Accurate Word Extraction from Documents with Complex Layouts

Bachelor thesis

Author: Tanyu Tanev
Supervisor: Claudius Korzen
Examiner: Prof. Dr. Hannah Bast

Chair for Algorithms and Data Structures
Department of Computer Science
University of Freiburg

October 11, 2019

Table of Contents

Table of Contents

PDF files

- Self-contained file format, which shows a document **exactly** how its author intended

PDF files

- Self-contained file format, which shows a document **exactly** how its author intended

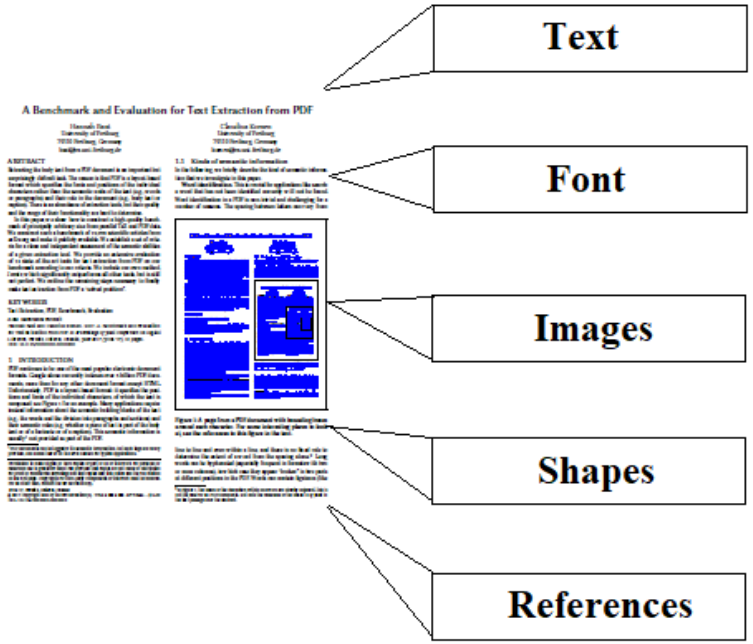


Figure: Contents of the PDF document "A Benchmark and Evaluation for Text Extraction from PDF" [2]

PDF files

- Self-contained file format, which shows a document **exactly** how its author intended
- They comprise a bigger part of the Internet with every day [1]

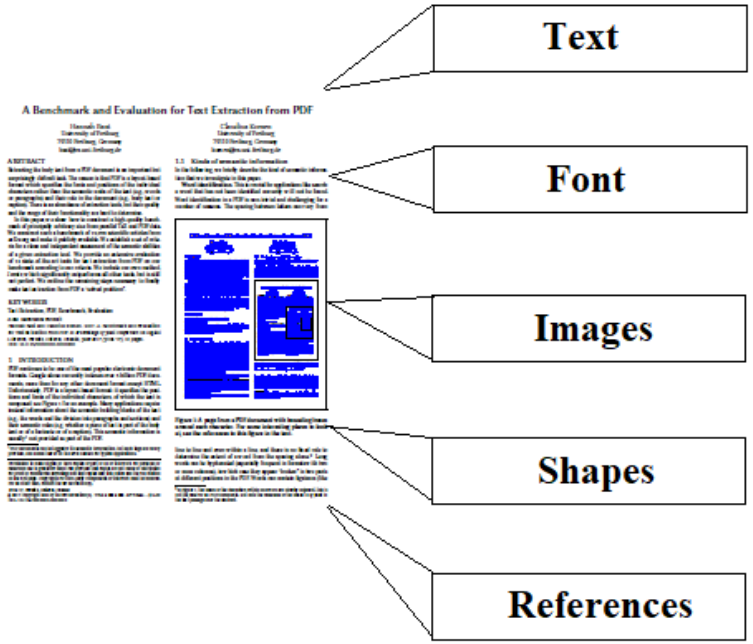


Figure: Contents of the PDF document "A Benchmark and Evaluation for Text Extraction from PDF" [2]

Content extraction 1/2

- Automatic **content extraction** is a a challenging topic with many sub-problems:

Content extraction 1/2

- Automatic **content extraction** is a a challenging topic with many sub-problems:
  - extraction of **textual content**
  - grouping text blocks according to **semantic role**
  - **dehyphenation**
  - and others...

Content extraction 1/2

- Automatic **content extraction** is a a challenging topic with many sub-problems:
  - extraction of **textual content**
  - grouping text blocks according to **semantic role**
  - **dehyphenation**
  - and others...
- This bachelor thesis focuses on the extraction of **textual content**, which enables:

Content extraction 1/2

- Automatic **content extraction** is a a challenging topic with many sub-problems:
  - extraction of **textual content**
  - grouping text blocks according to **semantic role**
  - **dehyphenation**
  - and others...
- This bachelor thesis focuses on the extraction of **textual content**, which enables:
  - Search engine indexing,
  - Conversion from PDF to other file formats,
  - and others...

Content extraction 2/2

- **Problem:** PDF is layout based - main goal is showing its contents in their proper form and position

Content extraction 2/2

- **Problem:** PDF is layout based - main goal is showing its contents in their proper form and position
- **Consequence:** Textual information is only stored on **character level**, with:
  - no **connections** to their respective semantic blocks (e.g. **words** and **text lines**)
  - no information about **whitespaces**

Content extraction 2/2

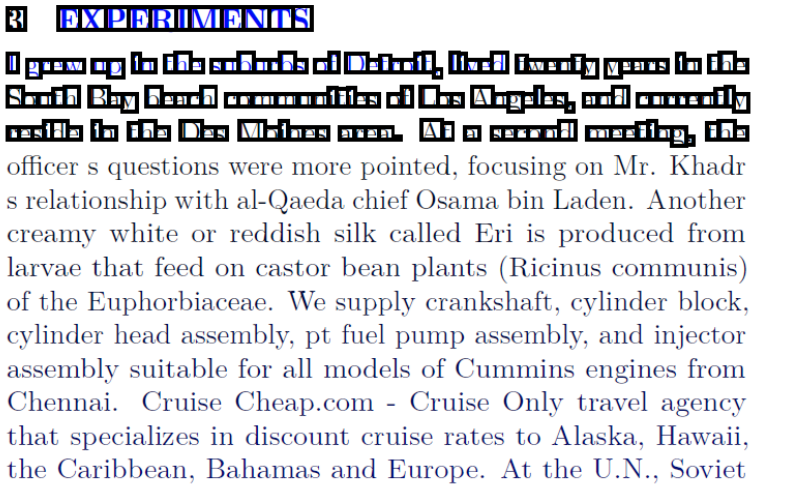


Figure: How textual content is stored on a character level

Content extraction 2/2

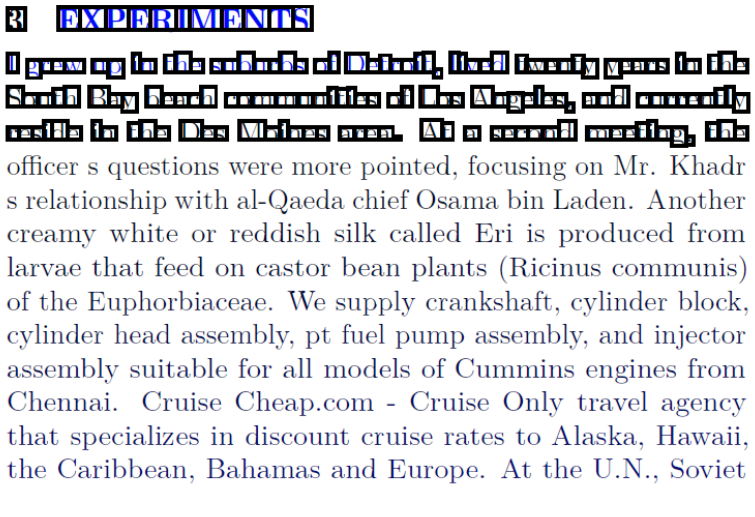


Figure: How textual content is stored on a character level

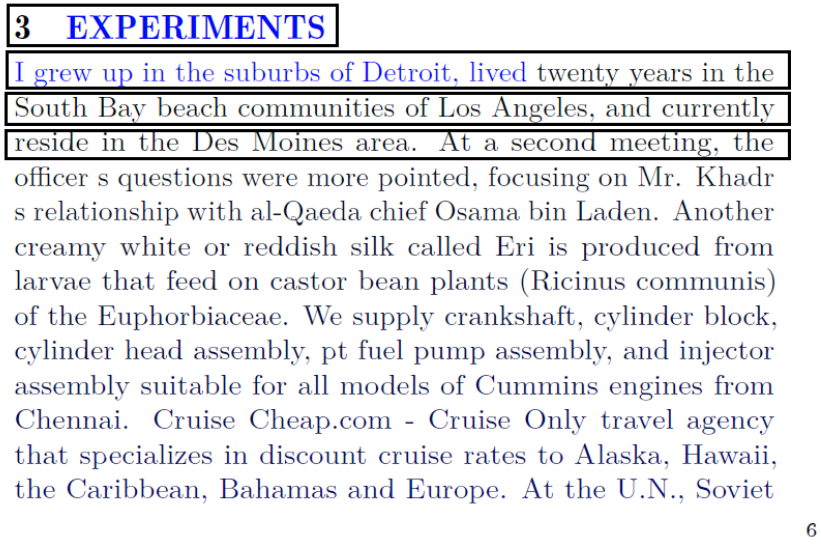


Figure: How the textual content from the left has to be extracted

Rule-based approaches

- **Solution:** Group characters to bigger blocks

Rule-based approaches

- **Solution:** Group characters to bigger blocks
- **How?**

Rule-based approaches

- **Solution:** Group characters to bigger blocks
- **How?**
  - **Top-down** approaches - segment a page into smaller blocks (e.g. **columns**) and continue way downwards
  - **Bottom-up** approaches - first group individual characters to **words** and **text lines** and work upwards to bigger blocks

Recursive X-Y cut 1/2

- Recursive **top-to-bottom** algorithm for page segmentation
- **Alternate** between "cutting" a PDF page horizontally and vertically
- **When to cut?**

Recursive X-Y cut 1/2

- Recursive **top-to-bottom** algorithm for page segmentation
- **Alternate** between "cutting" a PDF page horizontally and vertically
- **When to cut?** when there are gaps of white space,
  bigger than a **custom defined** threshold
  (e.g. most common **font size** or **character size**)
- **Termination:** There are no possible cuts left (in **both** directions)

Recursive X-Y cut 2/2

Efficient Generation of Geographically Accurate Transit Maps

Hannah Bast
University of Freiburg
Georges Köhler Allee 051
Freiburg 79110

Patrick Brosi
University of Freiburg
Georges Köhler Allee 051
Freiburg 79110

Markus Näther
University of Freiburg
Georges Köhler Allee 051
Freiburg 79110

Claudius Korzen
University of Freiburg
Georges Köhler Allee 051
Freiburg 79110

Elmar Haussmann
University of Freiburg
Georges Köhler Allee 051
Freiburg 79110

1 ABSTRACT

In January this year, Rosselli refused to participate in the election of officers of the 650,000-member SEIU California State Council, or to run for reelection as its president, accusing Stern of rigging the process to guarantee the success of his own handpicked choices. The morning after Captain Cook s arrival in Queen Charlotte s Sound, he went himself, at daybreak, to look for scurvy-grass, celery, and other vegetables; and he had the good fortune to return with a boatload, in a very short space of time. This encompassed Pell s lands south to Westchester, the settlement that he helped found in late 1654. This powerful quote that guided Nobel Peace Prize Laureate Wangari Maathai to create sustainable change in Africa provided the same inspiration to well over 1500 women across North America. The Java Servlet and PHP session mechanisms both use this method if cookies are not enabled. When Stabroek News was at Marlborough in the Pomeroon River on Tuesday, five families were preparing to make an eight-hour and 64-mile journey up the Akiwini Creek to obtain fresh water.

