TeSToP – Set of testing documents

TeSToP stands for "Testing of Support Tools for Plagiarism Detection". We want to investigate available software tools that claim to find plagiarism, in order to document their strengths, weaknesses, and limitations. The aim of this activity is twofold: (1) to expose vendors to competition and encourage them to improve their systems; (2) to inform (potential) users about limitations of these systems so they understand the reports better and don't blindly rely on numbers reported.

This guideline describes a set of documents in given language L to be used to test support tools for plagiarism detection. For each language, a set of at least four documents is to be prepared, as described below.
In order to test the ability of the systems to detect language-independent features such as automatic synonym replacement, homoglyph substitution, text-as-image obfuscation, access to open access repositories, or the ability to handle large documents, etc., a separate set of documents in English and German will be prepared. These obfuscation methods should not be used when preparing the following documents. All documents are expected to be about 4 to 5 pages long.

1) Plagiarism from Wikipedia

Take any long Wikipedia article in language L, copy it to Word. Remove all underlines, images, footnotes, and navigational matter, but keep hyperlinks (i.e. hyperlinks will still remain in the document, but they won't have any specific formatting, like blue colour or underline). Also keep in-text citations and list of references; you may use any referencing style according to your preference. Divide the article into three parts of approximately same length and mark them with sub-headings "Chapter 1", "Chapter 2", "Chapter 3" (translate the term "chapter" into language L to make sure whole document is in language L). If there are other sub-headings containing numbers, remove these numbers. The goal is that anyone who does not understand language L is able to identify these three chapters easily.
Now perform the following obfuscations:

- Leave Chapter 1, as it is – verbatim plagiarism.
- In Chapter 2, substitute 1-2 words in each sentence by their synonyms. Don't change the structure of sentences.
- Paraphrase Chapter 3, i.e. change structure of each sentence and substitute words by their synonyms.

2) Plagiarism from another source

Imagine you are a student plagiarizing an assignment in language L. Take any publicly available source in that language (e.g. a thesis publicly available from university webpage), extract a continuous part of 4-5 pages and save this as a separate document. Divide it into three parts with the same sub-headings and same obfuscation strategies as in Section 1.

Make sure you have the author's explicit consent and substantiate it (e.g. by screenshot with CC license, or an e-mail from the author). If you are the author, attach a document declaring your authorship and consent to use it for TeSToP.

3) Translation from English

Take the Wikipedia article on plagiarism detection in English (https://en.wikipedia.org/wiki/Plagiarism_detection) with all images, citations and references. Approximately in the middle of the article, there is an image *("Detection performance of CaPD approaches depending on the type of plagiarism being present"*). Leave the image exactly at this position. The image splits document into two parts. First part will be machine-translated; second part will be human-translated.
Translate the text before this image to the language L using Google Translate only. Don't make any corrections – leave the result of the machine translation as it is, even if it is terrible. On the contrary, the text below the image should be correct. You may either translate it manually or make use of machine translation. If you use machine translation, check the output and correct it to make sure it is both accurate and grammatically correct.

4) Original document

Write an original document in language L. You may use a document written previously, but make sure it was never publicly available from the Internet (including Google documents with shareable links) and it was never submitted to any plagiarism detection system. The document should be 4 to 5 pages long. Don't cite or paraphrase any other sources, whole text should be original. It does not necessarily have to make sense, but the sentences should be correct in the language L. The aim of this document is to measure false positives. Write a separate document in which you declare your authorship and consent to use this document for TeSToP.

5) Additional documents

If you have any language-specific or country-specific issues (e.g. translation from a different language then English, different alphabets, etc.), you may prepare one or two more documents in language L. Make sure that creation of these documents is documented properly and you have the consent of the authors of any original text you are using. Should you have any issues, please consult with TeSToP organizers.

General notes

Please provide all the above-mentioned documents in DOCX, PDF and TXT formats. Name the files by number (01 to 04), dash and language ISO code (EN, DE, ES…). So at the end, there will be at least 12 documents (e.g. for English 01-en.docx, 01-en.pdf, 01-en.txt, 02-en.docx, etc.)

Additionally, prepare a supporting document containing the following information:

- Name and contact information of person(s) who prepared the set

- For document 1: Link to original Wikipedia article and date of download (using “cite this” in the left-hand navigation, then copy the APA style reference)
- For document 2: Link to original document and date of download
- For document 3: Wikipedia article version number and date of download (using “cite this” in the left-hand navigation, then copy the APA style reference)
- For possible additional documents: Their purpose, description, link to all original pieces of work and date of their download.

In total, you should provide us with at least 15 files:

- At least 12 files to be used for testing (at least 4 testing documents x 3 versions)
- Substantiation of authorship and consent for document 2
- Substantiation of authorship and consent for document 4
- Possible substantiation and consent of materials used for additional files
- Supporting document with contact, links to original sources and dates

The TeSToP team is very grateful for your cooperation!