MET CS688

Web Analytics and Mining

Zlatko Vasilkoski

TEXT MINING EXERCISES

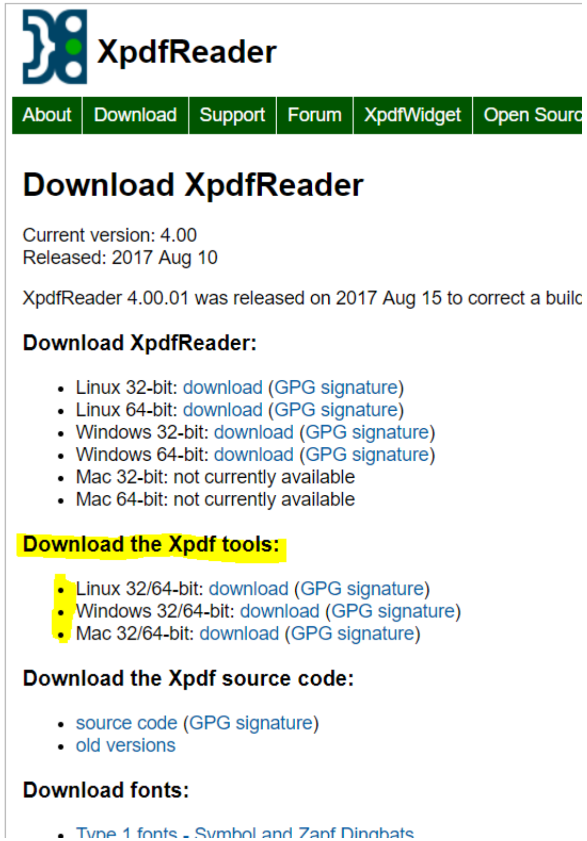
Exercise: Extracting the text content from a pdf file with Xpdf

Implement the following steps:

1. Install Xpdf Tools.
2. Test the installation on your OS by running "pdftotext.exe" in a terminal mode.
3. Run the terminal mode command "pdftotext.exe" from R using R’s *system()* function.
4. Use R to extract content from several pdf files in a folder by running the terminal mode command "pdftotext.exe" in a loop using *lapply()*.

1. Extracting the content from a pdf file

- Installing Xpdf - Extracts the content from a pdf file
- You might find it helpful to use it in Text mining projects on your own.
- Install pdftotext.exe (open-source PDF viewer) part of the Xpdf software suite.
- It can be downloaded from: http://www.xpdfreader.com/download.html
- Choose the download for your operating system.
- The Windows installation is more involved so it is illustrated on the next few slides.



Download XpdfReader

Current version: 4.00
Released: 2017 Aug 10

XpdfReader 4.00.01 was released on 2017 Aug 15 to correct a buil

Download XpdfReader:

- Linux 32-bit: download (GPG signature)
- Linux 64-bit: download (GPG signature)
- Windows 32-bit: download (GPG signature)
- Windows 64-bit: download (GPG signature)
- Mac 32-bit: not currently available
- Mac 64-bit: not currently available

Download the Xpdf tools:

- Linux 32/64-bit: download (GPG signature)
- Windows 32/64-bit: download (GPG signature)
- Mac 32/64-bit: download (GPG signature)

Download the Xpdf source code:

- source code (GPG signature)
- old versions

Download fonts:

- Type 1 fonts - Symbol and Zapf Dingbats

Additional Xpdf Utilities

These are all of the Xpdf utilities:

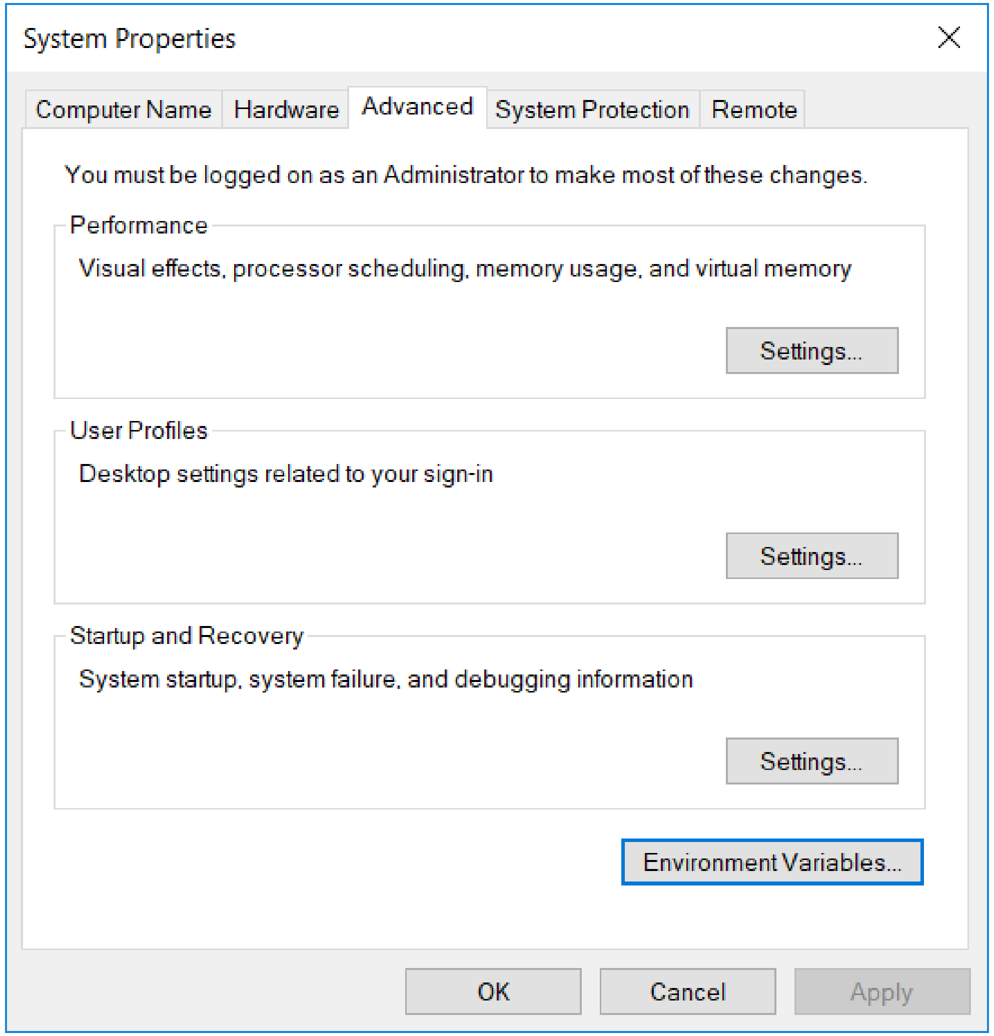
1. pdftotext -- generates a text file from a pdf file
2. pdftops -- generates a PostScript file from a pdf file
3. pdfinfo -- dumps a PDF file's Info dictionary (plus some other useful information)
4. pdffonts -- lists the fonts used in a PDF file along with various information for each font
5. pdfdetach -- lists or extracts embedded files (attachments) from a PDF (archived) file
6. pdftoppm -- converts a PDF file to a series of PPM/PGM/PBM-format bitmaps
7. pdfimages -- extracts the images from a PDF file

Installing Xpdf

- For mac users follow: http://macappstore.org/pdftotext/

- Un-compress the downloaded files in a folder and set the environmental variables and the path to point to the folder where downloaded files are so R can find them.

- On the Windows operating system you would do that by going to **"System"**, choosing **"Advanced system settings"**, and **"Environment Variables"**.



You must be logged on as an Administrator to make most of these changes.

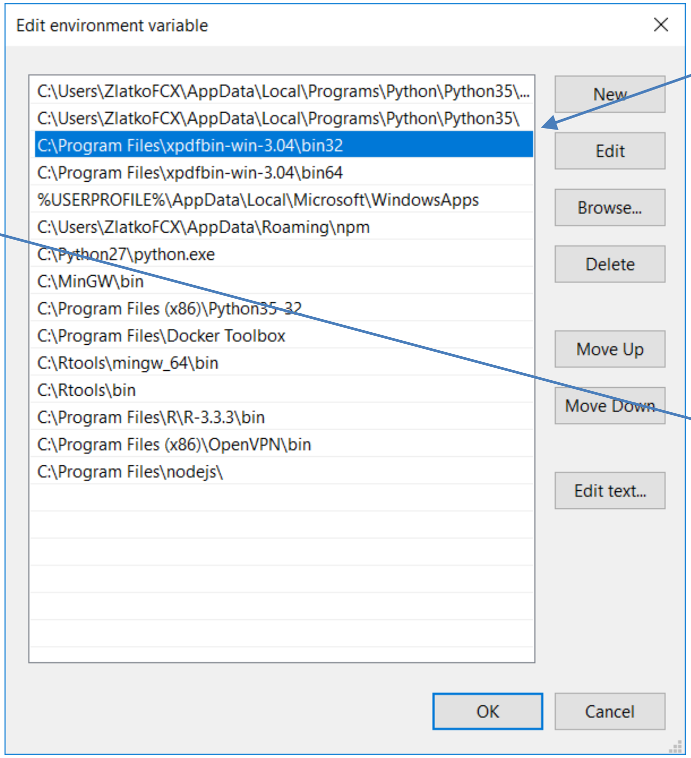
Visual effects, processor scheduling, memory usage, and virtual memory

Desktop settings related to your sign-in

System startup, system failure, and debugging information

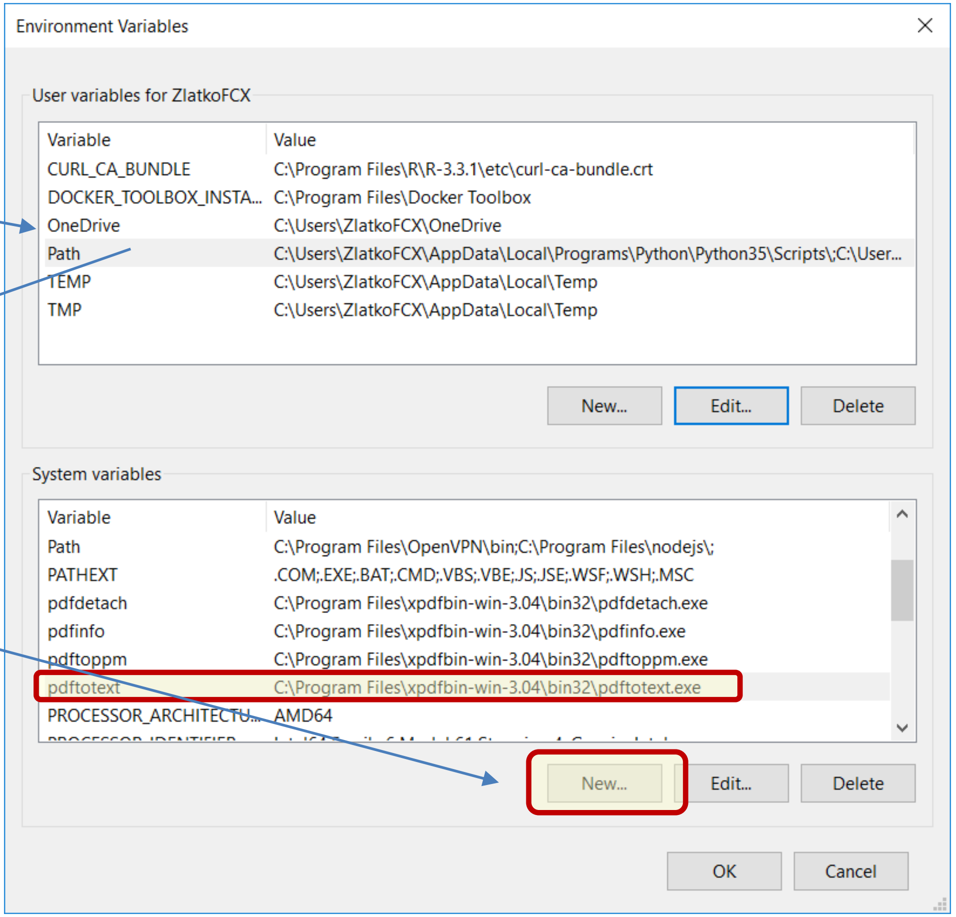
Installing Xpdf

- Add to the “Path” user-variable the **folder** where the downloaded Xpdf files are.
- By clicking “New” add to the System the 2 new environmental variables with “Variable name” "pdfinfo" and "pdftotext".
- For the “Variable value” enter the **path** to the executables "pdfinfo.exe" and "pdftotext.exe".



Edit environment variable

```
C:\Users\ZlatkoFCX\AppData\Local\Programs\Python\Python35\...
C:\Users\ZlatkoFCX\AppData\Local\Programs\Python\Python35\
C:\Program Files\xpdfbin-win-3.04\bin32
C:\Program Files\xpdfbin-win-3.04\bin64
%USERPROFILE%\AppData\Local\Microsoft\WindowsApps
C:\Users\ZlatkoFCX\AppData\Roaming\npm
C:\Python27\python.exe
C:\MinGW\bin
C:\Program Files (x86)\Python35-32
C:\Program Files\Docker Toolbox
C:\Rtools\mingw_64\bin
C:\Rtools\bin
C:\Program Files\R\R-3.3.3\bin
C:\Program Files (x86)\OpenVPN\bin
C:\Program Files\nodejs\
```



Environment Variables

User variables for ZlatkoFCX

Variable	Value
CURL_CA_BUNDLE	C:\Program Files\R\R-3.3.1\etc\curl-ca-bundle.crt
DOCKER_TOOLBOX_INSTA...	C:\Program Files\Docker Toolbox
OneDrive	C:\Users\ZlatkoFCX\OneDrive
Path	C:\Users\ZlatkoFCX\AppData\Local\Programs\Python\Python35\Scripts\;C:\User...
TEMP	C:\Users\ZlatkoFCX\AppData\Local\Temp
TMP	C:\Users\ZlatkoFCX\AppData\Local\Temp

System variables

Variable	Value
Path	C:\Program Files\OpenVPN\bin;C:\Program Files\nodejs\;
PATHEXT	.COM;.EXE;.BAT;.CMD;.VBS;.VBE;.JS;.JSE;.WSF;.WSH;.MSC
pdfdetach	C:\Program Files\xpdfbin-win-3.04\bin32\pdfdetach.exe
pdfinfo	C:\Program Files\xpdfbin-win-3.04\bin32\pdfinfo.exe
pdftoppm	C:\Program Files\xpdfbin-win-3.04\bin32\pdftoppm.exe
pdftotext	C:\Program Files\xpdfbin-win-3.04\bin32\pdftotext.exe
PROCESSOR_ARCHITECTU...	AMD64

Install Xpdf on Mac

- Download the Mac source code to your computer (Desktop, etc.).
- Uncompressed source code will expand into a folder.
- Open up a Terminal and follow the directions in the “INSTALL” text file (shown below):
  - To install this binary package:
  - 1. Copy the executables (xpdf, pdftotext, etc.) to to /usr/local/bin.
  - 2. Copy the man pages (*.1 and *.5) to /usr/local/man/man1 and /usr/local/man/man5.
  - 3. Copy the sample-xpdfrc file to /usr/local/etc/xpdfrc. You’ll probably want to edit its contents (as distributed, everything is commented out) -- see xpdfrc(5) for details.
- To test the installation:
- In the Terminal using “cd /location of file”, navigate to the directory where the PDF file is and then type:
  - pdftotext -layout pdfname.pdf
- Depending on the size of the PDF file, your output text file (with the same name as the original) will be in the same folder in a matter of seconds.