While I have done Linux and Windows 2000, I have not seen the Solaris 7 and Windows 2000 combination. Let s welcome the Chinese New Year ringing in the Year of The Rabbit on February 3, 2011. He commanded the Australians for much of the First World War, including during the Gallipoli Campaign, and as a result the collection is in constant use by researchers. The Isotope s annual event coinciding with San Francisco s Alternative Press Expo featuring the awarding of the Isotope Award For Excellence in Mini-Comics. It has a population of 39 million people, more or less equal to that of New England plus New York State and New Jersey. In the presidential election of 1999, opposition leader Abdoulaye Wade defeated Diouf in an election deemed free and fair by international observers. The workshop was led by Sue Becker (Psychology, McMaster University) with a lecture by Dieter Jaeger (Biology, Emory University). This is just a guess on the reason for the implementation difference but I know that Cisco pushed LDP and Juniper pushed RSVP at the standard when MPLS first came around. Yet Rowling s publisher, Scholastic, had the good sense to low-key her tour, teaming with the independent booksellers association.

2 INTRODUCTION

Sambasiva R. Kottamasu, MD, of Saginaw, Mich, was recently appointed to the Central Michigan University Board of Control. The project also provides coverage in four downtown Seattle parks: Occidental, Freeway, Westlake and Victor Steinbrueck, as well as the City Hall lobby area. February 2nd 2009 05:01 Midway through season 2008, I ran an review of the top 10 handballers in the league HERE. The voters in California and Arizona have recognized this at the ballot box.

Freeware.For Windows 95, 98, ME, NT, 2000, XP. From the latter part of the 1970s, Royal Ordnance investigated the development of light space launch vehicles, based on the use of the Stonechat motor. What was of particular interest was the discovery by him in 1850 of names and events that occurred in the Old Testament, such as Judah, Hezekiah, Jerusalem and Samaria. What we needed to understand was that the good works that the sheep did flowed from a heart that truly loved God, and that loving heart was given to them by the God who elected them to put their faith in Jesus and to do those good works for his own glory. He also encouraged Northwestern students who had Quaker ties to attend Evanston Meeting and, with Royal, frequently hosted teas or breakfasts to welcome them. The city of Belgrade (the White City) was a center for the Ottoman administration, a janissary garrison and Muslim shopkeepers. He s the fellow who claimed that his own work invalidated Rupert Sheldrake s research on pet telepathy, but when Wiseman s results were analyzed, they were found to match the same patterns Sheldrake had identified! For instance, Seam offers a wide variety of ways to express these restrictions through roles, declarative rules, and access control lists (ACLs). Holland, McMeel Family Professor in Shakespeare Studies, was elected an honorary fellow at Trinity Hall, his alma mater and one of the 31 colleges that comprise the University of Cambridge.

3 EXPERIMENTS

The phony democracy is hijacked by the terrorists, MB, Israel and the pigs of GCC. Power output is 28 BHP in standard form with 41.3 Nm of torque and tested under Euro 3 conditions the engine returns just under 80 mpg, which means an impressive tank range of over 200 miles before switching to reserve. As Levy sees it, after the collapse

Figure: A page from a randomly generated PDF document

Recursive X-Y cut 2/2

Efficient Generation of Geographically Accurate Transit Maps

Hannah Bast
University of Freiburg
Georges Köhler Allee 051
Freiburg 79110

Patrick Brosi
University of Freiburg
Georges Köhler Allee 051
Freiburg 79110

Markus Näther
University of Freiburg
Georges Köhler Allee 051
Freiburg 79110

Claudius Korzen
University of Freiburg
Georges Köhler Allee 051
Freiburg 79110

Elmar Haussmann
University of Freiburg
Georges Köhler Allee 051
Freiburg 79110

1 ABSTRACT

In January this year, Rosselli refused to participate in the election of officers of the 650,000-member SEIU California State Council, or to run for reelection as its president, accusing Stern of rigging the process to guarantee the success of his own handpicked choices. The morning after Captain Cook s arrival in Queen Charlotte s Sound, he went himself, at daybreak, to look for scurvy-grass, celery, and other vegetables; and he had the good fortune to return with a boatload, in a very short space of time. This encompassed Pell s lands south to Westchester, the settlement that he helped found in late 1654. This powerful quote that guided Nobel Peace Prize Laureate Wangari Maathai to create sustainable change in Africa provided the same inspiration to well over 1500 women across North America. The Java Servlet and PHP session mechanisms both use this method if cookies are not enabled. When Stabroek News was at Marlborough in the Pomeroon River on Tuesday, five families were preparing to make an eight-hour and 64-mile journey up the Akiwini Creek to obtain fresh water.

While I have done Linux and Windows 2000, I have not seen the Solaris 7 and Windows 2000 combination. Let s welcome the Chinese New Year ringing in the Year of The Rabbit on February 3, 2011. He commanded the Australians for much of the First World War, including during the Gallipoli Campaign, and as a result the collection is in constant use by researchers. The Isotope s annual event coinciding with San Francisco s Alternative Press Expo featuring the awarding of the Isotope Award For Excellence in Mini-Comics. It has a population of 39 million people, more or less equal to that of New England plus New York State and New Jersey. In the presidential election of 1999, opposition leader Abdoulaye Wade defeated Diouf in an election deemed free and fair by international observers. The workshop was led by Sue Becker (Psychology, McMaster University) with a lecture by Dieter Jaeger (Biology, Emory University). This is just a guess on the reason for the implementation difference but I know that Cisco pushed LDP and Juniper pushed RSVP at the standard when MPLS first came around. Yet Rowling s publisher, Scholastic, had the good sense to low-key her tour, teaming with the independent booksellers association.

2 INTRODUCTION

Sambasiva R. Kottamasu, MD, of Saginaw, Mich, was recently appointed to the Central Michigan University Board of Control. The project also provides coverage in four downtown Seattle parks: Occidental, Freeway, Westlake and Victor Steinbrueck, as well as the City Hall lobby area. February 2nd 2009 05:01 Midway through season 2008, I ran an review of the top 10 handballers in the league HERE. The voters in California and Arizona have recognized this at the ballot box.

Freeware.For Windows 95, 98, ME, NT, 2000, XP. From the latter part of the 1970s, Royal Ordnance investigated the development of light space launch vehicles, based on the use of the Stonechat motor. What was of particular interest was the discovery by him in 1850 of names and events that occurred in the Old Testament, such as Judah, Hezekiah, Jerusalem and Samaria. What we needed to understand was that the good works that the sheep did flowed from a heart that truly loved God, and that loving heart was given to them by the God who elected them to put their faith in Jesus and to do those good works for his own glory. He also encouraged Northwestern students who had Quaker ties to attend Evanston Meeting and, with Royal, frequently hosted teas or breakfasts to welcome them. The city of Belgrade (the White City) was a center for the Ottoman administration, a janissary garrison and Muslim shopkeepers. He s the fellow who claimed that his own work invalidated Rupert Sheldrake s research on pet telepathy, but when Wiseman s results were analyzed, they were found to match the same patterns Sheldrake had identified! For instance, Seam offers a wide variety of ways to express these restrictions through roles, declarative rules, and access control lists (ACLs). Holland, McMeel Family Professor in Shakespeare Studies, was elected an honorary fellow at Trinity Hall, his alma mater and one of the 31 colleges that comprise the University of Cambridge.

3 EXPERIMENTS

The phony democracy is hijacked by the terrorists, MB, Israel and the pigs of GCC. Power output is 28 BHP in standard form with 41.3 Nm of torque and tested under Euro 3 conditions the engine returns just under 80 mpg, which means an impressive tank range of over 200 miles before switching to reserve. As Levy sees it, after the collapse

Figure: First (vertical) cut of the PDF page

Recursive X-Y cut 2/2

Hannah Bast
University of Freiburg
Georges Köhler Allee 051
Freiburg 79110

Patrick Brosi
University of Freiburg
Georges Köhler Allee 051
Freiburg 79110

Markus Näther
University of Freiburg
Georges Köhler Allee 051
Freiburg 79110

Claudius Korzen
University of Freiburg
Georges Köhler Allee 051
Freiburg 79110

Elmar Haussmann
University of Freiburg
Georges Köhler Allee 051
Freiburg 79110

1 ABSTRACT

In January this year, Rosselli refused to participate in the election of officers of the 650,000-member SEIU California State Council, or to run for reelection as its president, accusing Stern of rigging the process to guarantee the success of his own handpicked choices. The morning after Captain Cook s arrival in Queen Charlotte s Sound, he went himself, at daybreak, to look for scurvy-grass, celery, and other vegetables; and he had the good fortune to return with a boatload, in a very short space of time. This encompassed Pell s lands south to Westchester, the settlement that he helped found in late 1654. This powerful quote that guided Nobel Peace Prize Laureate Wangari Maathai to create sustainable change in Africa provided the same inspiration to well over 1500 women across North America. The Java Servlet and PHP session mechanisms both use this method if cookies are not enabled. When Stabroek News was at Marlborough in the Pomeroon River on Tuesday, five families were preparing to make an eight-hour and 64-mile journey up the Akiwini Creek to obtain fresh water.

