Emoveo

Functional Specification

Jackie Li

Overview

Emoveo is a perl script that supports automated and user controlled redaction of documents for declassification and other legal uses. Emoveo attempts focus on low cost content and syntax analysis so end-users are able to perform high linguistic analysis to control the content and quality of the redacted document.

User Scenarios

Scenario 1: Lawson the Lawyer

Lawson is a top lawyer working on a sensitive case where he needs to protect the identity of his witnesses in the court documents. In order to protect the identities of his witnesses, Lawson prepares each document into a text file. With the identity issue, Lawson knows the names of his witnesses. He enters their name into the field to do single word redactions. The witness names are removed and he prepares the redacted document.

Scenario 2: Wendy the Worker

Wendy is a top-secret clearance government employee who writes secret documents that hold sensitive information about the United States. Every year 69% of her documents are put through mandatory redaction requests by various agencies that request to review the data. Each document that needs to be released have different requirements from heavy redaction to partial redaction to limited redaction. Wendy pushes each document through the perl script. She first does a topic analysis on each paragraph and removes paragraphs with sensitive topics. If her file is too large to go through so many paragraphs, she can automatically remove words that she does not want to appear in the document. Each document is granted 20 keywords from the words in the text and Wendy can remove unwanted keywords until she gets keywords that relate the most to the document. She then is able to remove sentences and phrases containing the keyword. Wendy is worried she might have missed some things with the keywords so she decides to search all proper nouns and phrases and then does some manual searchers through Emoveo. Now, she has the option of automatically removing time elements, country names, and organization names from her redacted document. Before this document can be released, it has to be reviewd so Wendy uses Emoveo to create a separate text file for the review so that they can double check and go over Wendy’s work before release.

Scenario 3: Rita the Reporter

Rita received an anonymous document leak containing sensitive email information about the Asia region. It fits directly with her report on a conspiracy theory she reported on last week in the New York Times. But, being the ethical report that she is, she can't just release the documents. However, being crushed for time and having many small documents to go through, she inputs the documents through Emoveo. It automatically generates 5 key words for each paragraph of the document so that Rita know exactly what the paragraph is talking about. Having an idea of what to look for to redact, Rita effectively has an automatic system that helps her remove large portions of information for release to the public. Additionally, she can continue the automated redacted with time, country, and organization elements.

Non-Goals

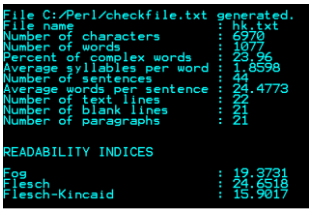
Emoveo does not support non-English texts. While the search and the keyword generation does work with a select amount of other languages, the proper output for the redacted document will not be formatted properly. Additionally, Emoveo depends on a properly formatted plaintext to do a proper analysis. Redaction will not complete properly and consistently if original plaintext format is off.

Workflow

First, the user loads Emoveo in the directory that the program is located and enters the location of the document for input.

```
C:\Perl>perl emoveo8.pl
Make sure that the pdf is saved as a txt file using the save as other function b
uilt into Adobe Reader.
Enter txt file location:        hk.txt
```

Secondly, Emoveo performs a content and structural analysis of the document input to decide whether the user can perform automatic paragraph redaction or is too large and gives the option of doing single word removal. In the case of paragraph redaction, Emoveo presents the user with 5 keywords about the paragraph and gives the user the choice of whether or not to remove that paragraph. In the choice of word redaction, the user enters three words delimited by commas to remove.

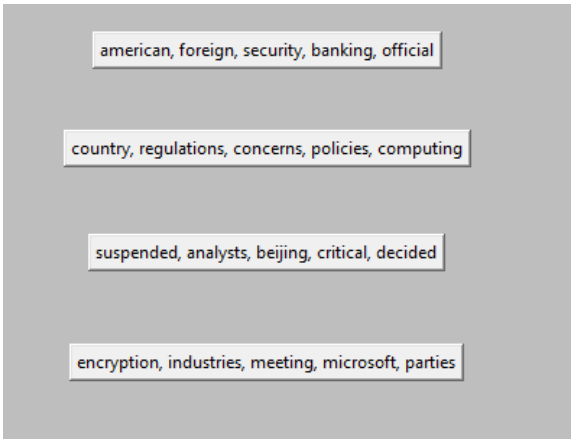


```
according
banking
chinese
companies
country

Paragraph topics listed. Remove paragraph?      no
brewing
chinese
companies
computer
conflict

Paragraph topics listed. Remove paragraph?      yes
```

The user is then presented with 20 keywords that describe the entire document. The user is given the choice to select words to exclude from the list and revise the 20 keywords that are extracted until they get their desired list which comes in the form of a pop-up. In the background, Emoveo begins to build word trees by placing each keyword into a Wikipedia engine and generating keywords from the Wikipedia page to associate with each of the 20 keywords the user selected.



The user has now arrived at the redaction stage. They can redact based on keywords, all noun search, and user input search. With each search, they are presented with a sentence or phrase depending on the size of the document and can decide how much to cut off from the right end to create the phrase to redact.

```
s of dollars of business for major American companies that make the advanced com
puting
Would you like to keep this line?       no

How many chars to remove from the end?  0

of dollars of business for major American companies that make the advanced compu
ting
pment is only a small reprieve for American tech companies. The suspension is te
mporar
Would you like to keep this line?       no

How many chars to remove from the end?  10
```

Last part of the redaction, the user can choose to automatically remove time, country, and organization entities. The user is prompted if they are finish with redaction and can choose to save a plaintext for later use.

```
Would you like Month, days, and weekdays to be removed? yes
Task completed.
Press any key to continue . . .

Would you like the year to be removed?  yes
Task completed.

Would you like to remove specific times such as 24:00?
Press any key to continue . . .

Would you like to be more thorough and remove more specific time indicatiors?
yes
Task completed.

Would you like to remove country names? yes
Task completed.
Press any key to continue . . .

Would you like to do organizational redaction?  yes
Task completed.
Press any key to continue . . .
```

The last analysis that Emoveo gives is the keyword comparisons of the original keyword list that the user created to the keyword list of the redacted document weighted by all the exclusions the user made in previous steps. Then, you get the redacted document.

> ***||||||||| — ||||||||| suspended a policy that would have effectively pushed foreign technology companies out of the country's banking sector, according to a note sent by ||||||||| regulators to banks.***
>
> ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
>
> ***At stake is billions ||||||||| hardware and software that crunches numbers for banks across |||||||||. Trade groups representing companies including Microsoft, ||||||||| and Apple have complained that such policies are protectionist.***