2. Test the "pdftotext.exe" installation

- To test the installation, convert a pdf file in a specific folder to a text file.

- The "pdftotext.exe" conversion from pdf to text in a terminal mode is implemented at as:

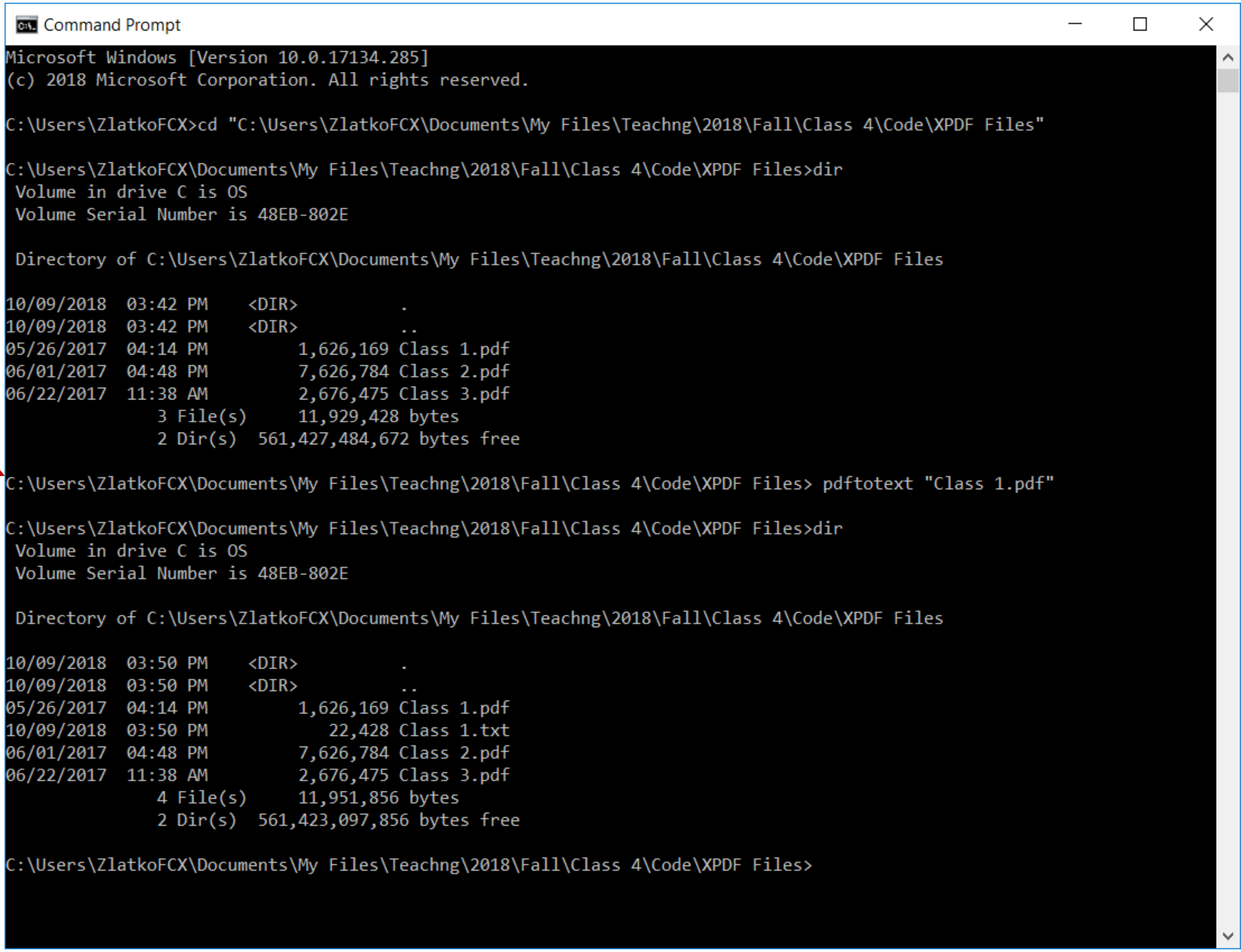
```
> pdftotext "file.pdf"
```

- Note:
  - This usage produces a text file with the same name as the input file
  - The text file is created in the same directory as the PDFs.

Test the "pdftotext.exe" installation

- Open a terminal window (cmd on Windows).
- Either navigate to the directory where the PDF file is or include the path to it in the filename.
- List all the files in the folder using “dir”
- Type: pdftotext "Class 1.pdf"
- This produces a text file, with the same name as the pdf file, created in the same directory as the PDFs.
- List all the files in the folder again using “dir” to see it.

- Note that the file name (and the path) needs to be enclosed in quotation.
- In R we can use the function *system()* to implement a cmd command as you would implement in a terminal window.



```
Microsoft Windows [Version 10.0.17134.285]
(c) 2018 Microsoft Corporation. All rights reserved.

C:\Users\ZlatkoFCX>cd "C:\Users\ZlatkoFCX\Documents\My Files\Teachng\2018\Fall\Class 4\Code\XPDF Files"

C:\Users\ZlatkoFCX\Documents\My Files\Teachng\2018\Fall\Class 4\Code\XPDF Files>dir
 Volume in drive C is OS
 Volume Serial Number is 48EB-802E

 Directory of C:\Users\ZlatkoFCX\Documents\My Files\Teachng\2018\Fall\Class 4\Code\XPDF Files

10/09/2018  03:42 PM    <DIR>          .
10/09/2018  03:42 PM    <DIR>          ..
05/26/2017  04:14 PM         1,626,169 Class 1.pdf
06/01/2017  04:48 PM         7,626,784 Class 2.pdf
06/22/2017  11:38 AM         2,676,475 Class 3.pdf
               3 File(s)     11,929,428 bytes
               2 Dir(s)  561,427,484,672 bytes free

C:\Users\ZlatkoFCX\Documents\My Files\Teachng\2018\Fall\Class 4\Code\XPDF Files> pdftotext "Class 1.pdf"

C:\Users\ZlatkoFCX\Documents\My Files\Teachng\2018\Fall\Class 4\Code\XPDF Files>dir
 Volume in drive C is OS
 Volume Serial Number is 48EB-802E

 Directory of C:\Users\ZlatkoFCX\Documents\My Files\Teachng\2018\Fall\Class 4\Code\XPDF Files

10/09/2018  03:50 PM    <DIR>          .
10/09/2018  03:50 PM    <DIR>          ..
05/26/2017  04:14 PM         1,626,169 Class 1.pdf
10/09/2018  03:50 PM            22,428 Class 1.txt
06/01/2017  04:48 PM         7,626,784 Class 2.pdf
06/22/2017  11:38 AM         2,676,475 Class 3.pdf
               4 File(s)     11,951,856 bytes
               2 Dir(s)  561,423,097,856 bytes free

C:\Users\ZlatkoFCX\Documents\My Files\Teachng\2018\Fall\Class 4\Code\XPDF Files>
```

3. Try implementing these R code examples

Download several pdf class notes from Blackboard and place them in a folder “PDF Files”, then

Task1: Use R to get the text content by implementing terminal mode command such as

```
> pdftotext "file.pdf"
```

- To accomplish this from an R script you can use R’s function *system()*
- To insert the "" around the pdf filename you need to escape them with '' . To merge, use *paste0()*. Here is the R code example:

```
system(paste(Sys.which("pdftotext"), paste0('"', myPDFfiles[1], '"')), wait=FALSE)
```

Task2: How to extend this to multiple files in a folder?

- A wildcards (*), for example *pdftotext "*pdf"*, for converting multiple files, cannot be used because pdftotext expects only one file name.

- Using R’s *lapply()* several pdf files that are contained in a single folder (specified by the R object “myPDFfiles”) can be converted with an in line function such as this,

```
> lapply(myPDFfiles, function(i) system(paste(Sys.which("pdftotext") , paste0('"', i, '"')), wait = FALSE))
```

- Note how each PDF file converted into a text file is indexed by *"i"*
- Note: Quotes (“”) in R are tricky. Make sure you properly “escape” them with single quotes such as in paste0('"', i, '"')) . **Copy/pasting the above code in R may not work. You need to type it!**

4. Getting several pdf files from a folder

Task: How to get path to **several** pdf files that are contained in a **single** folder.

```
# Example 1: Convert to text single pdf files that is contained in a single folder.
exe.loc <- Sys.which("pdftotext")  # location of "pdftotext.exe"
pdf.loc=file.path(getwd(),"PDF Files")  # folder "PDF Files" with PDFs

# Get the path (character vector) of PDF file names
myPDFfiles <- normalizePath(list.files(path = pdf.loc, pattern = "pdf",  full.names = TRUE))

# Convert single pdf file to text by placing "" around the character vector of PDF file name
system(paste(exe.loc, paste0('"', myPDFfiles[1], '"')), wait=FALSE)
```

Note:

- *Sys.which* gives the path to *pdftotext.exe* (the one you set it up during installation).
- *file.path(getwd(),"PDF Files")* Gets the current folder (*getwd()*) and forms a path.
- *normalizePath()* Converts the file paths to a canonical form for the operating system.
- *list.files()* Lists the files in a Directory/Folder.
- *system()* Invoke a system command.
- myPDFfiles[1] Access the first element of vector “myPDFfiles” (contains path to the PDF file).

Using the “pdftools” package

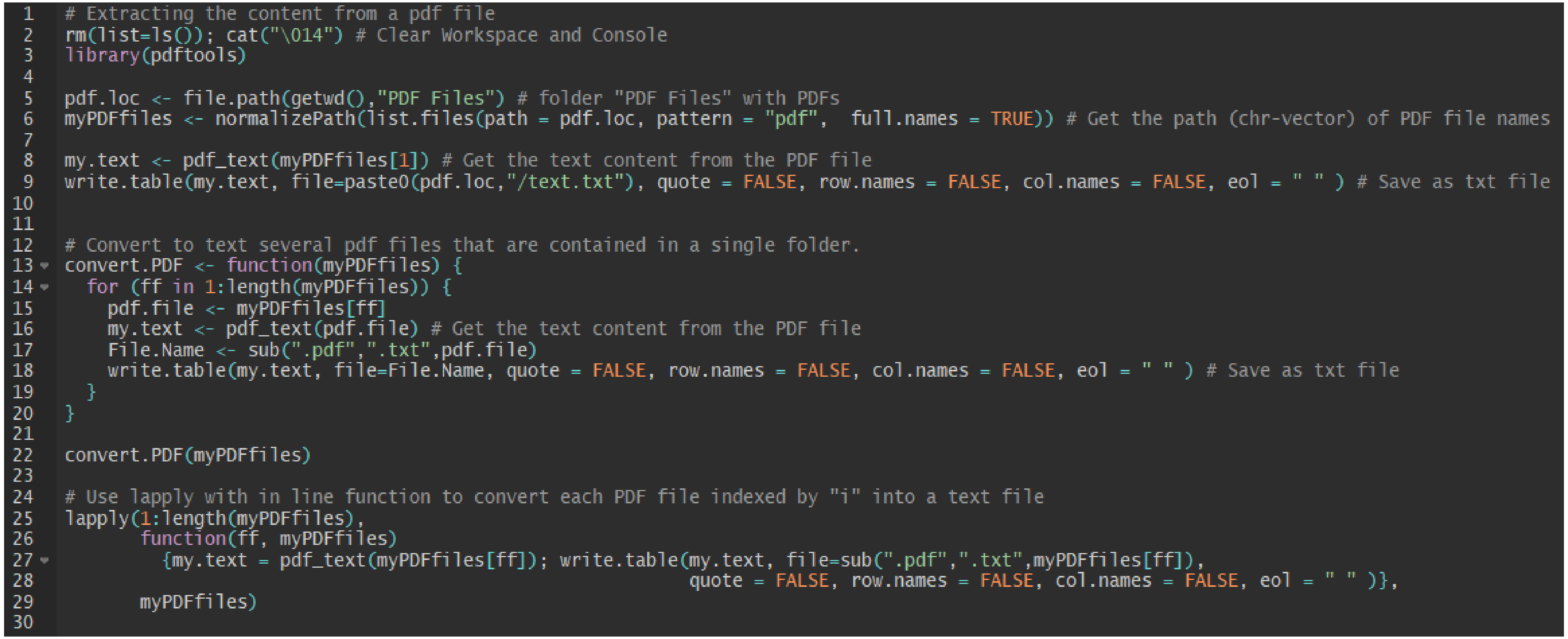
- Another way is to use the package “pdftools”.
- To get the text content from a PDF file use

  > my.text <- pdf_text(myPDFfiles[1])

- To save text content as txt file use

  > write.table(my.text, file=paste0(pdf.loc,"/text.txt"), quote = FALSE, row.names = FALSE, col.names = FALSE, eol = " ")

- To convert and save several pdf files you can create a function as illustrated below.

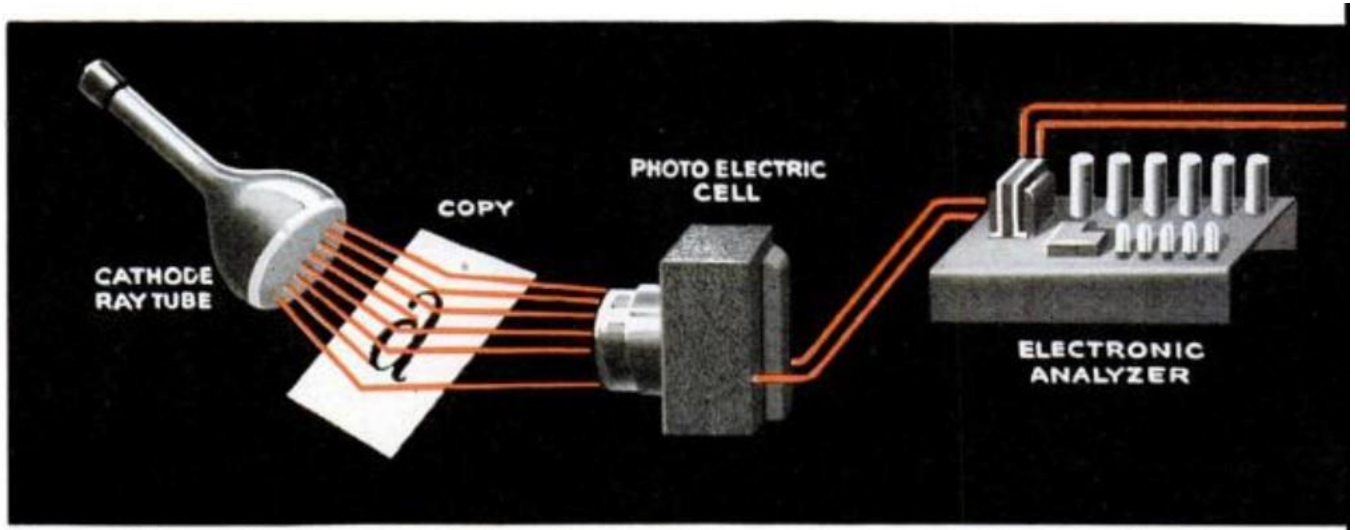


```r
# Extracting the content from a pdf file
rm(list=ls()); cat("\014") # Clear Workspace and Console
library(pdftools)

pdf.loc <- file.path(getwd(),"PDF Files") # folder "PDF Files" with PDFs
myPDFfiles <- normalizePath(list.files(path = pdf.loc, pattern = "pdf",  full.names = TRUE)) # Get the path (chr-vector) of PDF file names

my.text <- pdf_text(myPDFfiles[1]) # Get the text content from the PDF file
write.table(my.text, file=paste0(pdf.loc,"/text.txt"), quote = FALSE, row.names = FALSE, col.names = FALSE, eol = " " ) # Save as txt file


# Convert to text several pdf files that are contained in a single folder.
convert.PDF <- function(myPDFfiles) {
  for (ff in 1:length(myPDFfiles)) {
    pdf.file <- myPDFfiles[ff]
    my.text <- pdf_text(pdf.file) # Get the text content from the PDF file
    File.Name <- sub(".pdf",".txt",pdf.file)
    write.table(my.text, file=File.Name, quote = FALSE, row.names = FALSE, col.names = FALSE, eol = " " ) # Save as txt file
  }
}

convert.PDF(myPDFfiles)

# Use lapply with in line function to convert each PDF file indexed by "i" into a text file
lapply(1:length(myPDFfiles),
       function(ff, myPDFfiles)
         {my.text = pdf_text(myPDFfiles[ff]); write.table(my.text, file=sub(".pdf",".txt",myPDFfiles[ff]),
                                                          quote = FALSE, row.names = FALSE, col.names = FALSE, eol = " " )},
       myPDFfiles)

```

History of OCR

Google Book Search initiative – a servicer that searches the full text of books and magazines that Google has scanned, converted to text using optical character recognition (OCR), and stored in its digital database.