While I have done Linux and Windows 2000, I have not seen the Solaris 7 and Windows 2000 combination. Let s welcome the Chinese New Year ringing in the Year of The Rabbit on February 3, 2011. He commanded the Australians for much of the First World War, including during the Gallipoli Campaign, and as a result the collection is in constant use by researchers. The Isotope s annual event coinciding with San Francisco s Alternative Press Expo featuring the awarding of the Isotope Award For Excellence in Mini-Comics. It has a population of 39 million people, more or less equal to that of New England plus New York State and New Jersey. In the presidential election of 1999, opposition leader Abdoulaye Wade defeated Diouf in an election deemed free and fair by international observers. The workshop was led by Sue Becker (Psychology, McMaster University) with a lecture by Dieter Jaeger (Biology, Emory University). This is just a guess on the reason for the implementation difference but I know that Cisco pushed LDP and Juniper pushed RSVP at the standard when MPLS first came around. Yet Rowling s publisher, Scholastic, had the good sense to low-key her tour, teaming with the independent booksellers association.

2 INTRODUCTION

Sambasiva R. Kottamasu, MD, of Saginaw, Mich, was recently appointed to the Central Michigan University Board of Control. The project also provides coverage in four downtown Seattle parks: Occidental, Freeway, Westlake and Victor Steinbrueck, as well as the City Hall lobby area. February 2nd 2009 05:01 Midway through season 2008, I ran an review of the top 10 handballers in the league HERE. The voters in California and Arizona have recognized this at the ballot box.

Freeware.For Windows 95, 98, ME, NT, 2000, XP. From the latter part of the 1970s, Royal Ordnance investigated the development of light space launch vehicles, based on the use of the Stonechat motor. What was of particular interest was the discovery by him in 1850 of names and events that occurred in the Old Testament, such as Judah, Hezekiah, Jerusalem and Samaria. What we needed to understand was that the good works that the sheep did flowed from a heart that truly loved God, and that loving heart was given to them by the God who elected them to put their faith in Jesus and to do those good works for his own glory. He also encouraged Northwestern students who had Quaker ties to attend Evanston Meeting and, with Royal, frequently hosted teas or breakfasts to welcome them. The city of Belgrade (the White City) was a center for the Ottoman administration, a janissary garrison and Muslim shopkeepers. He s the fellow who claimed that his own work invalidated Rupert Sheldrake s research on pet telepathy, but when Wiseman s results were analyzed, they were found to match the same patterns Sheldrake had identified! For instance, Seam offers a wide variety of ways to express these restrictions through roles, declarative rules, and access control lists (ACLs). Holland, McMeel Family Professor in Shakespeare Studies, was elected an honorary fellow at Trinity Hall, his alma mater and one of the 31 colleges that comprise the University of Cambridge.

3 EXPERIMENTS

The phony democracy is hijacked by the terrorists, MB, Israel and the pigs of GCC. Power output is 28 BHP in standard form with 41.3 Nm of torque and tested under Euro 3 conditions the engine returns just under 80 mpg, which means an impressive tank range of over 200 miles before switching to reserve. As Levy sees it, after the collapse

Figure: Second (vertical) cut of the PDF page

Recursive X-Y cut 2/2

Claudius Korzen
University of Freiburg
Georges Köhler Allee 051
Freiburg 79110

1 ABSTRACT

In January this year, Rosselli refused to participate in the election of officers of the 650,000-member SEIU California State Council, or to run for reelection as its president, accusing Stern of rigging the process to guarantee the success of his own handpicked choices. The morning after Captain Cook s arrival in Queen Charlotte s Sound, he went himself, at daybreak, to look for scurvy-grass, celery, and other vegetables; and he had the good fortune to return with a boatload, in a very short space of time. This encompassed Pell s lands south to Westchester, the settlement that he helped found in late 1654. This powerful quote that guided Nobel Peace Prize Laureate Wangari Maathai to create sustainable change in Africa provided the same inspiration to well over 1500 women across North America. The Java Servlet and PHP session mechanisms both use this method if cookies are not enabled. When Stabroek News was at Marlborough in the Pomeroon River on Tuesday, five families were preparing to make an eight-hour and 64-mile journey up the Akiwini Creek to obtain fresh water.

While I have done Linux and Windows 2000, I have not seen the Solaris 7 and Windows 2000 combination. Let s welcome the Chinese New Year ringing in the Year of The Rabbit on February 3, 2011. He commanded the Australians for much of the First World War, including during the Gallipoli Campaign, and as a result the collection is in constant use by researchers. The Isotope s annual event coinciding with San Francisco s Alternative Press Expo featuring the awarding of the Isotope Award For Excellence in Mini-Comics. It has a population of 39 million people, more or less equal to that of New England plus New York State and New Jersey. In the presidential election of 1999, opposition leader Abdoulaye Wade defeated Diouf in an election deemed free and fair by international observers. The workshop was led by Sue Becker (Psychology, McMaster University) with a lecture by Dieter Jaeger (Biology, Emory University). This is just a guess on the reason for the implementation difference but I know that Cisco pushed LDP and Juniper pushed RSVP at the standard when MPLS first came around. Yet Rowling s publisher, Scholastic, had the good sense to low-key her tour, teaming with the independent booksellers association.

Elmar Haussmann
University of Freiburg
Georges Köhler Allee 051
Freiburg 79110

2 INTRODUCTION

Sambasiva R. Kottamasu, MD, of Saginaw, Mich, was recently appointed to the Central Michigan University Board of Control. The project also provides coverage in four downtown Seattle parks: Occidental, Freeway, Westlake and Victor Steinbrueck, as well as the City Hall lobby area. February 2nd 2009 05:01 Midway through season 2008, I ran an review of the top 10 handballers in the league HERE. The voters in California and Arizona have recognized this at the ballot box.

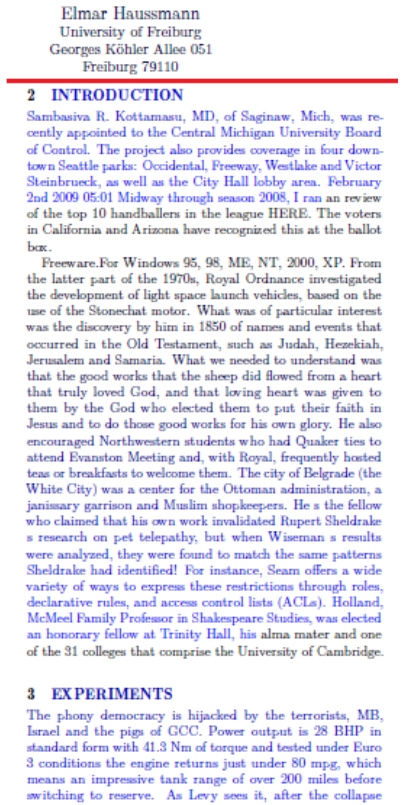
Freeware.For Windows 95, 98, ME, NT, 2000, XP. From the latter part of the 1970s, Royal Ordnance investigated the development of light space launch vehicles, based on the use of the Stonechat motor. What was of particular interest was the discovery by him in 1850 of names and events that occurred in the Old Testament, such as Judah, Hezekiah, Jerusalem and Samaria. What we needed to understand was that the good works that the sheep did flowed from a heart that truly loved God, and that loving heart was given to them by the God who elected them to put their faith in Jesus and to do those good works for his own glory. He also encouraged Northwestern students who had Quaker ties to attend Evanston Meeting and, with Royal, frequently hosted teas or breakfasts to welcome them. The city of Belgrade (the White City) was a center for the Ottoman administration, a janissary garrison and Muslim shopkeepers. He s the fellow who claimed that his own work invalidated Rupert Sheldrake s research on pet telepathy, but when Wiseman s results were analyzed, they were found to match the same patterns Sheldrake had identified! For instance, Seam offers a wide variety of ways to express these restrictions through roles, declarative rules, and access control lists (ACLs). Holland, McMeel Family Professor in Shakespeare Studies, was elected an honorary fellow at Trinity Hall, his alma mater and one of the 31 colleges that comprise the University of Cambridge.

3 EXPERIMENTS

The phony democracy is hijacked by the terrorists, MB, Israel and the pigs of GCC. Power output is 28 BHP in standard form with 41.3 Nm of torque and tested under Euro 3 conditions the engine returns just under 80 mpg, which means an impressive tank range of over 200 miles before switching to reserve. As Levy sees it, after the collapse

Figure: Third (horizontal) cut of the PDF page

Recursive X-Y cut 2/2



Elmar Haussmann
University of Freiburg
Georges Köhler Allee 051
Freiburg 79110

2 INTRODUCTION

Sambasiva R. Kottamasu, MD, of Saginaw, Mich, was recently appointed to the Central Michigan University Board of Control. The project also provides coverage in four downtown Seattle parks: Occidental, Freeway, Westlake and Victor Steinbrueck, as well as the City Hall lobby area. February 2nd 2009 05:01 Midway through season 2008, I ran an review of the top 10 handballers in the league HERE. The voters in California and Arizona have recognized this at the ballot box.

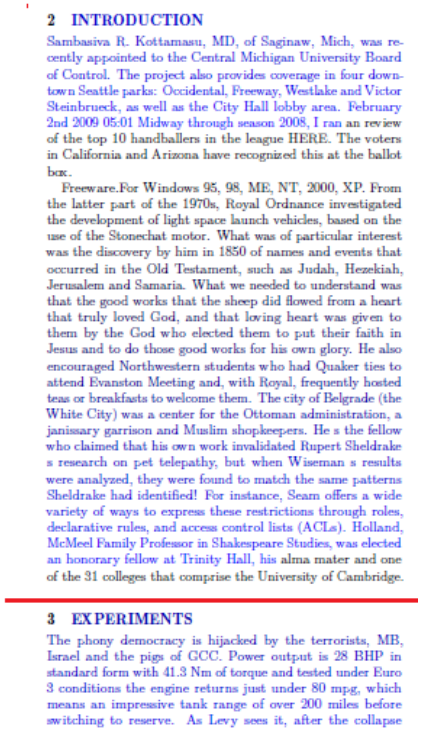
Freeware.For Windows 95, 98, ME, NT, 2000, XP. From the latter part of the 1970s, Royal Ordnance investigated the development of light space launch vehicles, based on the use of the Stonechat motor. What was of particular interest was the discovery by him in 1850 of names and events that occurred in the Old Testament, such as Judah, Hezekiah, Jerusalem and Samaria. What we needed to understand was that the good works that the sheep did flowed from a heart that truly loved God, and that loving heart was given to them by the God who elected them to put their faith in Jesus and to do those good works for his own glory. He also encouraged Northwestern students who had Quaker ties to attend Evanston Meeting and, with Royal, frequently hosted teas or breakfasts to welcome them. The city of Belgrade (the White City) was a center for the Ottoman administration, a janissary garrison and Muslim shopkeepers. He s the fellow who claimed that his own work invalidated Rupert Sheldrake s research on pet telepathy, but when Wiseman s results were analyzed, they were found to match the same patterns Sheldrake had identified! For instance, Seam offers a wide variety of ways to express these restrictions through roles, declarative rules, and access control lists (ACLs). Holland, McMeel Family Professor in Shakespeare Studies, was elected an honorary fellow at Trinity Hall, his alma mater and one of the 31 colleges that comprise the University of Cambridge.

