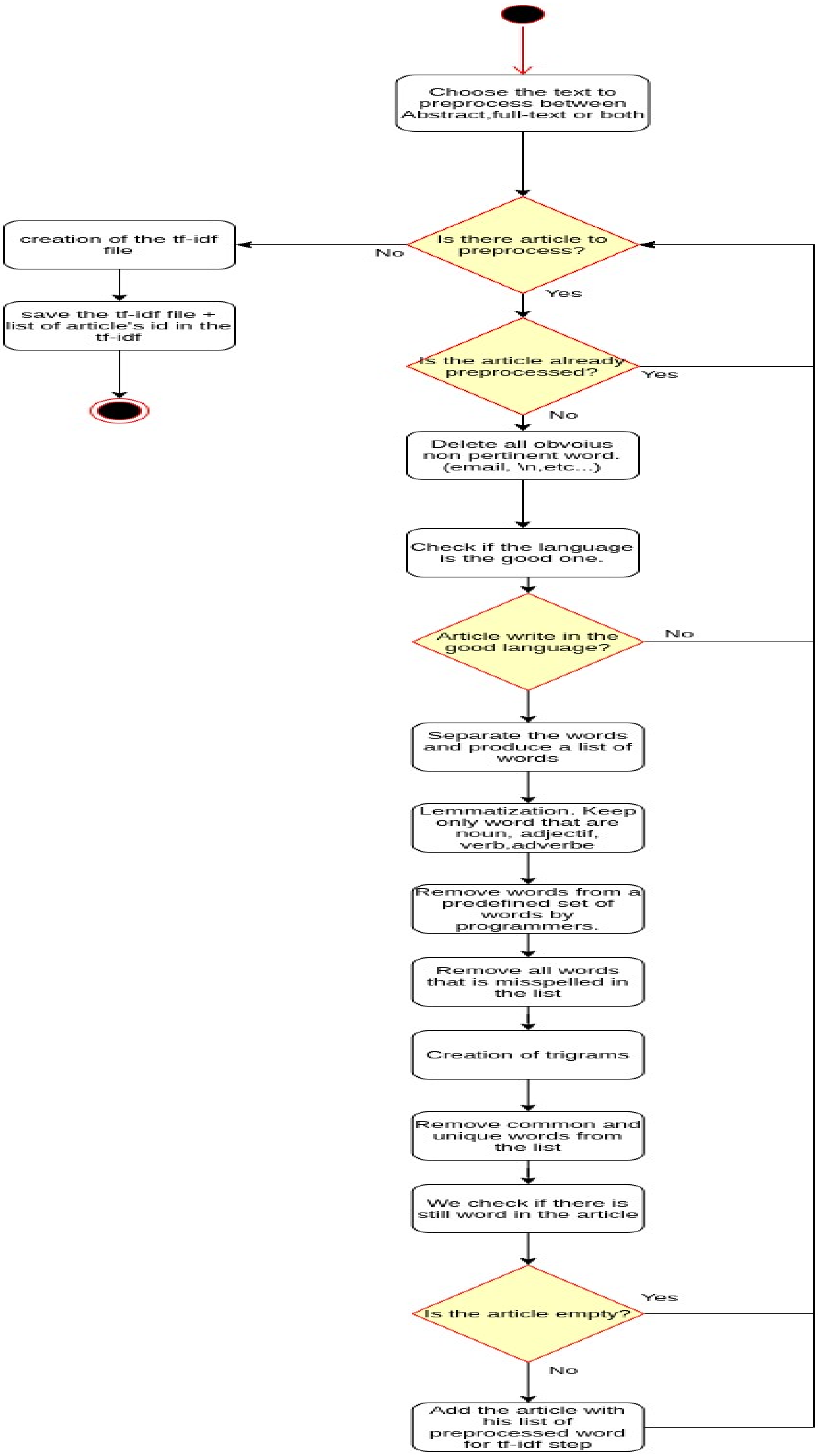
Preprocessing article

The goal is to keep only the pertinant word that represent the best the article and create a file tf-idf of all article. This will be the arguments for the article’s clustering.

At the end, we obtain a file with tf-idf and the final word list that used to create the tf-idf so we don’t have to make the preprocessing again if there are bugs. We keep a list of the id of our article so we can know which article correspond to each row of the tf-idf.

During the preprocessing, when an article is done, we keep his list of words so if there is a bug, when the preprocessing meet the article again, it continue to next and save up time.


Choose the text to preprocess between Abstract,full-text or both
Is there article to preprocess?
No
Yes
creation of the tf-idf file
save the tf-idf file + list of article's id in the tf-idf
Is the article already preprocessed?
Yes
No
Delete all obvoius non pertinent word. (email, \n,etc...)
Check if the language is the good one.
Article write in the good language?
No
Separate the words and produce a list of words
Lemmatization. Keep only word that are noun, adjectif, verb,adverbe
Remove words from a predefined set of words by programmers.
Remove all words that is misspelled in the list
Creation of trigrams
Remove common and unique words from the list
We check if there is still word in the article
Is the article empty?
Yes
No
Add the article with his list of preprocessed word for tf-idf step