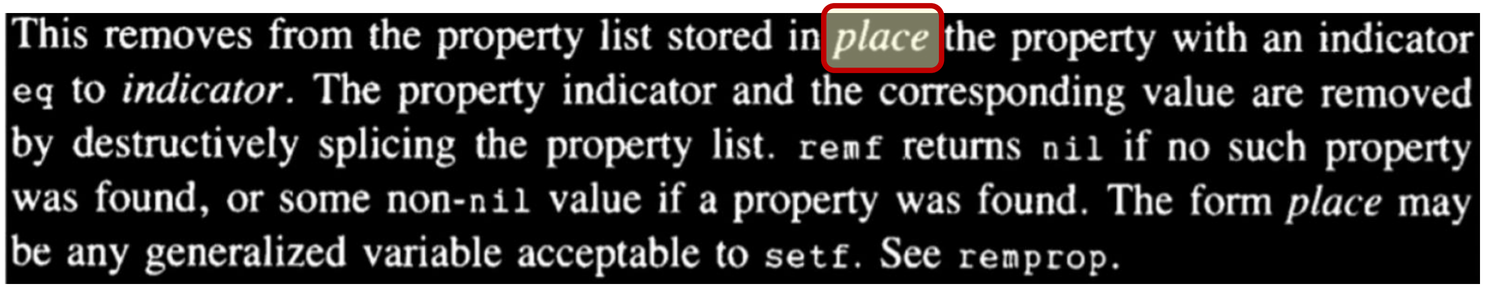
- It has opened up many avenues for future research in document understanding and recognition.
- Resulted in developing Google's Tesseract software.



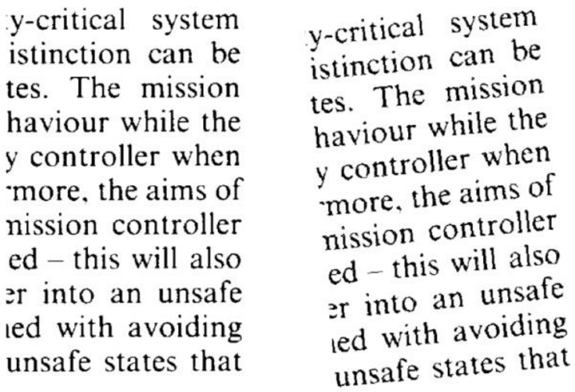
How machine reads: Electric eye looks at letter and reports its shape to electronic "brain."

An image of RCA's 1949 OCR system: M. Martin, "Reading Machine Speaks Out Loud," Popular Science, vol. 154, no. 2, Feb 1949, pp. 125-127. Used under fair use, 2014. The system was discontinued prior to completion due to its high costs .

Types of problems OCR encounters

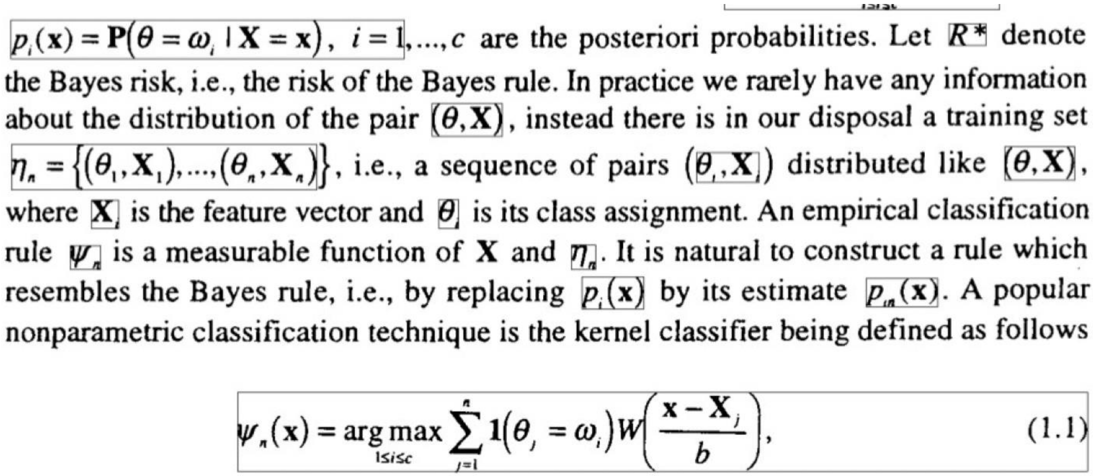


This removes from the property list stored in *place* the property with an indicator eq to *indicator*. The property indicator and the corresponding value are removed by destructively splicing the property list. `remf` returns `nil` if no such property was found, or some non-`nil` value if a property was found. The form *place* may be any generalized variable acceptable to `setf`. See `remprop`.



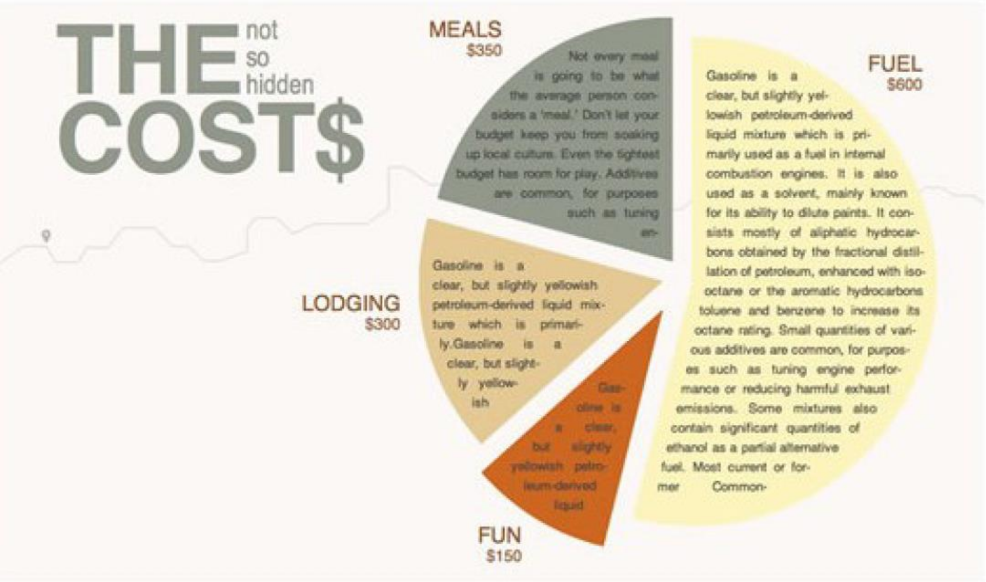
y-critical system
istinction can be
tes. The mission
haviour while the
y controller when
-more, the aims of
nission controller
ed – this will also
er into an unsafe
ied with avoiding
unsafe states that

y-critical system
istinction can be
tes. The mission
haviour while the
y controller when
-more, the aims of
nission controller
ed – this will also
er into an unsafe
ied with avoiding
unsafe states that



$p_i(\mathbf{x}) = \mathbf{P}(\theta = \omega_i \mid \mathbf{X} = \mathbf{x})$, $i = 1, \ldots, c$ are the posteriori probabilities. Let R^* denote the Bayes risk, i.e., the risk of the Bayes rule. In practice we rarely have any information about the distribution of the pair $(\theta, \mathbf{X})$, instead there is in our disposal a training set $\eta_n = \{(\theta_1, \mathbf{X}_1), \ldots, (\theta_n, \mathbf{X}_n)\}$, i.e., a sequence of pairs $(\theta_i, \mathbf{X}_i)$ distributed like $(\theta, \mathbf{X})$, where $\mathbf{X}_i$ is the feature vector and θ_i is its class assignment. An empirical classification rule ψ_n is a measurable function of $\mathbf{X}$ and η_n. It is natural to construct a rule which resembles the Bayes rule, i.e., by replacing $p_i(\mathbf{x})$ by its estimate $p_{in}(\mathbf{x})$. A popular nonparametric classification technique is the kernel classifier being defined as follows

$$\psi_n(\mathbf{x}) = \arg\max_{1 \le i \le c} \sum_{j=1}^{n} \mathbf{1}(\theta_j = \omega_i) W\left(\frac{\mathbf{x} - \mathbf{X}_j}{b}\right), \qquad (1.1)$$



Not every meal is going to be what the average person considers a 'meal.' Don't let your budget keep you from soaking up local culture. Even the tightest budget has room for play. Additives are common, for purposes such as tuning en-

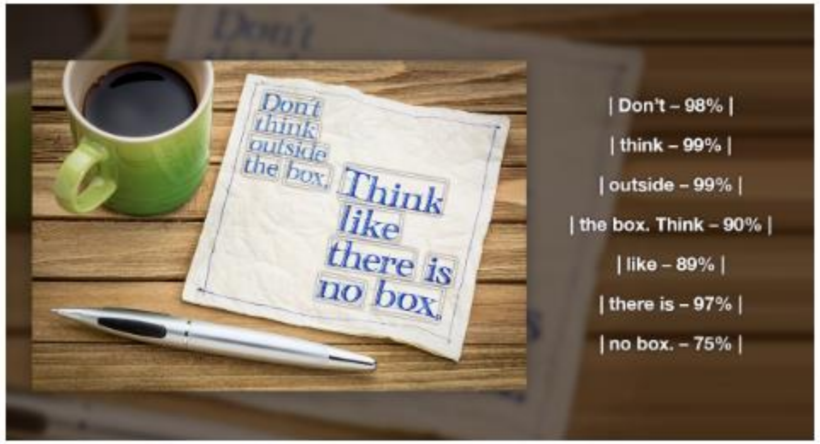
Gasoline is a clear, but slightly yellowish petroleum-derived liquid mixture which is primarily used as a fuel in internal combustion engines. It is also used as a solvent, mainly known for its ability to dilute paints. It consists mostly of aliphatic hydrocarbons obtained by the fractional distillation of petroleum, enhanced with iso-octane or the aromatic hydrocarbons toluene and benzene to increase its octane rating. Small quantities of various additives are common, for purposes such as tuning engine performance or reducing harmful exhaust emissions. Some mixtures also contain significant quantities of ethanol as a partial alternative fuel. Most current or former Common-

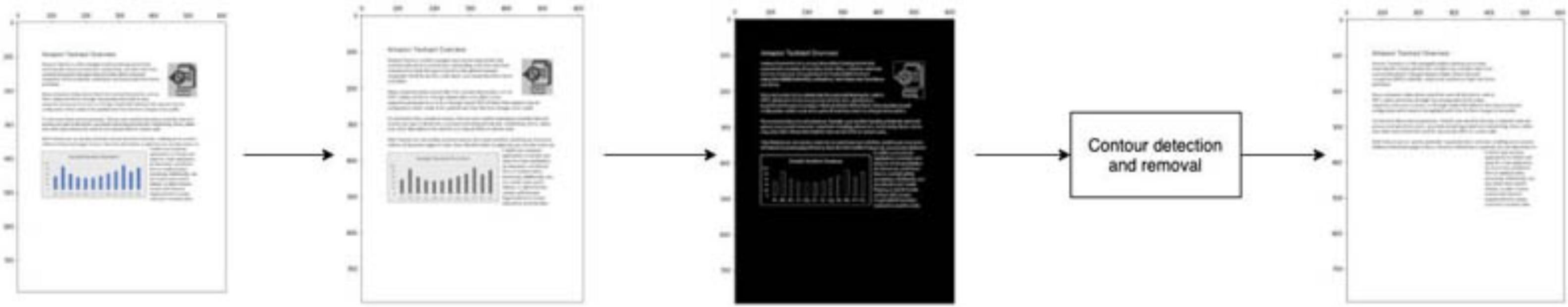
Gasoline is a clear, but slightly yellowish petroleum-derived liquid mixture which is primarily.Gasoline is a clear, but slightly yellowish

Gasoline is a clear, but slightly yellowish petroleum-derived liquid

Amazon Textract

Textract is a machine learning service that automatically extracts text, handwriting and data from scanned documents that goes beyond simple optical character recognition (OCR) to identify, understand, and extract data from forms and tables.




Contour detection and removal

Exercise: OCR with Tesseract

OCR (pattern recognition in general) is a very difficult problem for computers.

The R tesseract package provides R bindings to Google's OCR library Tesseract.

- It is a powerful optical character recognition (OCR) engine that supports over 100 languages. The engine is highly configurable in order to tune the detection algorithms and obtain the best possible results.

- Results are rarely perfect and the accuracy rapidly decreases with the quality of the input image. But if you can enhance your input images to reasonable quality, Tesseract can often help to extract most of the text from the image.
  - One such image enhancements used with together with tesseract is the “magick” package.

Tesseract

Using Tesseract, the following file types can be created:

- **alto** — Output in ALTO format (*OUTPUTBASE*.xml).
- **hocr** — Output in hOCR format (*OUTPUTBASE*.hocr).
- **pdf** — Output PDF (*OUTPUTBASE*.pdf).
- **tsv** — Output TSV (*OUTPUTBASE*.tsv).
- **txt** — Output plain text (*OUTPUTBASE*.txt).
- **makebox** — Write box file (*OUTPUTBASE*.box).
- **get.images** — Write processed input images to file (tessinput.tif).
- **lstm.train** — Output files used by LSTM training (*OUTPUTBASE*.lstmf).

The general syntax for converting a *eng* - English language *png.file* into a text file (***txt***) called *output.file* is:

```
system( paste("tesseract", png.file, output.file, ' --oem 1 -l eng txt'), wait=TRUE)
```

Install Tesseract

Search online where to find a link to install Tesseract for your OS

- One such site is: https://digi.bib.uni-mannheim.de/tesseract/

Once you download and run tesseract executable, on Windows OS you will need to add Tesseract path (most likely “C:\Program Files\Tesseract-OCR”) to your path in the system’s environment variable.