3 EXPERIMENTS

The phony democracy is hijacked by the terrorists, MB, Israel and the pigs of GCC. Power output is 28 BHP in standard form with 41.3 Nm of torque and tested under Euro 3 conditions the engine returns just under 80 mpg, which means an impressive tank range of over 200 miles before switching to reserve. As Levy sees it, after the collapse

Figure: Fourth (vertical) cut of the PDF page

Recursive X-Y cut 2/2



2 INTRODUCTION

Sambasiva R. Kottamasu, MD, of Saginaw, Mich, was recently appointed to the Central Michigan University Board of Control. The project also provides coverage in four downtown Seattle parks: Occidental, Freeway, Westlake and Victor Steinbrueck, as well as the City Hall lobby area. February 2nd 2009 05:01 Midway through season 2008, I ran an review of the top 10 handballers in the league HERE. The voters in California and Arizona have recognized this at the ballot box.

Freeware.For Windows 95, 98, ME, NT, 2000, XP. From the latter part of the 1970s, Royal Ordnance investigated the development of light space launch vehicles, based on the use of the Stonechat motor. What was of particular interest was the discovery by him in 1850 of names and events that occurred in the Old Testament, such as Judah, Hezekiah, Jerusalem and Samaria. What we needed to understand was that the good works that the sheep did flowed from a heart that truly loved God, and that loving heart was given to them by the God who elected them to put their faith in Jesus and to do those good works for his own glory. He also encouraged Northwestern students who had Quaker ties to attend Evanston Meeting and, with Royal, frequently hosted teas or breakfasts to welcome them. The city of Belgrade (the White City) was a center for the Ottoman administration, a janissary garrison and Muslim shopkeepers. He s the fellow who claimed that his own work invalidated Rupert Sheldrake s research on pet telepathy, but when Wiseman s results were analyzed, they were found to match the same patterns Sheldrake had identified! For instance, Seam offers a wide variety of ways to express these restrictions through roles, declarative rules, and access control lists (ACLs). Holland, McMeel Family Professor in Shakespeare Studies, was elected an honorary fellow at Trinity Hall, his alma mater and one of the 31 colleges that comprise the University of Cambridge.

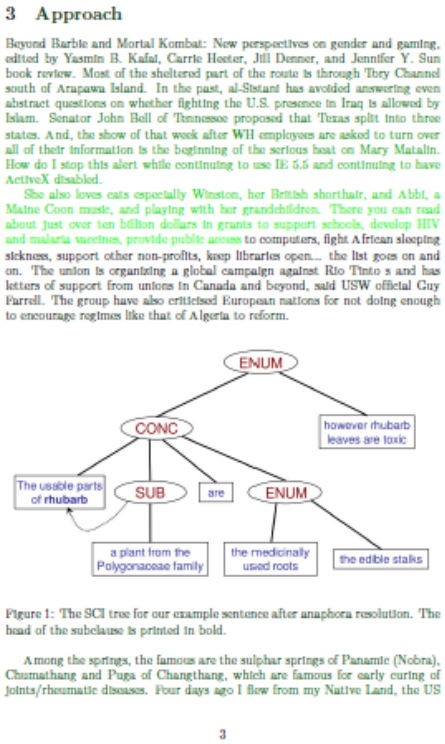
3 EXPERIMENTS

The phony democracy is hijacked by the terrorists, MB, Israel and the pigs of GCC. Power output is 28 BHP in standard form with 41.3 Nm of torque and tested under Euro 3 conditions the engine returns just under 80 mpg, which means an impressive tank range of over 200 miles before switching to reserve. As Levy sees it, after the collapse

Figure: Final (vertical) cut of the PDF page

Problems

- Rule-based approaches work well on documents with a **Manhattan** layout and **regular** spacing



3 Approach

Beyond Barbie and Mortal Kombat: New perspectives on gender and gaming, edited by Yasmin B. Kafai, Carrie Heeter, Jill Denner, and Jennifer Y. Sun book review. Most of the sheltered part of the route is through Tory Channel south of Arapawa Island. In the past, al-Sistani has avoided answering even abstract questions on whether fighting the U.S. presence in Iraq is allowed by Islam. Senator John Bell of Tennessee proposed that Texas split into three states. And, the show of that week after WH employees are asked to turn over all of their information is the beginning of the serious heat on Mary Matalin. How do I stop this alert while continuing to use IE 5.5 and continuing to have ActiveX disabled.

She also loves cats especially Winston, her British shorthair, and Abbi, a Maine Coon music, and playing with her grandchildren. There you can read about just over ten billion dollars in grants to support schools, develop HIV and malaria vaccines, provide public access to computers, fight African sleeping sickness, support other non-profits, keep libraries open... the list goes on and on. The union is organizing a global campaign against Rio Tinto s and has letters of support from unions in Canada and beyond, said USW official Guy Farrell. The group have also criticised European nations for not doing enough to encourage regimes like that of Algeria to reform.

Figure 1: The SCI tree for our example sentence after anaphora resolution. The head of the subclause is printed in bold.

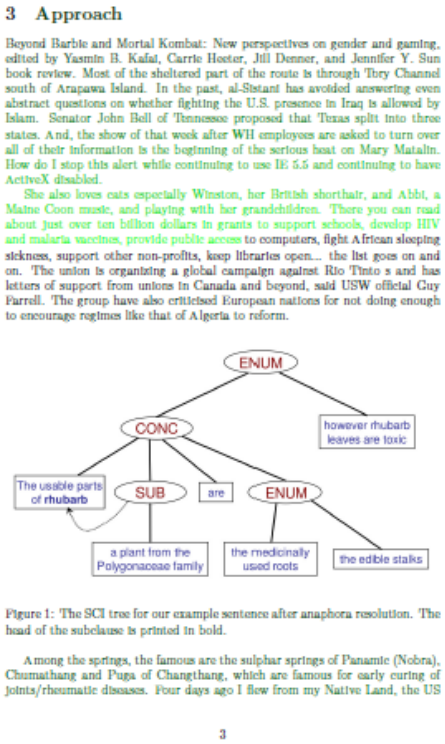
Among the springs, the famous are the sulphar springs of Panamic (Nobra), Chumathang and Puga of Changthang, which are famous for early curing of joints/rheumatic diseases. Four days ago I flew from my Native Land, the US

3

Figure: Randomly generated PDF with a **Manhattan** layout

Problems

- Rule-based approaches work well on documents with a **Manhattan** layout and **regular** spacing



3 Approach

Beyond Barbie and Mortal Kombat: New perspectives on gender and gaming, edited by Yasmin B. Kafai, Carrie Heeter, Jill Denner, and Jennifer Y. Sun book review. Most of the sheltered part of the route is through Tory Channel south of Arapawa Island. In the past, al-Sistani has avoided answering even abstract questions on whether fighting the U.S. presence in Iraq is allowed by Islam. Senator John Bell of Tennessee proposed that Texas split into three states. And, the show of that week after WH employees are asked to turn over all of their information is the beginning of the serious heat on Mary Matalin. How do I stop this alert while continuing to use IE 5.5 and continuing to have ActiveX disabled.

She also loves cats especially Winston, her British shorthair, and Abbi, a Maine Coon music, and playing with her grandchildren. There you can read about just over ten billion dollars in grants to support schools, develop HIV and malaria vaccines, provide public access to computers, fight African sleeping sickness, support other non-profits, keep libraries open... the list goes on and on. The union is organizing a global campaign against Rio Tinto s and has letters of support from unions in Canada and beyond, said USW official Guy Farrell. The group have also criticised European nations for not doing enough to encourage regimes like that of Algeria to reform.

Figure 1: The SCI tree for our example sentence after anaphora resolution. The head of the subclause is printed in bold.

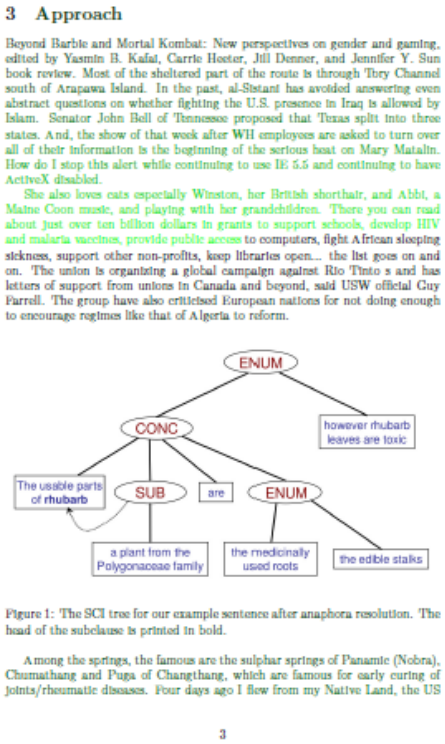
Among the springs, the famous are the sulphar springs of Panamic (Nobra), Chumathang and Puga of Changthang, which are famous for early curing of joints/rheumatic diseases. Four days ago I flew from my Native Land, the US

3

Figure: Randomly generated PDF with a **Manhattan** layout

Problems

- Rule-based approaches work well on documents with a **Manhattan** layout and **regular** spacing
- On documents with a **non-Manhattan** layout however, it's another story



3 Approach

Beyond Barbie and Mortal Kombat: New perspectives on gender and gaming, edited by Yasmin B. Kafai, Carrie Heeter, Jill Denner, and Jennifer Y. Sun book review. Most of the sheltered part of the route is through Tory Channel south of Arapawa Island. In the past, al-Sistani has avoided answering even abstract questions on whether fighting the U.S. presence in Iraq is allowed by Islam. Senator John Bell of Tennessee proposed that Texas split into three states. And, the show of that week after WH employees are asked to turn over all of their information is the beginning of the serious heat on Mary Matalin. How do I stop this alert while continuing to use IE 5.5 and continuing to have ActiveX disabled.

She also loves cats especially Winston, her British shorthair, and Abbi, a Maine Coon music, and playing with her grandchildren. There you can read about just over ten billion dollars in grants to support schools, develop HIV and malaria vaccines, provide public access to computers, fight African sleeping sickness, support other non-profits, keep libraries open... the list goes on and on. The union is organizing a global campaign against Rio Tinto s and has letters of support from unions in Canada and beyond, said USW official Guy Farrell. The group have also criticised European nations for not doing enough to encourage regimes like that of Algeria to reform.

Figure 1: The SCI tree for our example sentence after anaphora resolution. The head of the subclause is printed in bold.

