



EDTA, iText and INBATEK Conference

Removing Confidential Data

iText Presentation 26/07/2017



≡ Removing Confidential Data

- What is Redaction
- Case of Paper documents
- Case of PDF documents
- The Process
- pdfSweep Examples
- Extensibility

What is redaction?

- ≡ Process of cleaning your PDF of sensitive data

- ≡ Social security numbers

- ≡ Phone numbers

- ≡ Bank account details

- ≡ Healthcare information

- ≡ Proprietary information

- ≡ Trade secrets

- ≡ “Absolute” removal of sensitive data

- ≡ Recreating redacted content should not be possible

Case of Paper Documents

≡ For paper documents, Redaction meant :

- ≡ Printing a document
- ≡ Blacking out the necessary information
- ≡ Making a photocopy of the document

- Information that gets covered by black ink is actually gone
- Fits the basic idea of “covering up data”

The President said that there is every evidence that our position in Berlin is strongly supported by the people there, and we are committed to that area. Mr. Khrushchev says that we are for a state of war. This is incorrect. It would be well if relations between East Germany and West Germany improved and if the development of US-USSR relations were such as to permit solution of the whole German problem. During his stay in office, Mr. Khrushchev has seen many changes, and changes will go on. But now he wants a peace treaty in six months, an action which would drive us out of Berlin.

Mr. Khrushchev had said that the President was a young man, but, the President continued, he had not assumed office to accept arrangements totally inimical to US interests. The President said he

Case of PDF document

- ≡ Pdf contain instructions for rendering the document
 - ≡ Drawing a black rectangle DOES NOT erase text-rendering instructions
 - ≡ Covering up information no longer works
 - ≡ Actual removal/replacement is needed

- ≡ pdfSweep : What does it do?
 - ≡ Removes data and the metadata
 - ≡ Content does not reflow
 - ≡ Redacted content is (typically) replaced with colored rectangle

Case of PDF document

≡ Redaction Problems

- ≡ Text rendering instructions do not need to appear in logical (reading) order
- ≡ These do not always constitute complete words
- ≡ Images can be defined as a series of drawing operations
- ≡ Images can be added to a pdf document under many formats
- ≡ Presence of meta data : Subset of Fonts embedded *

```
20.1149 T 105.253 o -28.7356 n 17.2414 y 74.7126 8.62069 S 27.2652 o -28.7356 p 28.7356 r -11.8254 a 18.7118 n 17.2414 o 500] TJ
```

```
[a, -28.7356, p, 27.2652, p, 27.2652, e, -27.2652, a, -28.7356, r, 64.6889, a, -28.7356, n, 27.2652, c,
```

```
-38.7594, e, 444] TJ
```

```
/R10 10.44 Tf
```

```
68.16 0.24 Td
```

```
[", 17.1965, P, -18.7118, i, -9.35592, l, -9.35592, o, -17.2414, t, -9.35636, ", 17.1965, , 250] TJ
```

Example input

Tony Soprano

From Wikipedia, the free encyclopedia

Anthony John "Tony" Soprano is a fictional character and the protagonist in the HBO television drama series *The Sopranos*, portrayed by James Gandolfini. The Italian-American character was conceived by *Sopranos* creator and show runner David Chase, who was also largely responsible for the character's story arc throughout the show's six seasons. The character is loosely based on real-life New Jersey mobster Vincent "Vinny Ocean" Palermo (born 1944), a former caporegime (capo) and *de facto* boss of the DeCavalcante crime family of New Jersey. Considered to be the model for the DiMeo crime family, several incidents involving the DeCavalcantes were incorporated into *Sopranos* scripts. Bobby Boriello portrayed Soprano as a child in one episode and Danny Petrillo played the character as a teenager in three episodes.

Tony Soprano



Example output

Tony [REDACTED]

From Wikipedia, the free encyclopedia

Anthony John "Tony" [REDACTED] is a fictional character and the protagonist in the HBO television drama series *The [REDACTED]* portrayed by James Gandolfini. The Italian-American character was conceived by [REDACTED] creator and show runner David Chase, who was also largely responsible for the character's story arc throughout the show's six seasons. The character is loosely based on real-life New Jersey mobster Vincent "Vinny Ocean" Palermo (born 1944), a former caporegime (capo) and *de facto* boss of the DeCavalcante crime family of New Jersey. Considered to be the model for the DiMeo crime family, several incidents involving the DeCavalcantes were incorporated into [REDACTED] scripts. Bobby Boriello portrayed [REDACTED] as a child in one episode and Danny Petrillo played the character as a teenager in three episodes.