- Add “C:\Program Files\Tesseract-OCR”

To verify if the installation worked open command prompt in “C:\Program Files\Tesseract-OCR” and type

```
tesseract --version
```

If you see any error like

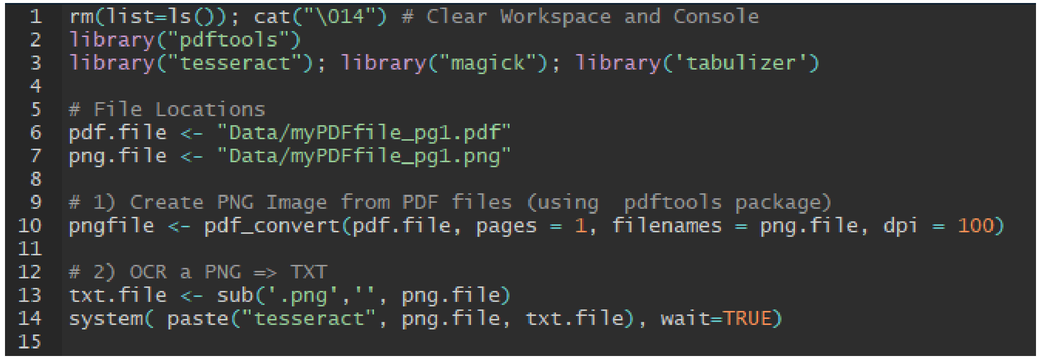
```
tesseract command not found
```

most probably you have made some mistake in the installation.

Example 1: OCR Image to Extract Text

OCR a PNG File to extract the text and save it as text file.

1. Start with a pdf that we are going to turn into a png file (image)
2. OCR the PNG File to extract the text and save it into a file.



```r
rm(list=ls()); cat("\014") # Clear Workspace and Console
library("pdftools")
library("tesseract"); library("magick"); library('tabulizer')

# File Locations
pdf.file <- "Data/myPDFfile_pg1.pdf"
png.file <- "Data/myPDFfile_pg1.png"

# 1) Create PNG Image from PDF files (using  pdftools package)
pngfile <- pdf_convert(pdf.file, pages = 1, filenames = png.file, dpi = 100)

# 2) OCR a PNG => TXT
txt.file <- sub('.png','', png.file)
system( paste("tesseract", png.file, txt.file), wait=TRUE)

```

PDF File

SCIENCE ADVANCES | RESEARCH ARTICLE

NETWORK SCIENCE

Optimal network topology for responsive collective behavior

David Mateo[1]*, Nikolaj Horsevad[1], Vahid Hassani[1], Mohammadreza Chamanbaz[1,2], Roland Bouffanais[1]

Animals, humans, and multi-robot systems operate in dynamic environments, where the ability to respond to changing circumstances is paramount. An effective collective response requires suitable information transfer among agents and thus critically depends on the interaction network. To investigate the influence of the network topology on collective response, we consider an archetypal model of distributed decision-making and study the capacity of the system to follow a driving signal for varying topologies and system sizes. Experiments with a swarm of robots reveal a nontrivial relationship between frequency of the driving signal and optimal network topology. The emergent collective response to slow-changing perturbations increases with the degree of the interaction network, but the opposite is true for the response to fast-changing ones. These results have far-reaching implications for the design and understanding of distributed systems: a dynamic rewiring of the interaction network is essential to effective collective operations at different time scales.

INTRODUCTION

A wide range of complex systems are characterized by relatively simple dynamical rules while still producing excessively complex emergent collective behaviors. Examples abound in the natural world [e.g., a flock of birds, a school of fish, a swarm of insects (*1–9*)], in social systems [e.g., social networks (*10–12*)], and in engineered multi-agent systems [e.g., self-organized networks of mobile sensors, multi-vehicle coordination, and swarm robotics systems (*13–16*)].

Historically, particular attention has been directed toward investigating varieties of collective behaviors obtained by testing a wide range of local agent-to-agent interaction rules (*6, 9*). Collective behaviors have also been investigated from the network-theoretic perspective (*4, 8, 17–21*). It is now clear that such rich collective behaviors are the outcome of a complex interplay between network topology—characteristic of the group-level organization—and the dynamical laws at the agent's level (*4, 8, 20–22*).

Many collective behaviors can be studied through the lens of distributed consensus problems, including collective motion in animal groups and multi-robot systems. Over the past decade, the number of studies on decentralized consensus and cooperation in networked multi-agent systems has experienced a spectacular growth, with concomitant developments in various fields of engineering and science (*2, 3, 23–26*). Consensus dynamics is the cornerstone of cooperative control strategies for vehicular formation (*13, 16, 23*), swarm robotics (*14, 15*), and synchronization of coupled oscillators (*23, 27*). Decentralized consensus is also at the core of collective opinion dynamics and complex contagion processes in social networks (*10–12*), as well as complex collective responses in biological swarms (*3–8*).

Previous studies focused on establishing the influence of the interaction network topology on (i) the capacity of the collective to reach consensus in the presence of noise, communication constraints, and time delays (*21, 23*); (ii) the speed of consensus (i.e., its convergence rate) (*18, 25, 28*); (iii) the stability and stabilization of consensus (*23*); and (iv) the ability to steer the system toward a particular consensus value by means of various control techniques such as pinning control, cooperative tracking control, or model reference consensus (*19, 20*).

However, the effects of the network topology on other dynamical properties of distributed multi-agent systems such as their adaptivity or responsiveness to external perturbations have received considerably less attention (*4*).

It is important to emphasize that a capacity for fast consensus is not necessarily indicative of a responsive collective behavior. For instance, ferromagnets at low temperature exhibit a global spontaneous magnetization—a process that can be described by a distributed consensus protocol. It is known that both the degree of consensus (i.e., magnetization) and the speed at which it is reached increase with decreasing temperature, but the capacity of the system to respond to external perturbations is maximized at a finite critical temperature.

Similarly, in the context of animal collective motion, it has been observed that midges exhibit low levels of ordering while maintaining large connected correlations, thus having a high collective response (*5*). With these observations, the authors eloquently argued that one must be careful in relating collective order (i.e., degree of consensus) with the collective responsiveness. The collective response of the animal group was obtained experimentally by measuring the correlations in the fluctuations of their behavior. While inferring a collective response to external perturbations from these fluctuations is not formally justified for out-of-equilibrium systems, extensive numerical studies (*29*) have shown that this equivalence holds in the context of collective motion based on distributed heading consensus. Moreover, simulations have shown that this measure of susceptibility is a good indicator of the group's performance in biologically relevant functions such as predator avoidance (*8*). These facts along with other empirical evidence have led to the conclusion that responsiveness, rather than high consensus or order, is the true hallmark of collective behavior (*3*).

To study how the responsiveness of a collective is affected by its interaction network topology, we consider an elementary example of distributed decision-making: a linear time-invariant (LTI) system of agents performing consensus over a scalar state variable. The agents—i.e., the nodes of the interaction network—are all identical, except for one "leader" [also known as "stubborn" agent in some contexts (*12, 25*)] with some arbitrary predefined dynamics. From the control-theoretic perspective, this leader introduces a time-varying control input signal into the system. In the biological context, this dynamical leader represents a member of a swarm with access to privileged information about a food

[1]Singapore University of Technology and Design, 8 Somapah Road, Singapore 487372, Singapore. [2]Arak University of Technology, Daneshgah Road, Arak, Iran.
*Corresponding author. Email: david.mateo.valderrama@gmail.com

Mateo *et al*., *Sci. Adv.* 2019; **5**:eaau0999   3 April 2019

1 of 10

PNG File

SCIENCE ADVANCES | RESEARCH ARTICLE

NETWORK SCIENCE

Optimal network topology for responsive collective behavior

David Mateo[1]*, Nikolaj Horsevad[1], Vahid Hassani[1], Mohammadreza Chamanbaz[1,2], Roland Bouffanais[1]

Copyright © 2019 The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works. Distributed under a Creative Commons Attribution NonCommercial License 4.0 (CC BY-NC).

Animals, humans, and multi-robot systems operate in dynamic environments, where the ability to respond to changing circumstances is paramount. An effective collective response requires suitable information transfer among agents and thus critically depends on the interaction network. To investigate the influence of the network topology on collective response, we consider an archetypal model of distributed decision-making and study the capacity of the system to follow a driving signal for varying topologies and system sizes. Experiments with a swarm of robots reveal a nontrivial relationship between frequency of the driving signal and optimal network topology. The emergent collective response to slow-changing perturbations increases with the degree of the interaction network, but the opposite is true for the response to fast-changing ones. These results have far-reaching implications for the design and understanding of distributed systems: a dynamic rewiring of the interaction network is essential to effective collective operations at different time scales.

INTRODUCTION

A wide range of complex systems are characterized by relatively simple dynamical rules while still producing excessively complex emergent collective behaviors. Examples abound in the natural world [e.g., a flock of birds, a school of fish, a swarm of insects (1–9)], in social systems [e.g., social networks (10–12)], and in engineered multi-agent systems [e.g., self-organized networks of mobile sensors, multi-vehicle coordination, and swarm robotics systems (13–16)].

Historically, particular attention has been directed toward investigating varieties of collective behaviors obtained by testing a wide range of local agent-to-agent interaction rules (6, 9). Collective behaviors have also been investigated from the network-theoretic perspective (4, 8, 17–21). It is now clear that such rich collective behaviors are the outcome of a complex interplay between network topology—characteristic of the group-level organization—and the dynamical laws at the agent's level (4, 8, 20–22).

Many collective behaviors can be studied through the lens of distributed consensus problems, including collective motion in animal groups and multi-robot systems. Over the past decade, the number of studies on decentralized consensus and cooperation in networked multi-agent systems has experienced a spectacular growth, with concomitant developments in various fields of engineering and science (2, 3, 23–26). Consensus dynamics is the cornerstone of cooperative control strategies for vehicular formation (13, 16, 23), swarm robotics (14, 15), and synchronization of coupled oscillators (23, 27). Decentralized consensus is also at the core of collective opinion dynamics and complex contagion processes in social networks (10–12), as well as complex collective responses in biological swarms (3–8).

Previous studies focused on establishing the influence of the interaction network topology on (i) the capacity of the collective to reach consensus in the presence of noise, communication constraints, and time delays (21, 23); (ii) the speed of consensus (i.e., its convergence rate) (18, 25, 28); (iii) the stability and stabilization of consensus (23); and (iv) the ability to steer the system toward a particular consensus value by means of various control techniques such as pinning control, cooperative tracking control, or model reference consensus (19, 20).

However, the effects of the network topology on other dynamical properties of distributed multi-agent systems such as their adaptivity or responsiveness to external perturbations have received considerably less attention (4).

It is important to emphasize that a capacity for fast consensus is not necessarily indicative of a responsive collective behavior. For instance, ferromagnets at low temperature exhibit a global spontaneous magnetization—a process that can be described by a distributed consensus protocol. It is known that both the degree of consensus (i.e., magnetization) and the speed at which it is reached increase with decreasing temperature, but the capacity of the system to respond to external perturbations is maximized at a finite critical temperature.

Similarly, in the context of animal collective motion, it has been observed that midges exhibit low levels of ordering while maintaining large connected correlations, thus having a high collective response (5). With these observations, the authors eloquently argued that one must be careful in relating collective order (i.e., degree of consensus) with the collective responsiveness. The collective response of the animal group was obtained experimentally by measuring the correlations in the fluctuations of their behavior. While inferring a collective response to external perturbations from these fluctuations is not formally justified for out-of-equilibrium systems, extensive numerical studies (29) have shown that this equivalence holds in the context of collective motion based on distributed heading consensus. Moreover, simulations have shown that this measure of susceptibility is a good indicator of the group's performance in biologically relevant functions such as predator avoidance (8). These facts along with other empirical evidence have led to the conclusion that responsiveness, rather than high consensus or order, is the true hallmark of collective behavior (3).

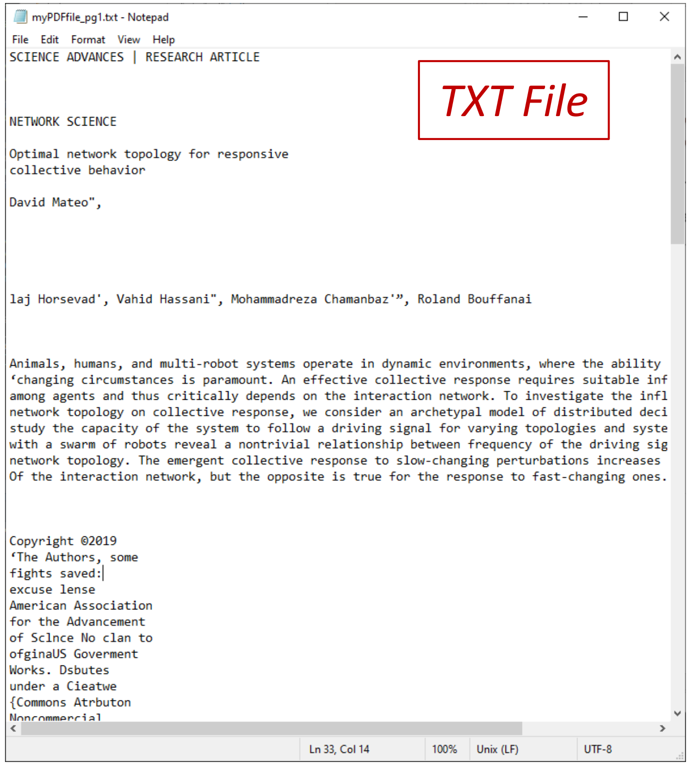
To study how the responsiveness of a collective is affected by its interaction network topology, we consider an elementary example of distributed decision-making: a linear time-invariant (LTI) system of agents performing consensus over a scalar state variable. The agents—i.e., the nodes of the interaction network—are all identical, except for one "leader" [also known as "stubborn" agent in some contexts (12, 25)] with some arbitrary predefined dynamics. From the control-theoretic perspective, this leader introduces a time-varying control input signal into the system. In the biological context, this dynamical leader represents a member of a swarm with access to privileged information about a food

[1]Singapore University of Technology and Design, 8 Somapah Road, Singapore 487372, Singapore. [2]Arak University of Technology, Daneshgah Road, Arak, Iran.
*Corresponding author. Email: david.mateo.valderrama@gmail.com

Mateo et al., Sci. Adv. 2019; 5:eaau0999   3 April 2019

1 of 10

Downloaded from http://advances.sciencemag.org/ on April 4, 2019

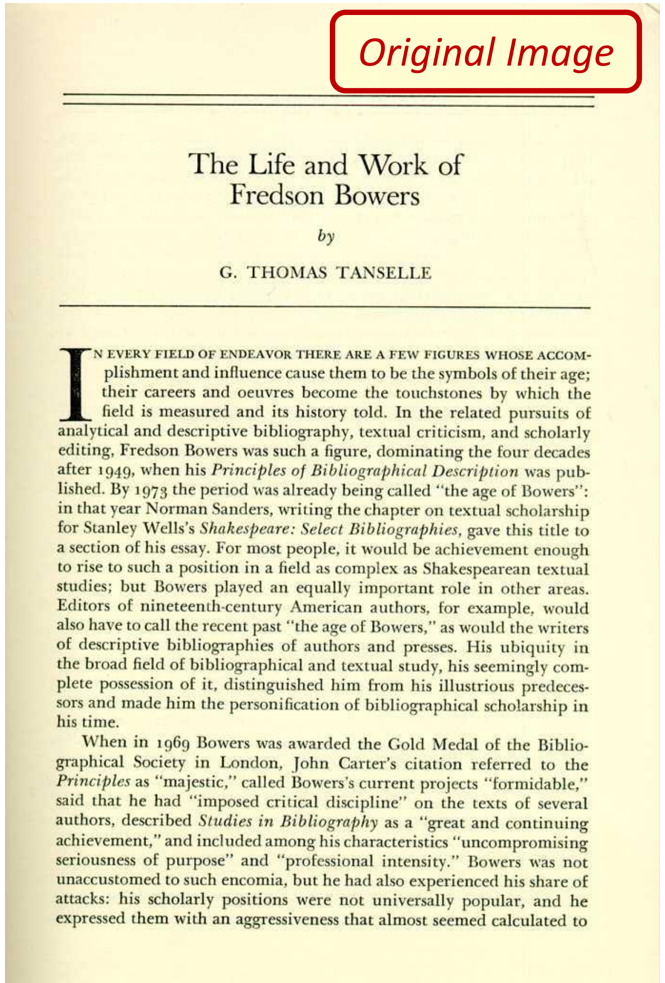


TXT File

```
SCIENCE ADVANCES | RESEARCH ARTICLE



NETWORK SCIENCE

Optimal network topology for responsive
collective behavior

David Mateo",






laj Horsevad', Vahid Hassani", Mohammadreza Chamanbaz'”, Roland Bouffanai




Animals, humans, and multi-robot systems operate in dynamic environments, where the ability
‘changing circumstances is paramount. An effective collective response requires suitable inf
among agents and thus critically depends on the interaction network. To investigate the infl
network topology on collective response, we consider an archetypal model of distributed deci
study the capacity of the system to follow a driving signal for varying topologies and syste
with a swarm of robots reveal a nontrivial relationship between frequency of the driving sig
network topology. The emergent collective response to slow-changing perturbations increases
Of the interaction network, but the opposite is true for the response to fast-changing ones.




Copyright ©2019
‘The Authors, some
fights saved:
excuse lense
American Association
for the Advancement
of Sclnce No clan to
ofginaUS Goverment
Works. Dsbutes
under a Cieatwe
{Commons Atrbuton
Noncommercial
```

Example 2: Tesseract and Magick

- Image cleaning and pre-processing before OCR is always advised
  - It improves the quality of the OCR output into text
- Cleaning and pre-processing steps typically involve cropping out the text area, rescaling, increasing contrast, etc.
- The "magick" package is an excellent tool for this task, as illustrated by the following code.