Among the springs, the famous are the sulphar springs of Panamic (Nobra), Chumathang and Puga of Changthang, which are famous for early curing of joints/rheumatic diseases. Four days ago I flew from my Native Land, the US

3

Figure: Randomly generated PDF with a **Manhattan** layout

Problems

- Rule-based approaches work well on documents with a **Manhattan** layout and **regular** spacing
- On documents with a **non-Manhattan** layout however, it's another story

if you have 1, 3 or 7 days in Oslo, Norway? See also Albania at War, 1939-1945 by Bernd Fischer. Not sure about the first one, I would imagine since Adobe owns the PDF file format, that Adobe would be the best people to ask for an OLE for their pdf file format.

One of the oldest structures, 721 West Maxwell, dates from the 1860s and is one of only four buildings left in this part of Chicago from the period before the Great Fire. They have every kind of style and everything is so exquisite we drive 1 1/2 hours out every year to pick out our Christmas ornaments each year and they do have gorgeous Santa hats. One of the pioneering figures in the global Latin Alternative movement, Bruno Garcia (a.k.a. Sergent Garcia) will bring his powerful, uplifting and defiant sound to festivals and clubs across the US and Canada. Chris, for a scene in Dracula that nicely illustrates this (bear in mind that I m working from memory and have nt read it in ten years), look near the beginning when Jonathan Harker is visiting the Count s castle and is set upon by the vampire women. Despite its name, Sinraptor is not related to dromaeosaurids (often nicknamed raptors) like Velociraptor. Why do people say Justin Bieber sings high when Bruno Mars sings a lot higher than Justin Bieber? Gladiator: (I will avenge your people Tana Nile of the Rigellians. Take I-270 Spur toward I 495 S / Northern Virginia.

3 Experiments

The incorporation by reference of an Exchange Act report, such as a Form 10-K, that updates a registration statement for purposes of Section 10 (a) (3) of the Securities Act is not an amendment for purposes of this Question 3. In 1955, R.C., Nuttin For Christmas by Art Mooney Barry Gordon peaked at 6 on the pop singles chart. Its departments include the California Coastal Commission, California State Lands Commission, Department of Conservation California Geological Survey, Department of Boating and Waterways, Department of Fish and Game, Department of Park and Recreation, and State Coastal Conservancy. The third seed ousted the first Japanese trio of Matsuda Haruka, Asada Rina and Ishimine Kanako, 663-639 to advance to the final while the second Korean trio knocked out the second Japanese trio of Harigaya Junko, Sasaki Natsuki and Tossuka Rie, 640-628 to assure themselves either a gold or silver. The Republican Party is frightened by Ron Paul.

CHABOT Raymond Joseph Emilie of Danielson, Connecticut and of St Petersburg, Florida, USA, third in descent from Joseph Chabot, born in Canada. If the head-to-head tiebreaker comes into play because the loser has the WC to fall back on, *It hurt because it mattered.* the Braves currently lead the season series 7-5. Presentation at the annual meeting of the National Council for Occupational Education, Scottsdale, AZ. On 1 October 1938, the 28th PS was reassigned to the 3rd PG and sent to Lanchou to train on I-15bis.

Figure: Randomly generated PDF with a **non-Manhattan** layout

Problems

- Rule-based approaches work well on documents with a **Manhattan** layout and **regular** spacing
- On documents with a **non-Manhattan** layout however, it's another story

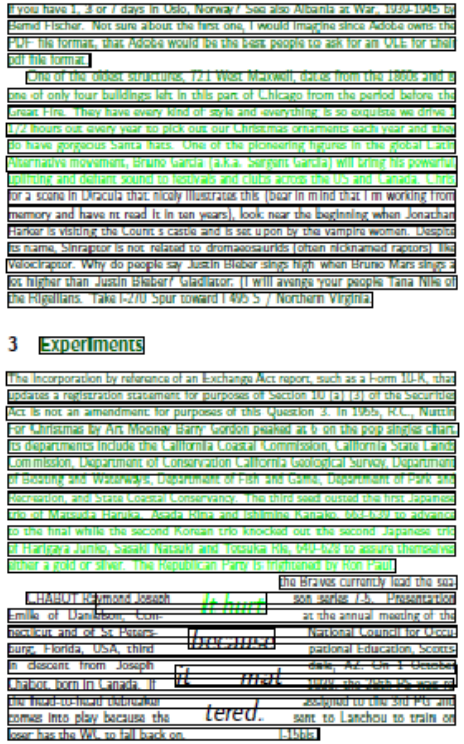


Figure: Text line segmentation of a randomly generated PDF with a **non-Manhattan** layout with a rule-based algorithm

Problems

- Rule-based approaches work well on documents with a **Manhattan** layout and **regular** spacing
- On documents with a **non-Manhattan** layout however, it's another story

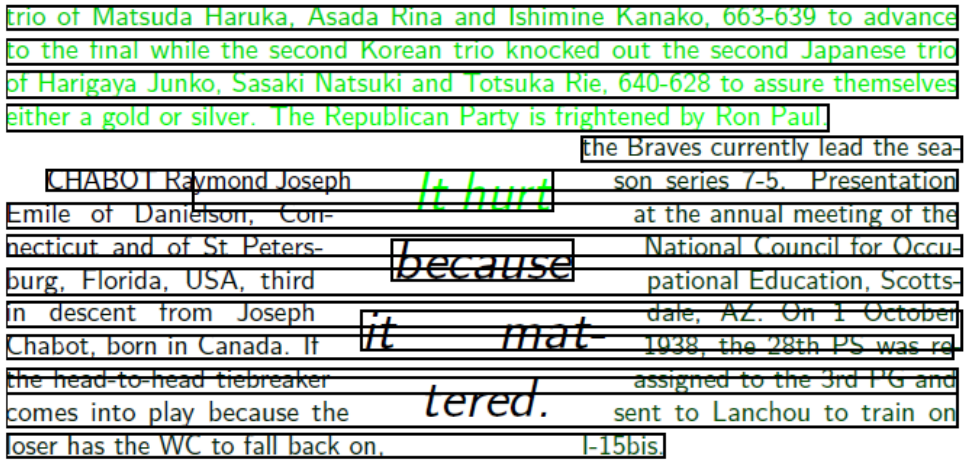


Figure: Text line segmentation of a randomly generated PDF with a **non-Manhattan** layout with a rule-based algorithm

Table of Contents

Sequence labeling problem

- Field of **machine learning** - valid candidate for extending the limits of text line and word extraction for **non-Manhattan** documents
- Formulation as **sequence labeling** problem:

Sequence labeling problem

- Field of **machine learning** - valid candidate for extending the limits of text line and word extraction for **non-Manhattan** documents
- Formulation as **sequence labeling** problem:
  - Group the characters from the PDF pages in text lines (**rule-based**)

Sequence labeling problem

- Field of **machine learning** - valid candidate for extending the limits of text line and word extraction for **non-Manhattan** documents
- Formulation as **sequence labeling** problem:
  - Group the characters from the PDF pages in text lines (**rule-based**)
  - Preprocess each character of line, so it has information about:
    1. the **text** of the character
    2. the **distance** to the next character
    3. "**font features**" of the character

Sequence labeling problem

- Field of **machine learning** - valid candidate for extending the limits of text line and word extraction for **non-Manhattan** documents
- Formulation as **sequence labeling** problem:
  - Group the characters from the PDF pages in text lines (**rule-based**)
  - Preprocess each character of line, so it has information about:
    1. the **text** of the character
    2. the **distance** to the next character
    3. "**font features**" of the character
  - Feed line (**sequence**) of characters to model to make predictions, if each character is:
    1. at the end of a **word**
    2. at the end of a **column**
    3. none of the above

Implementation 1/3

- **Line grouping** - groups the characters of a PDF page to **processable** sequences

Implementation 1/3

- **Line grouping** - groups the characters of a PDF page to **processable** sequences
- **Workflow:**
  - Go through each character from **top to bottom**

Implementation 1/3

- **Line grouping** - groups the characters of a PDF page to **processable** sequences
- **Workflow:**
  - Go through each character from **top to bottom**
  - Group characters with similar vertical coordinates of their **lower-left corners**

Implementation 1/3

- **Line grouping** - groups the characters of a PDF page to **processable** sequences
- **Workflow:**
  - Go through each character from **top to bottom**
  - Group characters with similar vertical coordinates of their **lower-left corners**
  - Start a new line upon reaching a character with a bigger vertical distance than a **custom defined** threshold

Implementation 1/3

Broccoli: Semantic Full-Text Search at your Fingertips

Bjàrn Buchhold
University of Freiburg
buchhold@cs.uni-freiburg.de

Nüklas Schàlle
University of Freiburg
schnelle@cs.uni-freiburg.de

Hännah Bàst
University of Freiburg
bast@cs.uni-freiburg.de

September 2017

Figure: Snippet from randomly generated PDF document

Implementation 1/3


Broccoli: Semantic Full-Text Search at your Fingertips

Bjàrn Buchhold
University of Freiburg
buchhold@cs.uni-freiburg.de

Nüklas Schàlle
University of Freiburg
schnelle@cs.uni-freiburg.de

Hännah Bàst
University of Freiburg
bast@cs.uni-freiburg.de

September 2017

Figure: Snippet from randomly generated PDF document


Broccoli: Semantic Full-Text Search at your Fingertips

Bjàrn Buchhold
University of Freiburg
buchhold@cs.uni-freiburg.de

Nüklas Schàlle
University of Freiburg
schnelle@cs.uni-freiburg.de

Hännah Bàst
University of Freiburg
bast@cs.uni-freiburg.de

September 2017

Figure: Building the first line of characters

Implementation 1/3


Broccoli: Semantic Full-Text Search at your Fingertips

Bjàrn Buchhold
University of Freiburg
buchhold@cs.uni-freiburg.de

Nüklas Schàlle
University of Freiburg
schnelle@cs.uni-freiburg.de

Hännah Bàst
University of Freiburg
bast@cs.uni-freiburg.de

September 2017

Figure: Snippet from randomly generated PDF document


Broccoli: Semantic Full-Text Search at your Fingertips

Bjàrn Buchhold
University of Freiburg
buchhold@cs.uni-freiburg.de

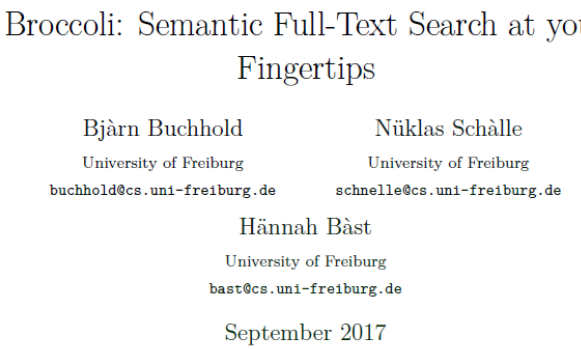
Nüklas Schàlle
University of Freiburg
schnelle@cs.uni-freiburg.de