Tony [REDACTED]



Redaction methods

☰ Redaction Rectangles

☰ Redaction Annotations

☰ Regular Expressions (AutoSweep)

Redaction Rectangles

- ≡ Specify dimensions of the Rectangle to erase, programmatically
- ≡ Specify page/pages to redact
- ≡ Specify redaction rectangle color

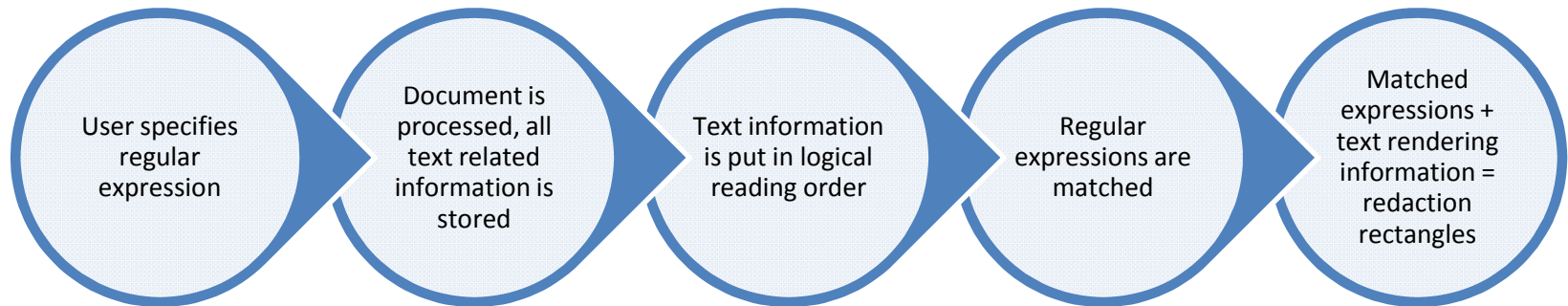
```
1. Rectangle region = new Rectangle(300f, 370f, 215f, 270f);
2. int page = 1;
3. PdfCleanUpLocation loc = new PdfCleanUpLocation(page, region, color);
4. PdfCleanUpTool cleanupTool = new PdfCleanUpTool(pdf);
5. cleanupTool.addCleanupLocation(location);
6. cleanupTool.cleanup();
```

Redaction Annotation

- Specify document, already marked with annotation rectangle

The screenshot shows a PDF viewer interface with a toolbar at the top containing icons for navigation, zoom (100%), and editing. Below the toolbar, three articles are displayed in a grid. The first article, 'Apple Encryption Engineers, if Ordered to Unlock iPhone, Might Resist', has its title highlighted with a red rectangle. The second article, 'With "Smog Jog" Through Beijing, Zuckerberg Stirs Debate on Air Pollution', has a paragraph of text highlighted with a red rectangle. The third article, 'Instagram May Change Your Feed, Personalizing It With an Algorithm', includes an image of two men sitting on a green sofa.

Auto-sweep Process



Text Redaction : AutoSweep

- ≡ Using Regular expressions Based Cleanup Strategy
- ≡ Specify the text pattern to match
- ≡ Specify the color of redaction rectangle

```
CompositeCleanupStrategy strategy = new CompositeCleanupStrategy();  
strategy.add(new RegexBasedCleanupStrategy("Tony( |_)Soprano"));  
strategy.add(new RegexBasedCleanupStrategy("Soprano"));  
strategy.add(new RegexBasedCleanupStrategy("Sopranos"));
```

Extensibility

redactiPhoneUserManualMatchColor.pdf - Adobe Acrobat Pro DC

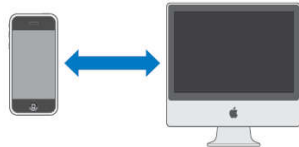
File Edit View Window Help

Home Tools redactiPhoneUser... x

7 / 130 100%

Syncing [redacted] with Your Computer

When you connect [redacted] to your computer, iTunes syncs [redacted] with the information and media on your computer, according to how you've configured the [redacted] sync settings in iTunes.



You can set iTunes to sync any or all of the following:

- Contacts—names, phone numbers, addresses, email addresses, and so on
- Calendars—appointments and events
- Email account settings
- Webpage bookmarks
- Ringtones
- Music and audiobooks
- Photos
- Podcasts
- Videos

Ringtones, music, audiobooks, podcasts, and video content are synced from your iTunes library. If you don't already have content in iTunes, the iTunes Store (available in some countries) makes it easy to purchase or subscribe to content and download it to iTunes. You can also add music to your iTunes library from your CDs. To learn about iTunes and the iTunes Store, open iTunes and choose Help > iTunes Help.

Contacts, calendars, webpage bookmarks, and photos are synced from applications on your computer, as described below. Contacts and calendars are synced both ways between your computer and [redacted]. New entries or changes you make on [redacted] are synced to your computer, and vice versa. Webpage bookmarks are also synced both ways.

Email account settings are only synced from your computer's email application to [redacted]. This allows you to customize your email accounts on [redacted] without affecting email account settings on your computer.

Introduction to pdfCalligraph

- ≡ iText add-on product
- ≡ Commercial License required
- ≡ Correct character Rendering
- ≡ Non Latin language
 - ≡ Thai
 - ≡ Arabic
 - ≡ Hindi etc

Thank You!

- ≡ Log on to <http://itextpdf.com/>
- ≡ Log on to <http://developers.itextpdf.com/>