Original Image

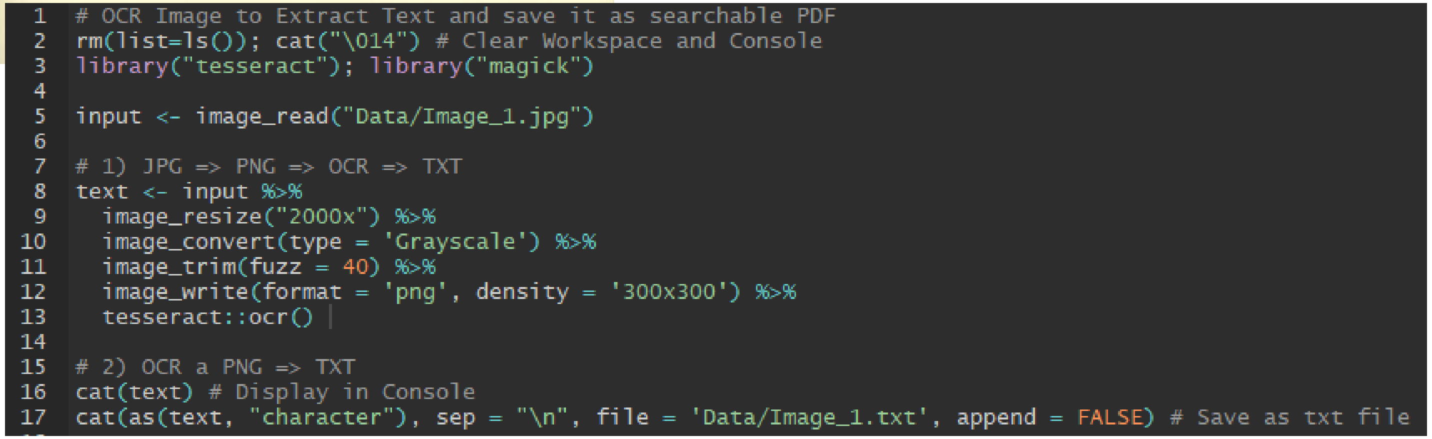
The Life and Work of
Fredson Bowers

by

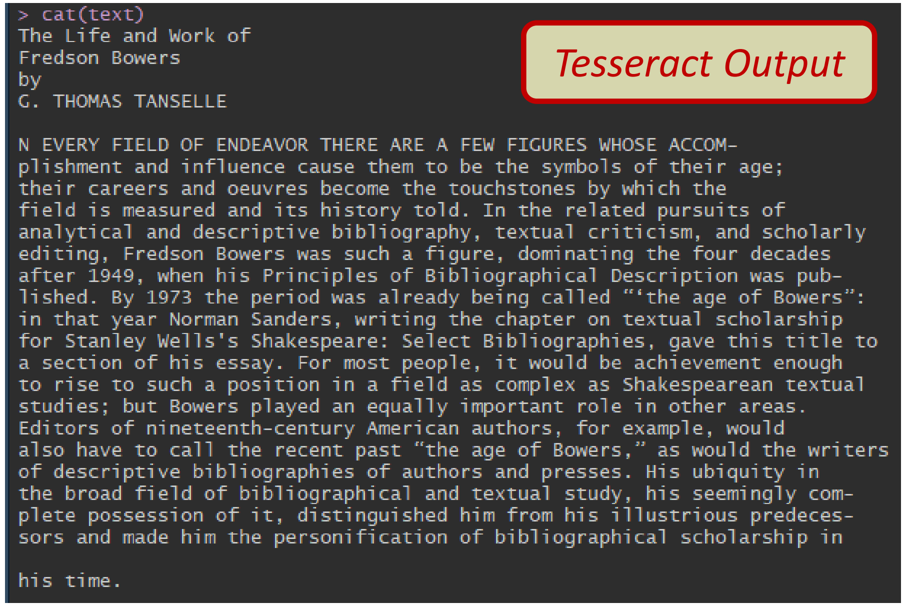
G. THOMAS TANSELLE

IN EVERY FIELD OF ENDEAVOR THERE ARE A FEW FIGURES WHOSE ACCOMplishment and influence cause them to be the symbols of their age; their careers and oeuvres become the touchstones by which the field is measured and its history told. In the related pursuits of analytical and descriptive bibliography, textual criticism, and scholarly editing, Fredson Bowers was such a figure, dominating the four decades after 1949, when his *Principles of Bibliographical Description* was published. By 1973 the period was already being called "the age of Bowers": in that year Norman Sanders, writing the chapter on textual scholarship for Stanley Wells's *Shakespeare: Select Bibliographies*, gave this title to a section of his essay. For most people, it would be achievement enough to rise to such a position in a field as complex as Shakespearean textual studies; but Bowers played an equally important role in other areas. Editors of nineteenth-century American authors, for example, would also have to call the recent past "the age of Bowers," as would the writers of descriptive bibliographies of authors and presses. His ubiquity in the broad field of bibliographical and textual study, his seemingly complete possession of it, distinguished him from his illustrious predecessors and made him the personification of bibliographical scholarship in his time.

When in 1969 Bowers was awarded the Gold Medal of the Bibliographical Society in London, John Carter's citation referred to the *Principles* as "majestic," called Bowers's current projects "formidable," said that he had "imposed critical discipline" on the texts of several authors, described *Studies in Bibliography* as a "great and continuing achievement," and included among his characteristics "uncompromising seriousness of purpose" and "professional intensity." Bowers was not unaccustomed to such encomia, but he had also experienced his share of attacks: his scholarly positions were not universally popular, and he expressed them with an aggressiveness that almost seemed calculated to



```r
# OCR Image to Extract Text and save it as searchable PDF
rm(list=ls()); cat("\014") # Clear Workspace and Console
library("tesseract"); library("magick")

input <- image_read("Data/Image_1.jpg")

# 1) JPG => PNG => OCR => TXT
text <- input %>%
  image_resize("2000x") %>%
  image_convert(type = 'Grayscale') %>%
  image_trim(fuzz = 40) %>%
  image_write(format = 'png', density = '300x300') %>%
  tesseract::ocr()

# 2) OCR a PNG => TXT
cat(text) # Display in Console
cat(as(text, "character"), sep = "\n", file = 'Data/Image_1.txt', append = FALSE) # Save as txt file
```

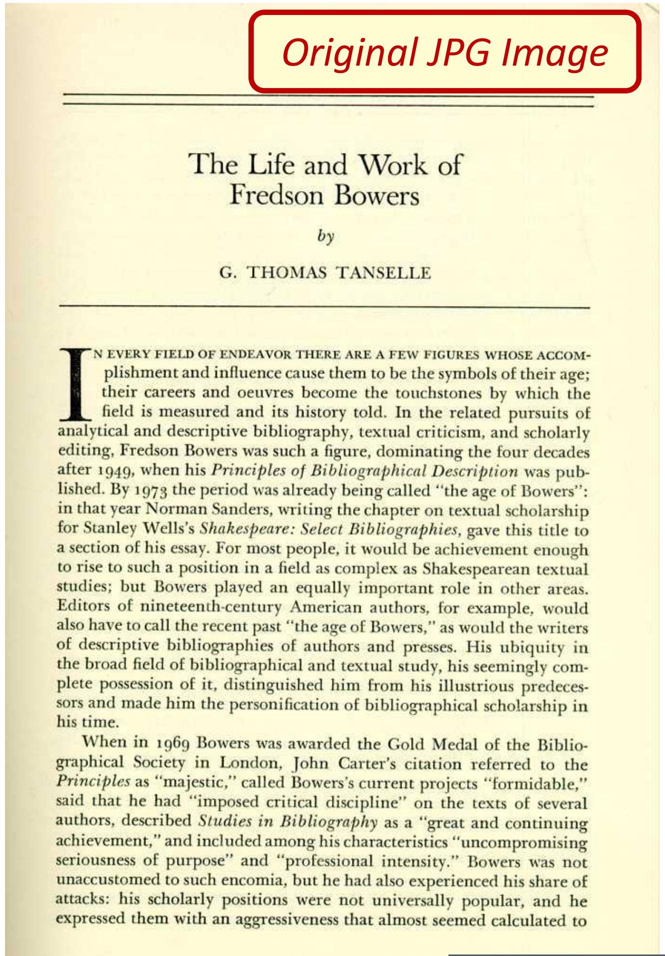


Tesseract Output

```
> cat(text)
The Life and Work of
Fredson Bowers
by
G. THOMAS TANSELLE

N EVERY FIELD OF ENDEAVOR THERE ARE A FEW FIGURES WHOSE ACCOM-
plishment and influence cause them to be the symbols of their age;
their careers and oeuvres become the touchstones by which the
field is measured and its history told. In the related pursuits of
analytical and descriptive bibliography, textual criticism, and scholarly
editing, Fredson Bowers was such a figure, dominating the four decades
after 1949, when his Principles of Bibliographical Description was pub-
lished. By 1973 the period was already being called "‘the age of Bowers”:
in that year Norman Sanders, writing the chapter on textual scholarship
for Stanley Wells's Shakespeare: Select Bibliographies, gave this title to
a section of his essay. For most people, it would be achievement enough
to rise to such a position in a field as complex as Shakespearean textual
studies; but Bowers played an equally important role in other areas.
Editors of nineteenth-century American authors, for example, would
also have to call the recent past “the age of Bowers,” as would the writers
of descriptive bibliographies of authors and presses. His ubiquity in
the broad field of bibliographical and textual study, his seemingly com-
plete possession of it, distinguished him from his illustrious predeces-
sors and made him the personification of bibliographical scholarship in

his time.
```

Example 3: Searchable PDF from JPG

- Image cleaning and pre-processing before OCR is always advised
  - It improves the quality of the OCR output into text
- Cleaning and pre-processing steps typically involve cropping out the text area, rescaling, increasing contrast, etc.
- The “magick” package is an excellent tool for this task, as illustrated by the following code.



Original JPG Image

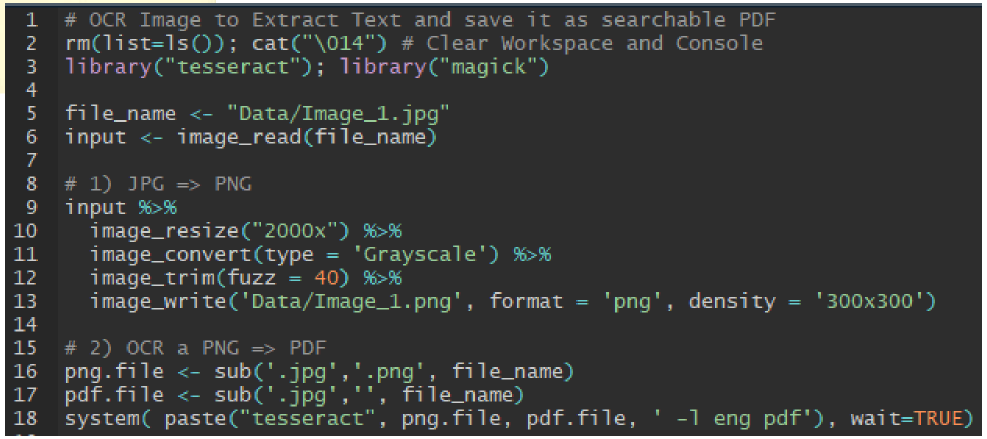
The Life and Work of
Fredson Bowers

by

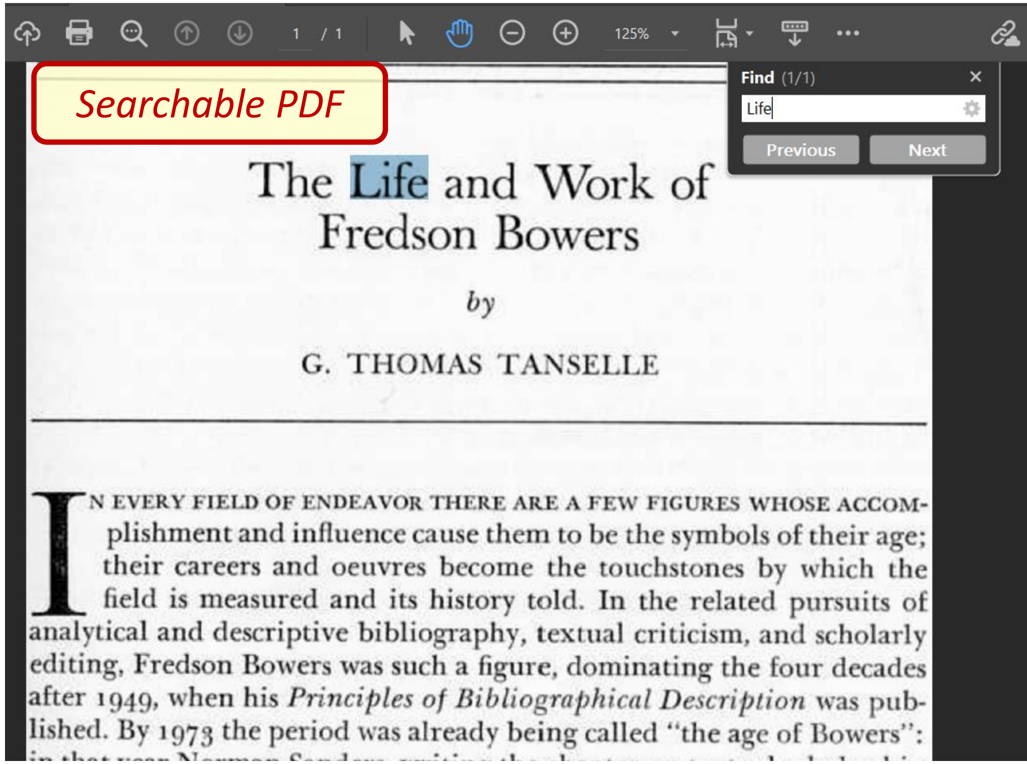
G. THOMAS TANSELLE

IN EVERY FIELD OF ENDEAVOR THERE ARE A FEW FIGURES WHOSE ACCOMplishment and influence cause them to be the symbols of their age; their careers and oeuvres become the touchstones by which the field is measured and its history told. In the related pursuits of analytical and descriptive bibliography, textual criticism, and scholarly editing, Fredson Bowers was such a figure, dominating the four decades after 1949, when his *Principles of Bibliographical Description* was published. By 1973 the period was already being called “the age of Bowers”: in that year Norman Sanders, writing the chapter on textual scholarship for Stanley Wells's *Shakespeare: Select Bibliographies*, gave this title to a section of his essay. For most people, it would be achievement enough to rise to such a position in a field as complex as Shakespearean textual studies; but Bowers played an equally important role in other areas. Editors of nineteenth-century American authors, for example, would also have to call the recent past “the age of Bowers,” as would the writers of descriptive bibliographies of authors and presses. His ubiquity in the broad field of bibliographical and textual study, his seemingly complete possession of it, distinguished him from his illustrious predecessors and made him the personification of bibliographical scholarship in his time.