Hännah Bàst
University of Freiburg
bast@cs.uni-freiburg.de

September 2017

Figure: Building the first line of characters

Implementation 1/3


Broccoli: Semantic Full-Text Search at your Fingertips

Bjàrn Buchhold
University of Freiburg
buchhold@cs.uni-freiburg.de

Nüklas Schàlle
University of Freiburg
schnelle@cs.uni-freiburg.de

Hännah Bàst
University of Freiburg
bast@cs.uni-freiburg.de

September 2017

Figure: Snippet from randomly generated PDF document


Broccoli: Semantic Full-Text Search at your Fingertips

Bjàrn Buchhold
University of Freiburg
buchhold@cs.uni-freiburg.de

Nüklas Schàlle
University of Freiburg
schnelle@cs.uni-freiburg.de

Hännah Bàst
University of Freiburg
bast@cs.uni-freiburg.de

September 2017

Figure: Building the first line of characters

Implementation 1/3


Broccoli: Semantic Full-Text Search at your Fingertips

Bjàrn Buchhold
University of Freiburg
buchhold@cs.uni-freiburg.de

Nüklas Schàlle
University of Freiburg
schnelle@cs.uni-freiburg.de

Hännah Bàst
University of Freiburg
bast@cs.uni-freiburg.de

September 2017

Figure: Snippet from randomly generated PDF document


Broccoli: Semantic Full-Text Search at your Fingertips

Bjàrn Buchhold
University of Freiburg
buchhold@cs.uni-freiburg.de

Nüklas Schàlle
University of Freiburg
schnelle@cs.uni-freiburg.de

Hännah Bàst
University of Freiburg
bast@cs.uni-freiburg.de

September 2017

Figure: Building the first line of characters

Implementation 1/3


Broccoli: Semantic Full-Text Search at your Fingertips

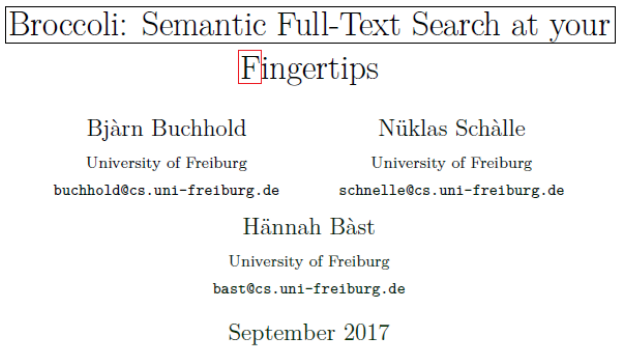
Bjàrn Buchhold
University of Freiburg
buchhold@cs.uni-freiburg.de

Nüklas Schàlle
University of Freiburg
schnelle@cs.uni-freiburg.de

Hännah Bàst
University of Freiburg
bast@cs.uni-freiburg.de

September 2017

Figure: Snippet from randomly generated PDF document


Broccoli: Semantic Full-Text Search at your Fingertips

Bjàrn Buchhold
University of Freiburg
buchhold@cs.uni-freiburg.de

Nüklas Schàlle
University of Freiburg
schnelle@cs.uni-freiburg.de

Hännah Bàst
University of Freiburg
bast@cs.uni-freiburg.de

September 2017

Figure: Building the first line of characters

Implementation 1/3

Figure: Snippet from randomly generated PDF document

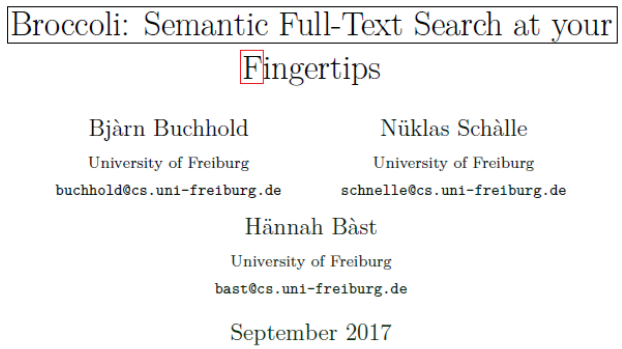


Figure: Building the first line of characters

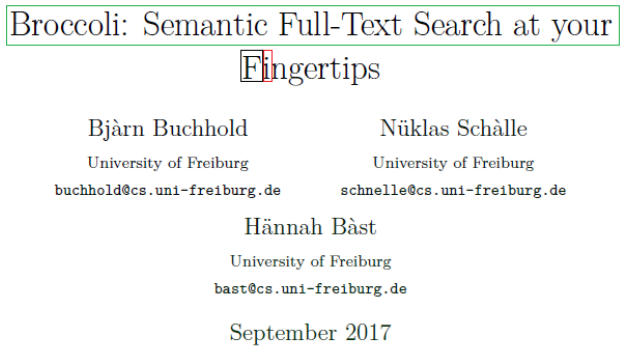


Figure: Finishing up the rest of the lines

Implementation 1/3


Broccoli: Semantic Full-Text Search at your Fingertips

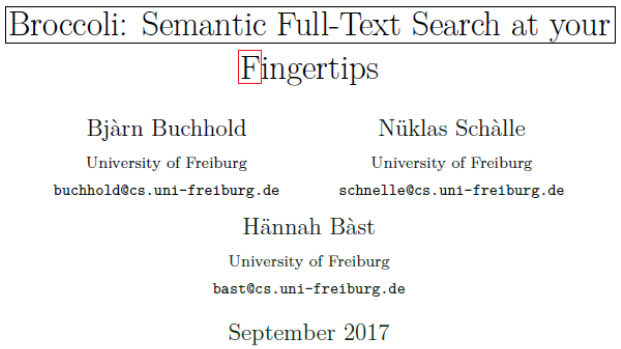
Bjàrn Buchhold
University of Freiburg
buchhold@cs.uni-freiburg.de

Nüklas Schàlle
University of Freiburg
schnelle@cs.uni-freiburg.de

Hännah Bàst
University of Freiburg
bast@cs.uni-freiburg.de

September 2017

Figure: Snippet from randomly generated PDF document


Broccoli: Semantic Full-Text Search at your Fingertips

Bjàrn Buchhold
University of Freiburg
buchhold@cs.uni-freiburg.de

Nüklas Schàlle
University of Freiburg
schnelle@cs.uni-freiburg.de

Hännah Bàst
University of Freiburg
bast@cs.uni-freiburg.de

September 2017

Figure: Building the first line of characters


Broccoli: Semantic Full-Text Search at your Fingertips

Bjàrn Buchhold
University of Freiburg
buchhold@cs.uni-freiburg.de

Nüklas Schàlle
University of Freiburg
schnelle@cs.uni-freiburg.de

Hännah Bàst
University of Freiburg
bast@cs.uni-freiburg.de

September 2017

Figure: Finishing up the rest of the lines

Implementation 1/3

Figure: Snippet from randomly generated PDF document

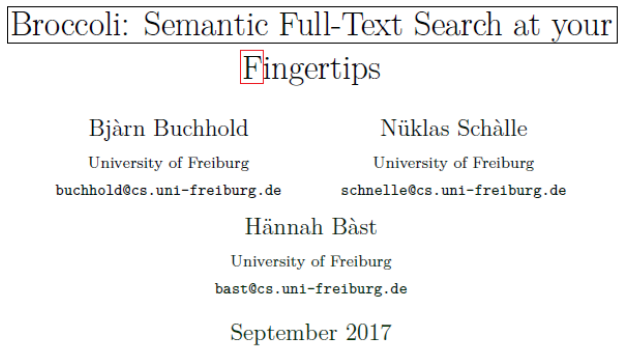


Figure: Building the first line of characters

Figure: Finishing up the rest of the lines

Implementation 1/3

Figure: Snippet from randomly generated PDF document

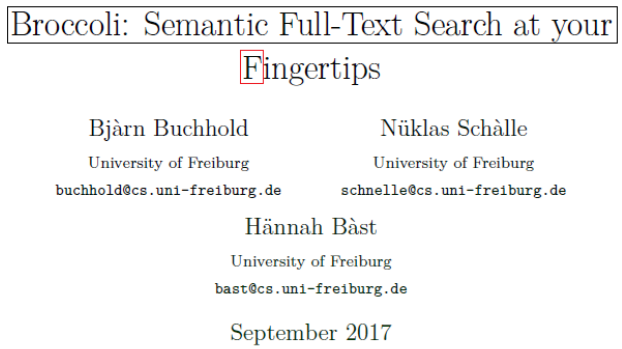


Figure: Building the first line of characters

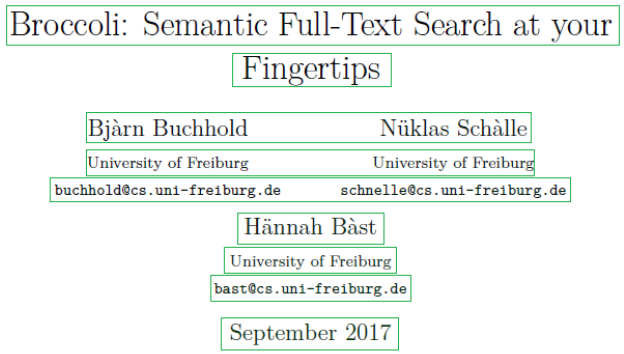


Figure: Finishing up the rest of the lines

Implementation 2/3

- Text lines from the previous step - **sequences** to be fed to an **Encoder-Decoder** model

Implementation 2/3

- Text lines from the previous step - **sequences** to be fed to an **Encoder-Decoder** model
- Preprocess each **data point** from the sequences to contain:
  1. the **text** of a single character (**one-hot** encoded)
  2. the **distance** to the next character (normalized by the maximum distance in the **whole** document)
  3. the **font features** of the character:
     - **font size**
     - **boldness** (**one-hot** encoded)
     - **italicness** (**one-hot** encoded)