When in 1969 Bowers was awarded the Gold Medal of the Bibliographical Society in London, John Carter's citation referred to the *Principles* as “majestic,” called Bowers's current projects “formidable,” said that he had “imposed critical discipline” on the texts of several authors, described *Studies in Bibliography* as a “great and continuing achievement,” and included among his characteristics “uncompromising seriousness of purpose” and “professional intensity.” Bowers was not unaccustomed to such encomia, but he had also experienced his share of attacks: his scholarly positions were not universally popular, and he expressed them with an aggressiveness that almost seemed calculated to



```r
# OCR Image to Extract Text and save it as searchable PDF
rm(list=ls()); cat("\014") # Clear Workspace and Console
library("tesseract"); library("magick")

file_name <- "Data/Image_1.jpg"
input <- image_read(file_name)

# 1) JPG => PNG
input %>%
  image_resize("2000x") %>%
  image_convert(type = 'Grayscale') %>%
  image_trim(fuzz = 40) %>%
  image_write('Data/Image_1.png', format = 'png', density = '300x300')

# 2) OCR a PNG => PDF
png.file <- sub('.jpg','.png', file_name)
pdf.file <- sub('.jpg','', file_name)
system( paste("tesseract", png.file, pdf.file, ' -l eng pdf'), wait=TRUE)
```



Searchable PDF

The Life and Work of
Fredson Bowers

by

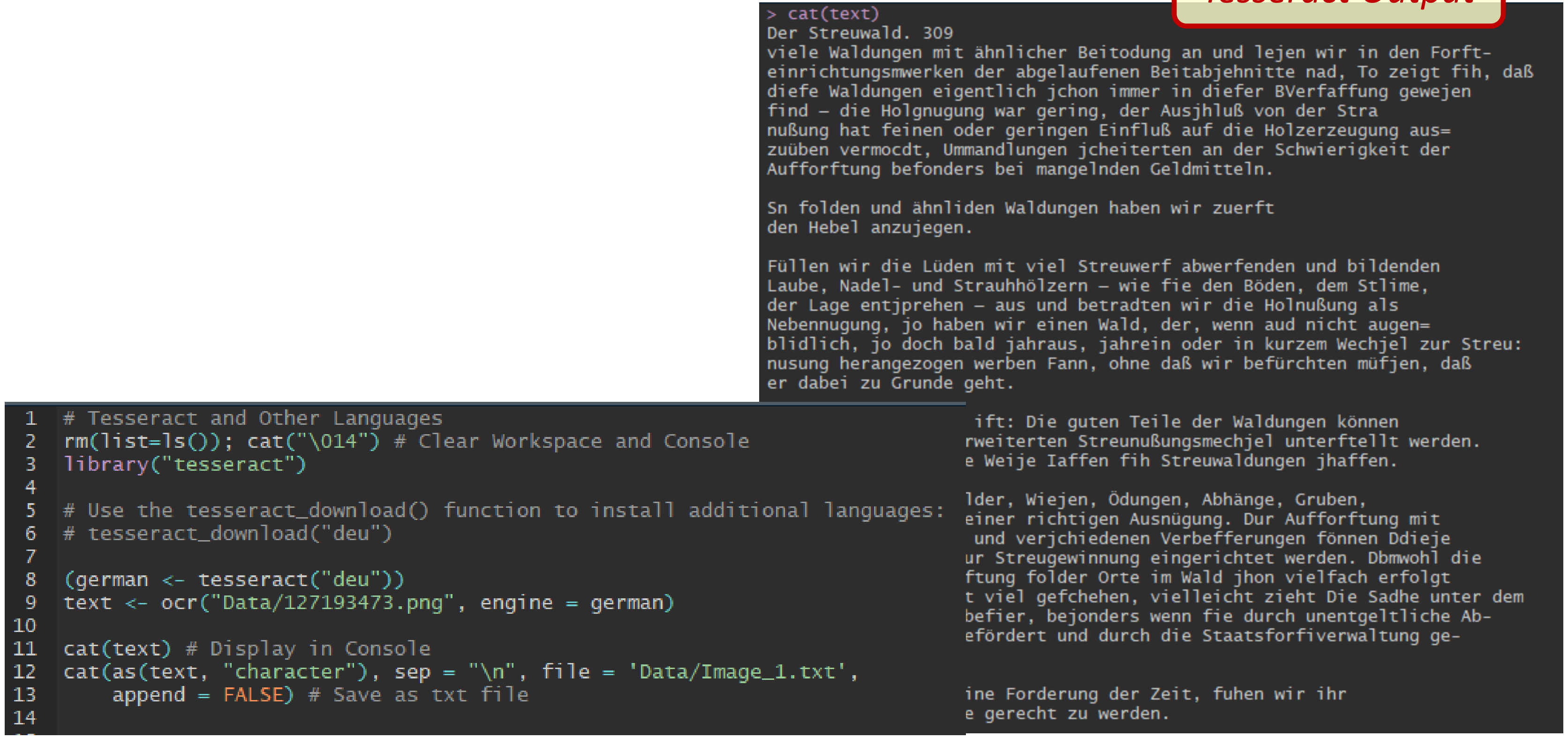
G. THOMAS TANSELLE

IN EVERY FIELD OF ENDEAVOR THERE ARE A FEW FIGURES WHOSE ACCOMplishment and influence cause them to be the symbols of their age; their careers and oeuvres become the touchstones by which the field is measured and its history told. In the related pursuits of analytical and descriptive bibliography, textual criticism, and scholarly editing, Fredson Bowers was such a figure, dominating the four decades after 1949, when his *Principles of Bibliographical Description* was published. By 1973 the period was already being called “the age of Bowers”:

Tesseract and Other Languages

- To better identify OCR-ed words, Tesseract has capability for installing and using vocabularies for additional languages. Use the tesseract_download() function to install additional languages:
  - tesseract_download("deu")

  To OCR German text use:
  - (german <- tesseract("deu"))
  - text <- ocr("Data/127193473.png", engine = german)

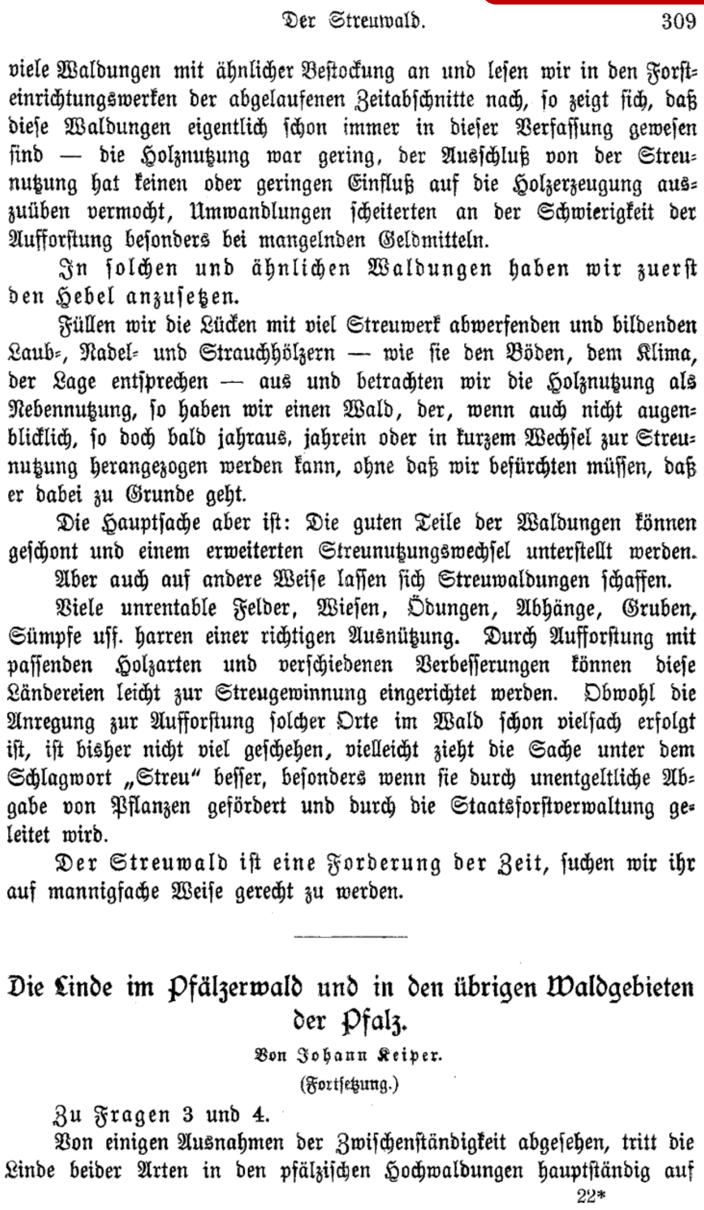


Tesseract Output

```
> cat(text)
Der Streuwald. 309
viele Waldungen mit ähnlicher Beitodung an und lejen wir in den Forft-
einrichtungsmwerken der abgelaufenen Beitabjehnitte nad, To zeigt fih, daß
diefe Waldungen eigentlich jchon immer in diefer BVerfaffung gewejen
find – die Holgnugung war gering, der Ausjhluß von der Stra
nußung hat feinen oder geringen Einfluß auf die Holzerzeugung aus=
zuüben vermocdt, Ummandlungen jcheiterten an der Schwierigkeit der
Aufforftung befonders bei mangelnden Geldmitteln.

Sn folden und ähnliden Waldungen haben wir zuerft
den Hebel anzujegen.

Füllen wir die Lüden mit viel Streuwerf abwerfenden und bildenden
Laube, Nadel- und Strauhhölzern – wie fie den Böden, dem Stlime,
der Lage entjprehen – aus und betradten wir die Holnußung als
Nebennugung, jo haben wir einen Wald, der, wenn aud nicht augen=
blidlich, jo doch bald jahraus, jahrein oder in kurzem Wechjel zur Streu:
nusung herangezogen werben Fann, ohne daß wir befürchten müfjen, daß
er dabei zu Grunde geht.

                           ift: Die guten Teile der Waldungen können
                          rweiterten Streunußungsmechjel unterftellt werden.
                          e Weije Iaffen fih Streuwaldungen jhaffen.

                          lder, Wiejen, Ödungen, Abhänge, Gruben,
                          einer richtigen Ausnügung. Dur Aufforftung mit
                           und verjchiedenen Verbefferungen fönnen Ddieje
                          ur Streugewinnung eingerichtet werden. Dbmwohl die
                          ftung folder Orte im Wald jhon vielfach erfolgt
                          t viel gefchehen, vielleicht zieht Die Sadhe unter dem
                          befier, bejonders wenn fie durch unentgeltliche Ab-
                          efördert und durch die Staatsforfiverwaltung ge-


                          ine Forderung der Zeit, fuhen wir ihr
                          e gerecht zu werden.
```

```r
# Tesseract and Other Languages
rm(list=ls()); cat("\014") # Clear Workspace and Console
library("tesseract")

# Use the tesseract_download() function to install additional languages:
# tesseract_download("deu")

(german <- tesseract("deu"))
text <- ocr("Data/127193473.png", engine = german)

cat(text) # Display in Console
cat(as(text, "character"), sep = "\n", file = 'Data/Image_1.txt',
    append = FALSE) # Save as txt file

```

Original Image



Der Streuwald. 309

viele Waldungen mit ähnlicher Beſtockung an und leſen wir in den Forſteinrichtungswerken der abgelaufenen Zeitabſchnitte nach, ſo zeigt ſich, daß dieſe Waldungen eigentlich ſchon immer in dieſer Verfaſſung geweſen ſind — die Holznutzung war gering, der Ausſchluß von der Streunutzung hat keinen oder geringen Einfluß auf die Holzerzeugung auszuüben vermocht, Umwandlungen ſcheiterten an der Schwierigkeit der Aufforſtung beſonders bei mangelnden Geldmitteln.

In ſolchen und ähnlichen Waldungen haben wir zuerſt den Hebel anzuſetzen.

Füllen wir die Lücken mit viel Streuwerk abwerfenden und bildenden Laub-, Nadel- und Strauchhölzern — wie ſie den Böden, dem Klima, der Lage entſprechen — aus und betrachten wir die Holznutzung als Nebennutzung, ſo haben wir einen Wald, der, wenn auch nicht augenblicklich, ſo doch bald jahraus, jahrein oder in kurzem Wechſel zur Streunutzung herangezogen werden kann, ohne daß wir befürchten müſſen, daß er dabei zu Grunde geht.

Die Hauptſache aber iſt: Die guten Teile der Waldungen können geſchont und einem erweiterten Streunutzungswechſel unterſtellt werden.

Aber auch auf andere Weiſe laſſen ſich Streuwaldungen ſchaffen.

Viele unrentable Felder, Wieſen, Ödungen, Abhänge, Gruben, Sümpfe uſſ. harren einer richtigen Ausnützung. Durch Aufforſtung mit paſſenden Holzarten und verſchiedenen Verbeſſerungen können dieſe Ländereien leicht zur Streugewinnung eingerichtet werden. Obwohl die Anregung zur Aufforſtung ſolcher Orte im Wald ſchon vielfach erfolgt iſt, iſt bisher nicht viel geſchehen, vielleicht zieht die Sache unter dem Schlagwort „Streu“ beſſer, beſonders wenn ſie durch unentgeltliche Abgabe von Pflanzen gefördert und durch die Staatsforſtverwaltung geleitet wird.

Der Streuwald iſt eine Forderung der Zeit, ſuchen wir ihr auf mannigfache Weiſe gerecht zu werden.

Die Linde im Pfälzerwald und in den übrigen Waldgebieten der Pfalz.

Von Johann Keiper.

(Fortſetzung.)

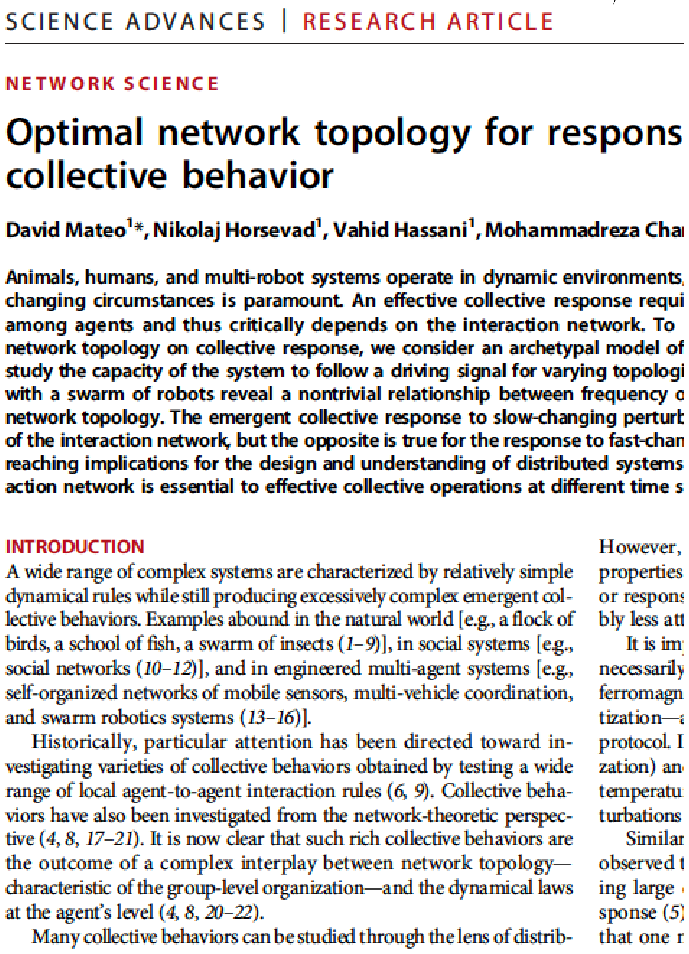
Zu Fragen 3 und 4.

Von einigen Ausnahmen der Zwiſchenſtändigkeit abgeſehen, tritt die Linde beider Arten in den pfälziſchen Hochwaldungen hauptſtändig auf

22*

Example 4: OCR a PNG and create HTML file that looks the same

Original PNG Image



SCIENCE ADVANCES | RESEARCH ARTICLE

NETWORK SCIENCE

Optimal network topology for respons
collective behavior

David Mateo[1]*, Nikolaj Horsevad[1], Vahid Hassani[1], Mohammadreza Cha

Animals, humans, and multi-robot systems operate in dynamic environments,
changing circumstances is paramount. An effective collective response requi
among agents and thus critically depends on the interaction network. To
network topology on collective response, we consider an archetypal model of
study the capacity of the system to follow a driving signal for varying topologi
with a swarm of robots reveal a nontrivial relationship between frequency o
network topology. The emergent collective response to slow-changing perturb
of the interaction network, but the opposite is true for the response to fast-chan
reaching implications for the design and understanding of distributed systems
action network is essential to effective collective operations at different time s

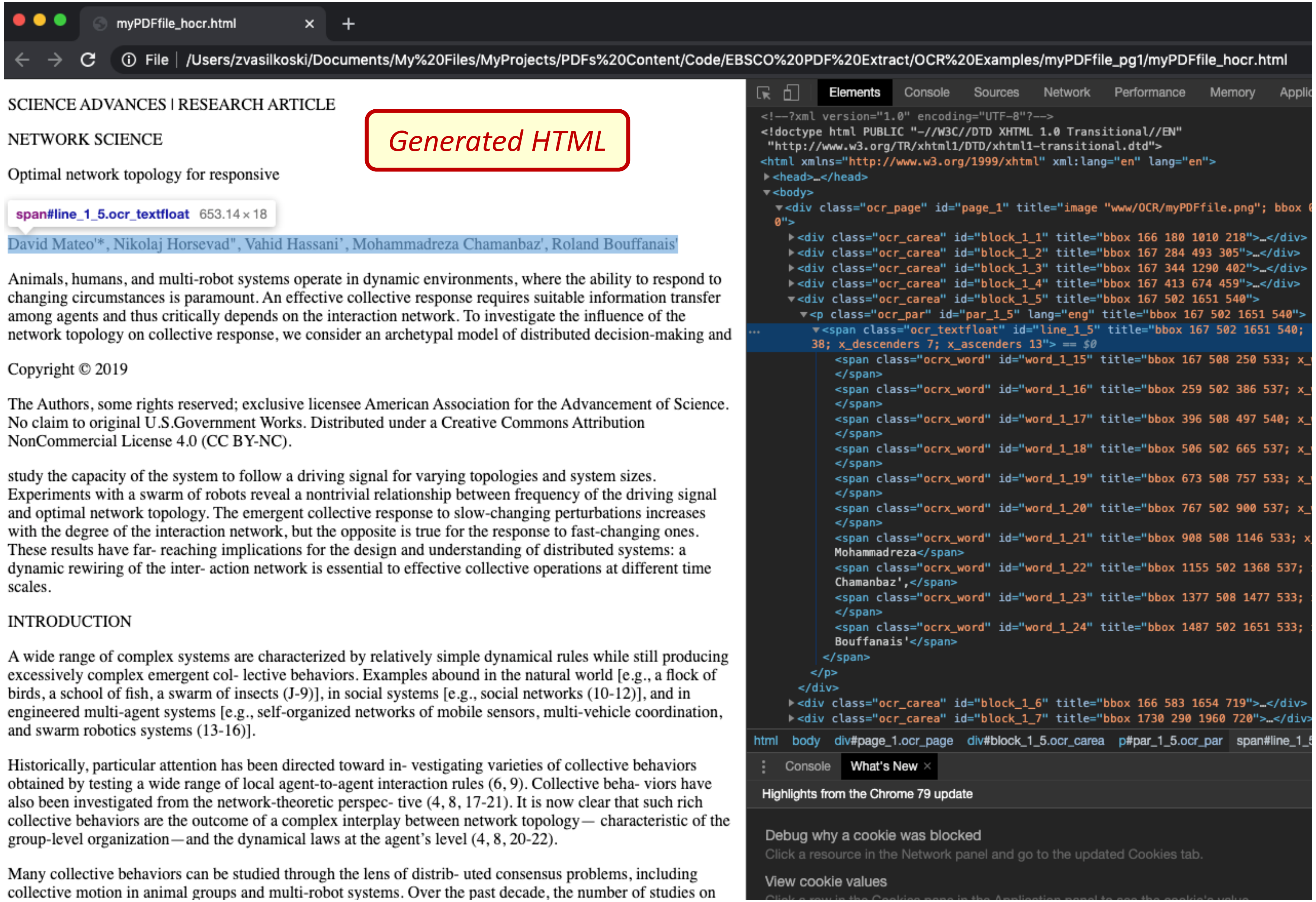
INTRODUCTION

A wide range of complex systems are characterized by relatively simple dynamical rules while still producing excessively complex emergent collective behaviors. Examples abound in the natural world [e.g., a flock of birds, a school of fish, a swarm of insects (*1–9*)], in social systems [e.g., social networks (*10–12*)], and in engineered multi-agent systems [e.g., self-organized networks of mobile sensors, multi-vehicle coordination, and swarm robotics systems (*13–16*)].

Historically, particular attention has been directed toward investigating varieties of collective behaviors obtained by testing a wide range of local agent-to-agent interaction rules (*6, 9*). Collective behaviors have also been investigated from the network-theoretic perspective (*4, 8, 17–21*). It is now clear that such rich collective behaviors are the outcome of a complex interplay between network topology—characteristic of the group-level organization—and the dynamical laws at the agent's level (*4, 8, 20–22*).

Many collective behaviors can be studied through the lens of distrib-

However,
properties
or respons
bly less att
It is im
necessarily
ferromagn
tization—
protocol. I
zation) an
temperatu
turbations
Similar
observed t
ing large
sponse (5)
that one r



/Users/zvasilkoski/Documents/My%20Files/MyProjects/PDFs%20Content/Code/EBSCO%20PDF%20Extract/OCR%20Examples/myPDFfile_pg1/myPDFfile_hocr.html

Generated HTML

SCIENCE ADVANCES | RESEARCH ARTICLE

NETWORK SCIENCE

Optimal network topology for responsive

span#line_1_5.ocr_textfloat 653.14 × 18

David Mateo'*, Nikolaj Horsevad", Vahid Hassani', Mohammadreza Chamanbaz', Roland Bouffanais'

Animals, humans, and multi-robot systems operate in dynamic environments, where the ability to respond to changing circumstances is paramount. An effective collective response requires suitable information transfer among agents and thus critically depends on the interaction network. To investigate the influence of the network topology on collective response, we consider an archetypal model of distributed decision-making and

Copyright © 2019

The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S.Government Works. Distributed under a Creative Commons Attribution NonCommercial License 4.0 (CC BY-NC).

study the capacity of the system to follow a driving signal for varying topologies and system sizes. Experiments with a swarm of robots reveal a nontrivial relationship between frequency of the driving signal and optimal network topology. The emergent collective response to slow-changing perturbations increases with the degree of the interaction network, but the opposite is true for the response to fast-changing ones. These results have far- reaching implications for the design and understanding of distributed systems: a dynamic rewiring of the inter- action network is essential to effective collective operations at different time scales.

INTRODUCTION

A wide range of complex systems are characterized by relatively simple dynamical rules while still producing excessively complex emergent col- lective behaviors. Examples abound in the natural world [e.g., a flock of birds, a school of fish, a swarm of insects (J-9)], in social systems [e.g., social networks (10-12)], and in engineered multi-agent systems [e.g., self-organized networks of mobile sensors, multi-vehicle coordination, and swarm robotics systems (13-16)].

Historically, particular attention has been directed toward in- vestigating varieties of collective behaviors obtained by testing a wide range of local agent-to-agent interaction rules (6, 9). Collective beha- viors have also been investigated from the network-theoretic perspec- tive (4, 8, 17-21). It is now clear that such rich collective behaviors are the outcome of a complex interplay between network topology— characteristic of the group-level organization—and the dynamical laws at the agent's level (4, 8, 20-22).

Many collective behaviors can be studied through the lens of distrib- uted consensus problems, including collective motion in animal groups and multi-robot systems. Over the past decade, the number of studies on

```
<!--?xml version="1.0" encoding="UTF-8"?-->
<!doctype html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"
 "http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
<html xmlns="http://www.w3.org/1999/xhtml" xml:lang="en" lang="en">
 ▶<head>…</head>
 ▼<body>
  ▼<div class="ocr_page" id="page_1" title="image "www/OCR/myPDFfile.png"; bbox 0
   0">
    ▶<div class="ocr_carea" id="block_1_1" title="bbox 166 180 1010 218">…</div>
    ▶<div class="ocr_carea" id="block_1_2" title="bbox 167 284 493 305">…</div>
    ▶<div class="ocr_carea" id="block_1_3" title="bbox 167 344 1290 402">…</div>
    ▶<div class="ocr_carea" id="block_1_4" title="bbox 167 413 674 459">…</div>
    ▼<div class="ocr_carea" id="block_1_5" title="bbox 167 502 1651 540">
     ▼<p class="ocr_par" id="par_1_5" lang="eng" title="bbox 167 502 1651 540">
…     ▼<span class="ocr_textfloat" id="line_1_5" title="bbox 167 502 1651 540;
       38; x_descenders 7; x_ascenders 13"> == $0
         <span class="ocrx_word" id="word_1_15" title="bbox 167 508 250 533; x_
         </span>
         <span class="ocrx_word" id="word_1_16" title="bbox 259 502 386 537; x_
         </span>
         <span class="ocrx_word" id="word_1_17" title="bbox 396 508 497 540; x_
         </span>
         <span class="ocrx_word" id="word_1_18" title="bbox 506 502 665 537; x_
         </span>
         <span class="ocrx_word" id="word_1_19" title="bbox 673 508 757 533; x_
         </span>
         <span class="ocrx_word" id="word_1_20" title="bbox 767 502 900 537; x_
         </span>
         <span class="ocrx_word" id="word_1_21" title="bbox 908 508 1146 533; x
         Mohammadreza</span>
         <span class="ocrx_word" id="word_1_22" title="bbox 1155 502 1368 537;
         Chamanbaz',</span>
         <span class="ocrx_word" id="word_1_23" title="bbox 1377 508 1477 533;
         </span>
         <span class="ocrx_word" id="word_1_24" title="bbox 1487 502 1651 533;
         Bouffanais'</span>
       </span>
     </p>
    </div>
    ▶<div class="ocr_carea" id="block_1_6" title="bbox 166 583 1654 719">…</div>
    ▶<div class="ocr_carea" id="block_1_7" title="bbox 1730 290 1960 720">…</div>
```

Highlights from the Chrome 79 update

Debug why a cookie was blocked

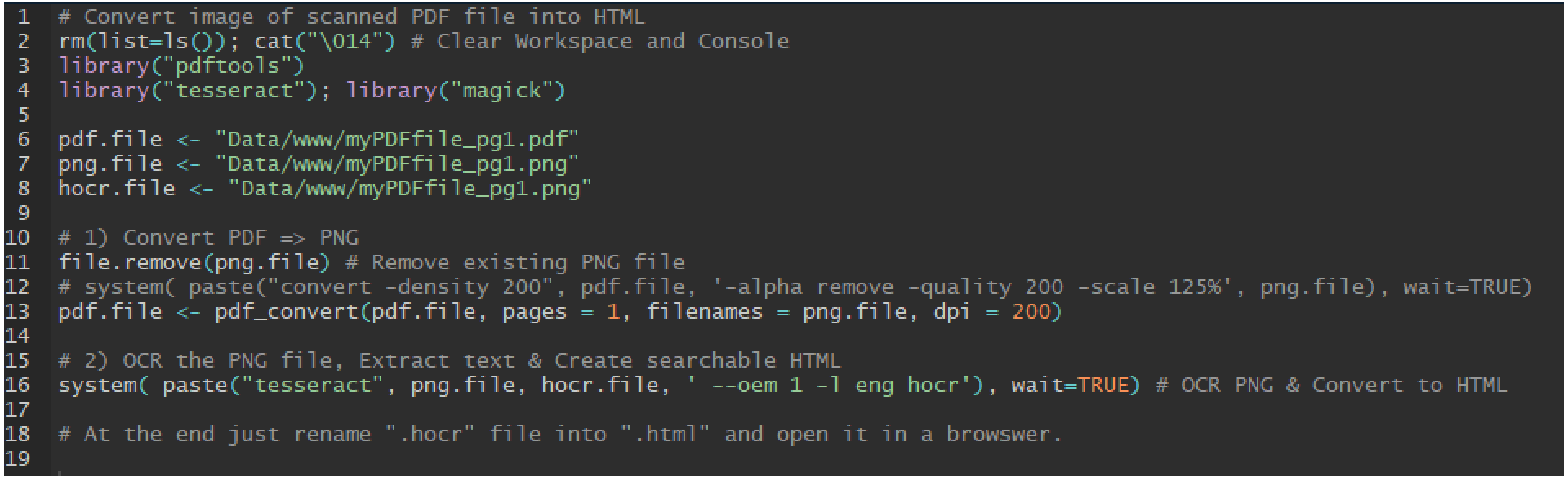
Click a resource in the Network panel and go to the updated Cookies tab.