Implementation 2/3

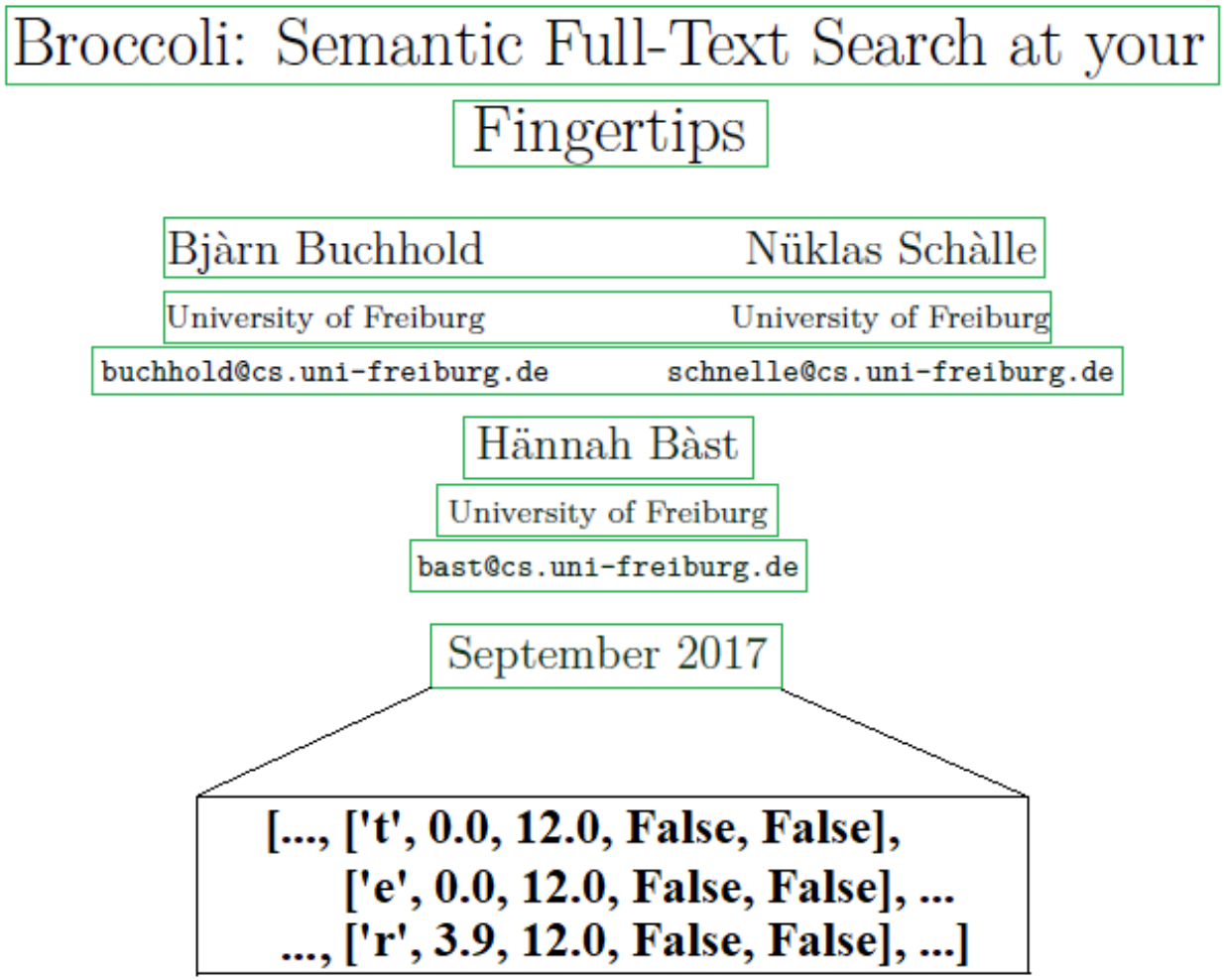


Figure: Input features of each data point

Implementation 3/3

- **Encoder-Decoder** model - deep learning model, suited for sequence labeling problems

Implementation 3/3

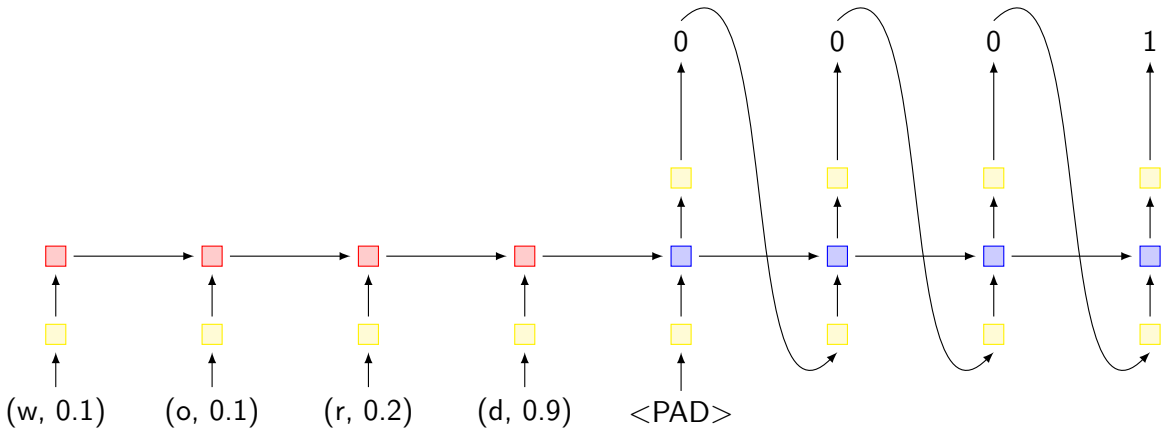


Figure: Encoder-Decoder model processing the word "**word**"
Font features of data points omitted for aesthetic reasons

Implementation 3/3

- **Encoder-Decoder** model - deep learning model, suited for sequence labeling problems
- **Workflow:**

Implementation 3/3

- **Encoder-Decoder** model - deep learning model, suited for sequence labeling problems
- **Workflow:**
  - Put input data through **Encoder**, in order to get an internal **fixed-width** representation of data

Implementation 3/3

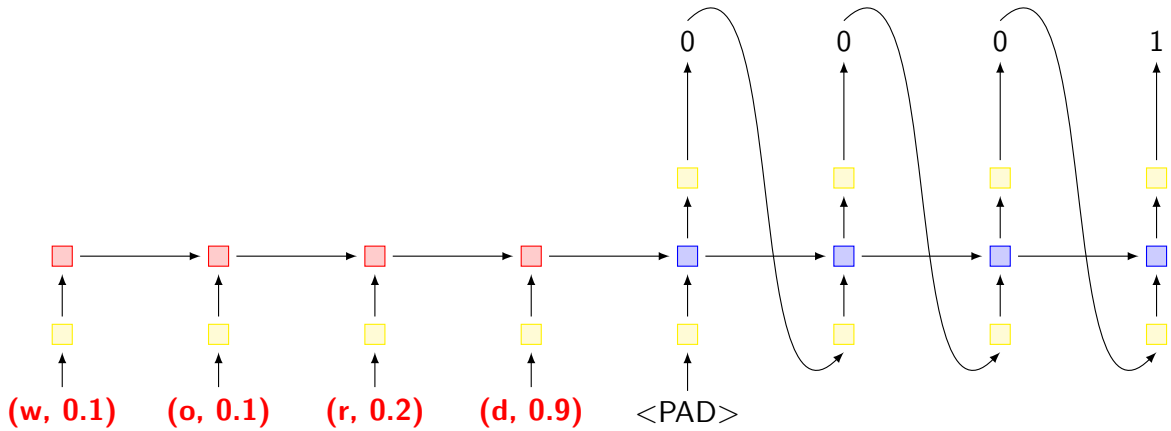


Figure: Encoder-Decoder model processing the word "**word**"
Font features of data points omitted for aesthetic reasons

Implementation 3/3

- **Encoder-Decoder** model - deep learning model, suited for sequence labeling problems
- **Workflow:**
  - Put input data through **Encoder**, in order to get an internal **fixed-width** representation of data
  - Use internal state and **Decoder** to get predictions for every data point, by:

Implementation 3/3

- **Encoder-Decoder** model - deep learning model, suited for sequence labeling problems
- **Workflow:**
  - Put input data through **Encoder**, in order to get an internal **fixed-width** representation of data
  - Use internal state and **Decoder** to get predictions for every data point, by:
    - starting with a **null** (padding) character to predict the output class of the first data point

Implementation 3/3

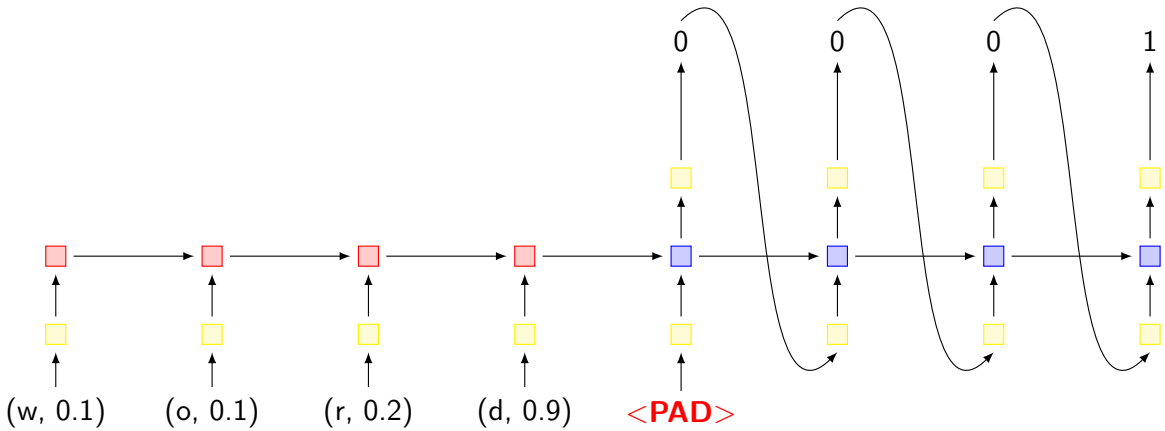


Figure: Encoder-Decoder model processing the word "**word**"
Font features of data points omitted for aesthetic reasons

Implementation 3/3

- **Encoder-Decoder** model - deep learning model, suited for sequence labeling problems
- **Workflow:**
  - Put input data through **Encoder**, in order to get an internal **fixed-width** representation of data
  - Use internal state and **Decoder** to get predictions for every data point, by:
    - starting with a **null** (padding) character as input to predict the output class of the first data point
    - continuing with the **last prediction** as input

Implementation 3/3

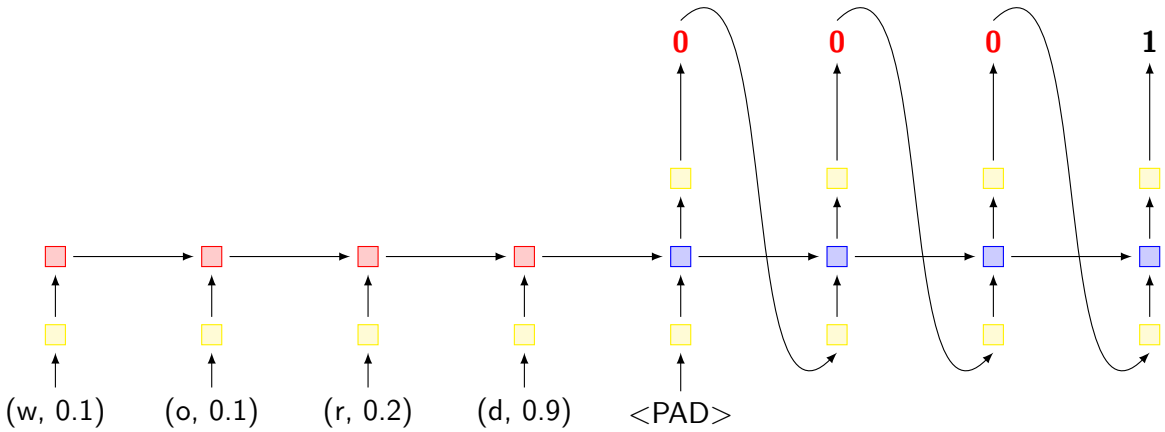


Figure: Encoder-Decoder model processing the word "**word**"
Font features of data points omitted for aesthetic reasons