View cookie values

Example 4: OCR a PNG to create HTML file

Convert the PNG image to a HTML file using the following syntax:

- "tesseract Data/www/myPDFfile_pg1.png Data/www/myPDFfile_pg1.png --oem 1 -l eng hocr“
- This will produce a file with “hocr” extension that you need to rename it with “html” extension before you open it in a browser.

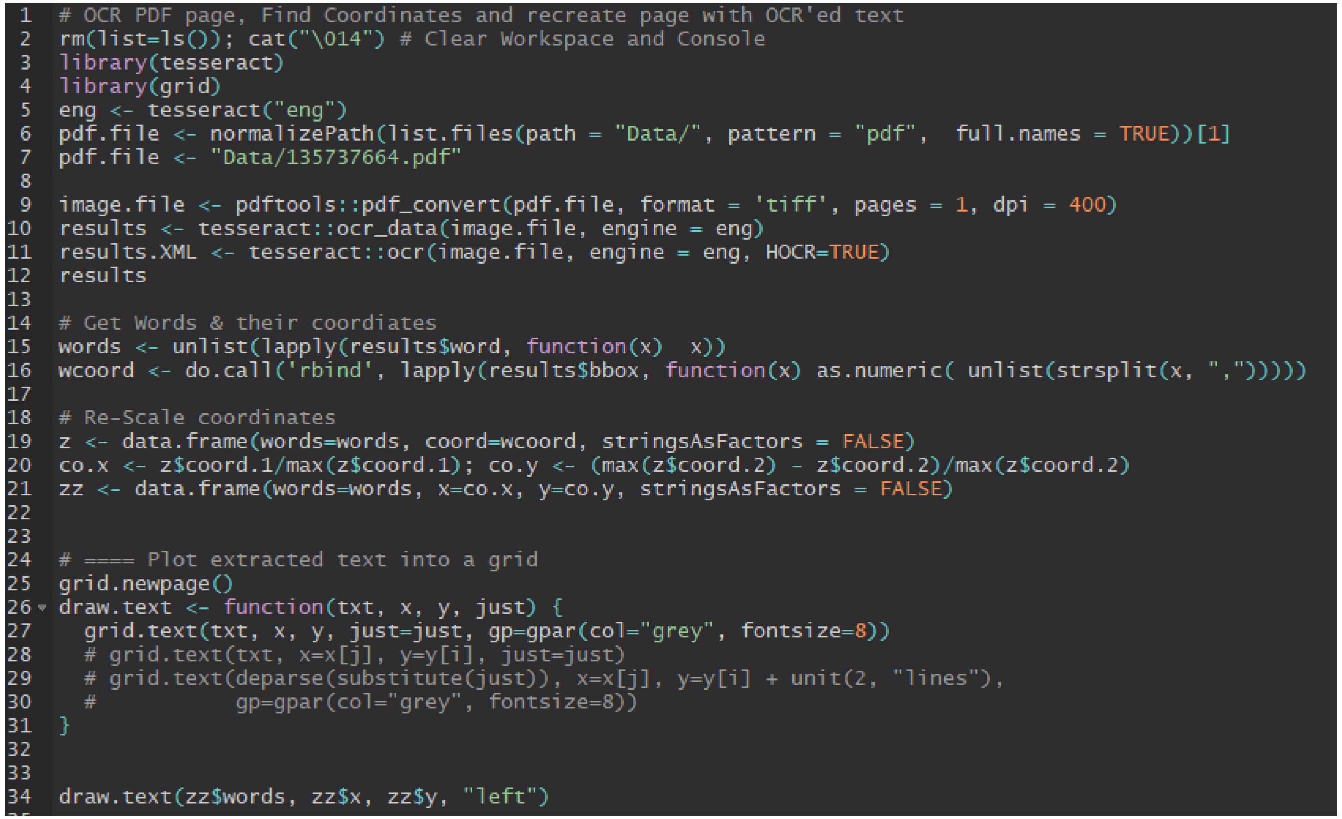


```r
# Convert image of scanned PDF file into HTML
rm(list=ls()); cat("\014") # Clear Workspace and Console
library("pdftools")
library("tesseract"); library("magick")

pdf.file <- "Data/www/myPDFfile_pg1.pdf"
png.file <- "Data/www/myPDFfile_pg1.png"
hocr.file <- "Data/www/myPDFfile_pg1.png"

# 1) Convert PDF => PNG
file.remove(png.file) # Remove existing PNG file
# system( paste("convert -density 200", pdf.file, '-alpha remove -quality 200 -scale 125%', png.file), wait=TRUE)
pdf.file <- pdf_convert(pdf.file, pages = 1, filenames = png.file, dpi = 200)

# 2) OCR the PNG file, Extract text & Create searchable HTML
system( paste("tesseract", png.file, hocr.file, ' --oem 1 -l eng hocr'), wait=TRUE) # OCR PNG & Convert to HTML

# At the end just rename ".hocr" file into ".html" and open it in a browswer.

```

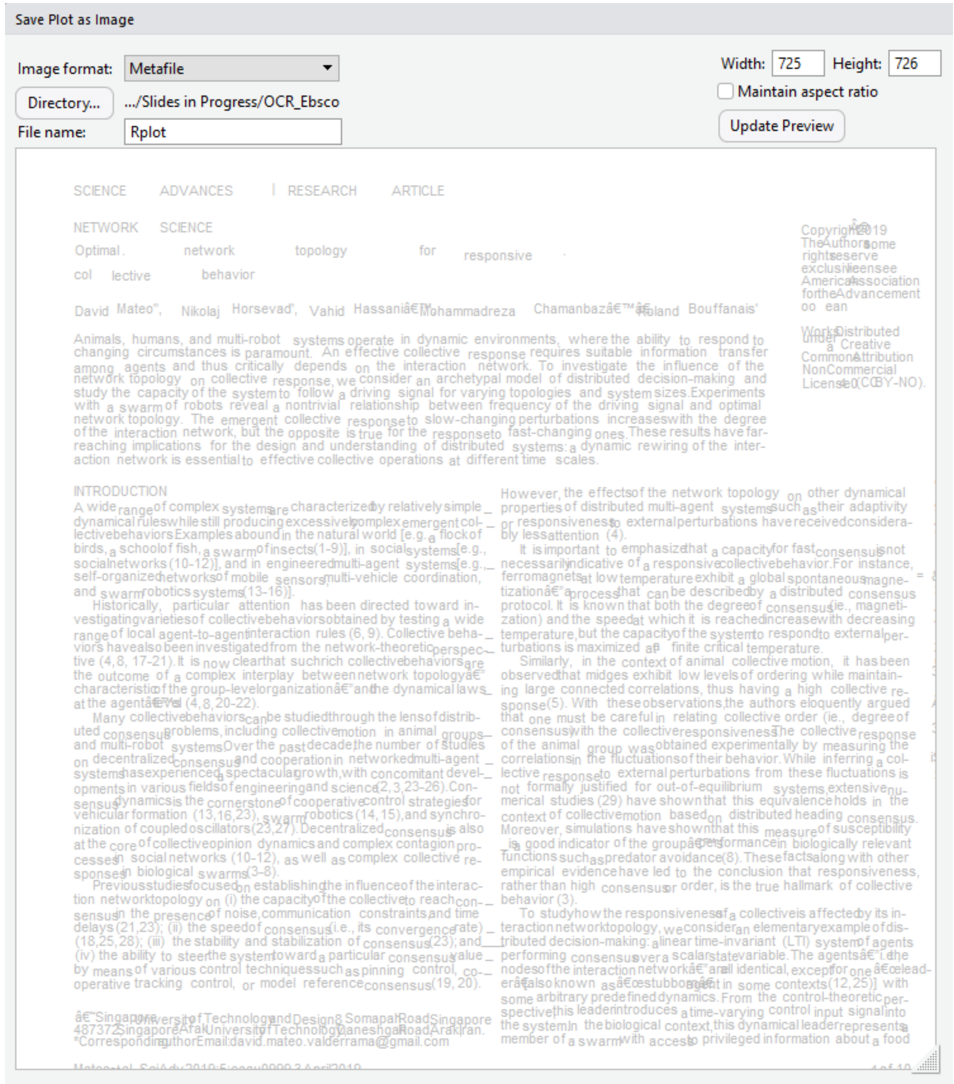
Find Coordinates of the OCR words

Another Tesseract feature is to

- OCR a PDF page,
- Find Coordinates and
- recreate page with OCR'ed text



```r
# OCR PDF page, Find Coordinates and recreate page with OCR'ed text
rm(list=ls()); cat("\014") # Clear Workspace and Console
library(tesseract)
library(grid)
eng <- tesseract("eng")
pdf.file <- normalizePath(list.files(path = "Data/", pattern = "pdf",  full.names = TRUE))[1]
pdf.file <- "Data/135737664.pdf"

image.file <- pdftools::pdf_convert(pdf.file, format = 'tiff', pages = 1, dpi = 400)
results <- tesseract::ocr_data(image.file, engine = eng)
results.XML <- tesseract::ocr(image.file, engine = eng, HOCR=TRUE)
results

# Get Words & their coordiates
words <- unlist(lapply(results$word, function(x)  x))
wcoord <- do.call('rbind', lapply(results$bbox, function(x) as.numeric( unlist(strsplit(x, ",")))))

# Re-Scale coordinates
z <- data.frame(words=words, coord=wcoord, stringsAsFactors = FALSE)
co.x <- z$coord.1/max(z$coord.1); co.y <- (max(z$coord.2) - z$coord.2)/max(z$coord.2)
zz <- data.frame(words=words, x=co.x, y=co.y, stringsAsFactors = FALSE)


# ==== Plot extracted text into a grid
grid.newpage()
draw.text <- function(txt, x, y, just) {
  grid.text(txt, x, y, just=just, gp=gpar(col="grey", fontsize=8))
  # grid.text(txt, x=x[j], y=y[i], just=just)
  # grid.text(deparse(substitute(just)), x=x[j], y=y[i] + unit(2, "lines"),
  #           gp=gpar(col="grey", fontsize=8))
}


draw.text(zz$words, zz$x, zz$y, "left")
```



Save Plot as Image

Image format: Metafile

Directory... .../Slides in Progress/OCR_Ebsco

File name: Rplot

Width: 725 Height: 726

□ Maintain aspect ratio

SCIENCE ADVANCES | RESEARCH ARTICLE

NETWORK SCIENCE

Optimal. network topology for responsive col lective behavior

David Mateo", Nikolaj Horsevad', Vahid Hassaniâ€™, Mohammadreza Chamanbazâ€™â€, Roland Bouffanais'

Copyright Â© 2019 The Authors, some rights reserved; exclusive licensee American Association for the Advancement oo ean Works. Distributed under a Creative Commons Attribution NonCommercial License 4.0 (CC BY-NO).

Animals, humans, and multi-robot systems operate in dynamic environments, where the ability to respond to changing circumstances is paramount. An effective collective response requires suitable information transfer among agents and thus critically depends on the interaction network. To investigate the influence of the network topology on collective response, we consider an archetypal model of distributed decision-making and study the capacity of the system to follow a driving signal for varying topologies and system sizes. Experiments with a swarm of robots reveal a nontrivial relationship between frequency of the driving signal and optimal network topology. The emergent collective response to slow-changing perturbations increases with the degree of the interaction network, but the opposite is true for the response to fast-changing ones. These results have far-reaching implications for the design and understanding of distributed systems: a dynamic rewiring of the interaction network is essential to effective collective operations at different time scales.

INTRODUCTION

A wide range of complex systems are characterized by relatively simple dynamical rules while still producing excessively complex emergent collective behaviors. Examples abound in the natural world [e.g., a flock of birds, a school of fish, a swarm of insects (1-9)], in social systems [e.g., social networks (10-12)], and in engineered multi-agent systems [e.g., self-organized networks of mobile sensors, multi-vehicle coordination, and swarm robotics systems (13-16)].

Historically, particular attention has been directed toward investigating varieties of collective behaviors obtained by testing a wide range of local agent-to-agent interaction rules (6, 9). Collective behaviors have also been investigated from the network-theoretic perspective (4, 8, 17-21). It is now clear that such rich collective behaviors are the outcome of a complex interplay between network topologyâ€" characteristic of the group-level organizationâ€" and the dynamical laws at the agentâ€™s level (4, 8, 20-22).

Many collective behaviors can be studied through the lens of distributed consensus problems, including collective motion in animal groups and multi-robot systems. Over the past decade, the number of studies on decentralized consensus and cooperation in networked multi-agent systems has experienced a spectacular growth, with concomitant developments in various fields of engineering and science (2, 3, 23-26). Consensus dynamics is the cornerstone of cooperative control strategies for vehicular formation (13, 16, 23), swarm robotics (14, 15), and synchronization of coupled oscillators (23, 27). Decentralized consensus is also at the core of collective opinion dynamics and complex contagion processes in social networks (10-12), as well as complex collective responses in biological swarms (3-8).

Previous studies focused on establishing the influence of the interaction network topology on (i) the capacity of the collective to reach consensus in the presence of noise, communication constraints, and time delays (21, 23); (ii) the speed of consensus (i.e., its convergence rate) (18, 25, 28); (iii) the stability and stabilization of consensus (23); and (iv) the ability to steer the system toward a particular consensus value by means of various control techniques such as pinning control, cooperative tracking control, or model reference consensus (19, 20).

â€˜Singapore University of Technology and Design, 8 Somapah Road, Singapore 487372, Singapore. Arak University of Technology, Daneshgah Road, Arak, Iran. *Corresponding author. Email: david.mateo.valderrama@gmail.com

However, the effects of the network topology on other dynamical properties of distributed multi-agent systems such as their adaptivity or responsiveness to external perturbations have received considerably less attention (4).

It is important to emphasize that a capacity for fast consensus is not necessarily indicative of a responsive collective behavior. For instance, ferromagnets at low temperature exhibit a global spontaneous magnetizationâ€"a process that can be described by a distributed consensus protocol. It is known that both the degree of consensus (ie., magnetization) and the speed at which it is reached increase with decreasing temperature, but the capacity of the system to respond to external perturbations is maximized at a finite critical temperature.

Similarly, in the context of animal collective motion, it has been observed that midges exhibit low levels of ordering while maintaining large connected correlations, thus having a high collective response (5). With these observations, the authors eloquently argued that one must be careful in relating collective order (ie., degree of consensus) with the collective responsiveness. The collective response of the animal group was obtained experimentally by measuring the correlations in the fluctuations of their behavior. While inferring a collective response to external perturbations from these fluctuations is not formally justified for out-of-equilibrium systems, extensive numerical studies (29) have shown that this equivalence holds in the context of collective motion based on distributed heading consensus. Moreover, simulations have shown that this measure of susceptibility is a good indicator of the groupâ€™s performance in biologically relevant functions such as predator avoidance (8). These facts along with other empirical evidence have led to the conclusion that responsiveness, rather than high consensus or order, is the true hallmark of collective behavior (3).

To study how the responsiveness of a collective is affected by its interaction network topology, we consider an elementary example of distributed decision-making: a linear time-invariant (LTI) system of agents performing consensus over a scalar state variable. The agentsâ€"i.e., the nodes of the interaction networkâ€"are all identical, except for one â€œleaderâ€ (also known as â€œstubbornâ€ agent in some contexts (12, 25)] with some arbitrary predefined dynamics. From the control-theoretic perspective, this leader introduces a time-varying control input signal into the system. In the biological context, this dynamical leader represents a member of a swarm with access to privileged information about a food

Mateo et al., Sci Adv 2019;5:eaau0000 3 April 2019

1 of 10