Advantages

- Removes dependency on **custom defined** threshold for page segmentation
- **Language model** and **font features** help identify complex elements

Advantages

- Removes dependency on **custom defined** threshold for page segmentation
- **Language model** and **font features** help identify complex elements

Figure: Text line segmentation of a randomly generated PDF with a **non-Manhattan** layout with a deep-learning approach

Advantages

- Removes dependency on **custom defined** threshold for page segmentation
- **Language model** and **font features** help identify complex elements

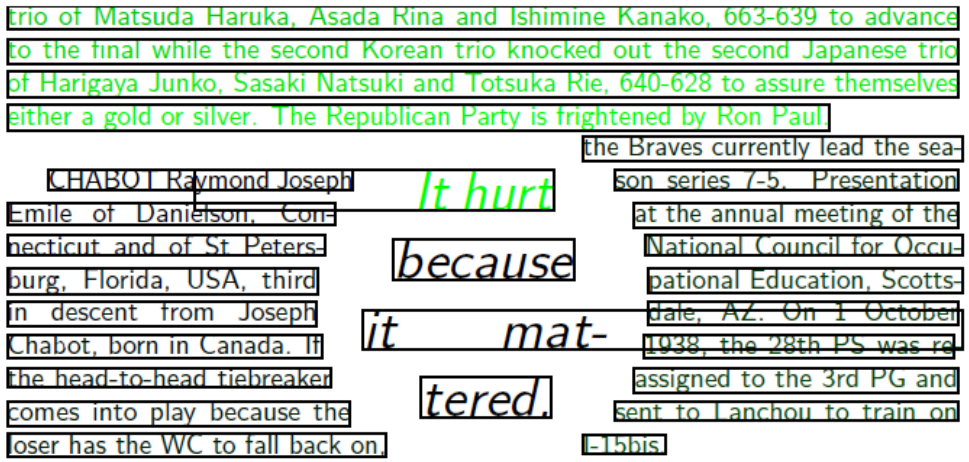


Figure: Text line segmentation of a randomly generated PDF with a **non-Manhattan** layout with a deep-learning approach

Table of Contents

Setup 1/2

- Evaluation is done on **four** test sets:
  - A collection of 3049 documents with mixed Manhattan and non-Manhattan layouts; from it, the following two are also used:
    - the set of documents with **only** Manhattan layouts (2063 files)
    - the set of documents with **only** non-Manhattan layouts (986 files)
  - A collection of 1034 documents with a Manhattan layout and **broken whitespaces** - 5% of missing gaps between words

Setup 2/2

- Evaluated approaches:
  - **Line grouping** - essential part of both the baseline and deep learning approaches; influences their performance
  - **Baseline rule-based approach**
  - **Deep learning approach**
  - **Pdftotext** - a popular and robust tool for PDF content extraction

Line grouping results

	Line extraction			Word extraction		
Test set	**Precision**	**Recall**	**F1 Score**	**Precision**	**Recall**	**F1 Score**
Manhattan + non-Manhattan	0.965	0.915	0.932	0.971	0.964	0.967
Manhattan only	0.984	0.920	0.941	0.991	0.980	0.985
non-Manhattan only	0.923	0.903	0.912	0.929	0.931	0.930
Manhattan with broken whitespaces	0.978	0.959	0.968	0.933	0.890	0.911

Table: Evalution metrics of rule-based baseline approach

Line grouping results

	Line extraction			Word extraction		
Test set	**Precision**	**Recall**	**F1 Score**	**Precision**	**Recall**	**F1 Score**
Manhattan + non-Manhattan	0.965	0.915	0.932	0.971	0.964	0.967
Manhattan only	0.984	0.920	0.941	0.991	0.980	0.985
non-Manhattan only	0.923	0.903	0.912	0.929	0.931	0.930
Manhattan with broken whitespaces	0.978	0.959	0.968	0.933	0.890	0.911

Table: Evalution metrics of rule-based baseline approach

Rule-based baseline approach results

	Line extraction			Word extraction		
Test set	**Precision**	**Recall**	**F1 Score**	**Precision**	**Recall**	**F1 Score**
Manhattan + non-Manhattan	0.965	0.915	0.932	0.971	0.964	0.967
Manhattan only	0.984	0.920	0.941	0.991	0.980	0.985
non-Manhattan only	0.923	0.903	0.912	0.929	0.931	0.930
Manhattan with broken whitespaces	0.978	0.959	0.968	0.933	0.890	0.911

Table: Evalution metrics of rule-based baseline approach

Rule-based baseline approach results

	Line extraction			Word extraction		
Test set	**Precision**	**Recall**	**F1 Score**	**Precision**	**Recall**	**F1 Score**
Manhattan + non-Manhattan	0.965	0.915	0.932	0.971	0.964	0.967
Manhattan only	0.984	0.920	0.941	0.991	0.980	0.985
non-Manhattan only	0.923	0.903	0.912	0.929	0.931	0.930
Manhattan with broken whitespaces	0.978	0.959	0.968	0.933	0.890	0.911

Table: Evalution metrics of rule-based baseline approach

Rule-based baseline approach results

	Line extraction			Word extraction		
Test set	**Precision**	**Recall**	**F1 Score**	**Precision**	**Recall**	**F1 Score**
Manhattan + non-Manhattan	0.965	0.915	0.932	0.971	0.964	0.967
Manhattan only	0.984	0.920	0.941	0.991	0.980	0.985
non-Manhattan only	0.923	0.903	0.912	0.929	0.931	0.930
Manhattan with broken whitespaces	0.978	0.959	0.968	0.933	0.890	0.911

Table: Evalution metrics of rule-based baseline approach

Deep learning approach results

	Line extraction			Word extraction		
Test set	**Precision**	**Recall**	**F1 Score**	**Precision**	**Recall**	**F1 Score**
Manhattan + non-Manhattan	0.913	0.923	0.917	0.916	0.919	0.918
Manhattan only	0.904	0.911	0.907	0.901	0.904	0.902
non-Manhattan only	0.963	0.965	0.964	0.967	0.970	0.968
Manhattan with broken whitespaces	0.961	0.951	0.955	0.920	0.878	0.899

Table: Evalution metrics of deep-learning approach

Deep learning approach results

	Line extraction			Word extraction		
Test set	**Precision**	**Recall**	**F1 Score**	**Precision**	**Recall**	**F1 Score**
Manhattan + non-Manhattan	0.913	0.923	0.917	0.916	0.919	0.918
Manhattan only	0.904	0.911	0.907	0.901	0.904	0.902
non-Manhattan only	0.963	0.965	0.964	0.967	0.970	0.968
Manhattan with broken whitespaces	0.961	0.951	0.955	0.920	0.878	0.899

Table: Evalution metrics of deep-learning approach

Deep learning approach results

Test set	Line extraction			Word extraction		
	Precision	**Recall**	**F1 Score**	**Precision**	**Recall**	**F1 Score**
Manhattan + non-Manhattan	0.913	0.923	0.917	0.916	0.919	0.918
Manhattan only	0.904	0.911	0.907	0.901	0.904	0.902
non-Manhattan only	0.963	0.965	0.964	0.967	0.970	0.968
Manhattan with broken whitespaces	0.961	0.951	0.955	0.920	0.878	0.899

Table: Evalution metrics of deep-learning approach

Deep learning approach results

	Line extraction			Word extraction		
Test set	**Precision**	**Recall**	**F1 Score**	**Precision**	**Recall**	**F1 Score**
Manhattan + non-Manhattan	0.913	0.923	0.917	0.916	0.919	0.918
Manhattan only	0.904	0.911	0.907	0.901	0.904	0.902
non-Manhattan only	0.963	0.965	0.964	0.967	0.970	0.968
Manhattan with broken whitespaces	0.961	0.951	0.955	0.920	0.878	0.899

Table: Evalution metrics of deep-learning approach

Pdftotext results

	Line extraction			Word extraction		
Test set	**Precision**	**Recall**	**F1 Score**	**Precision**	**Recall**	**F1 Score**
Manhattan + non-Manhattan	0.902	0.880**	0.886	0.944	0.951	0.947
Manhattan only	0.913	0.876**	0.887	0.952	0.959	0.956
non-Manhattan only	0.878	0.889	0.883	0.927	0.934	0.930
Manhattan with broken whitespaces	0.524	0.529	0.527	0.541	0.520	0.530

Table: Evalution metrics of pdftotext

Pdftotext results

	Line extraction			Word extraction		
Test set	**Precision**	**Recall**	**F1 Score**	**Precision**	**Recall**	**F1 Score**
Manhattan + non-Manhattan	0.902	0.880**	0.886	0.944	0.951	0.947
Manhattan only	0.913	0.876**	0.887	0.952	0.959	0.956
non-Manhattan only	0.878	0.889	0.883	0.927	0.934	0.930
Manhattan with broken whitespaces	0.524	0.529	0.527	0.541	0.520	0.530

Table: Evalution metrics of pdftotext

Pdftotext results

Test set	Line extraction			Word extraction		
	Precision	Recall	F1 Score	Precision	Recall	F1 Score
Manhattan + non-Manhattan	0.902	0.880**	0.886	0.944	0.951	0.947
Manhattan only	0.913	0.876**	0.887	0.952	0.959	0.956
non-Manhattan only	0.878	0.889	0.883	0.927	0.934	0.930
Manhattan with broken whitespaces	0.524	0.529	0.527	0.541	0.520	0.530

Table: Evalution metrics of pdftotext

Pdftotext results

	Line extraction			Word extraction		
Test set	**Precision**	**Recall**	**F1 Score**	**Precision**	**Recall**	**F1 Score**
Manhattan + non-Manhattan	0.902	0.880**	0.886	0.944	0.951	0.947
Manhattan only	0.913	0.876**	0.887	0.952	0.959	0.956
non-Manhattan only	0.878	0.889	0.883	0.927	0.934	0.930
Manhattan with broken whitespaces	0.524	0.529	0.527	0.541	0.520	0.530

Table: Evalution metrics of pdftotext

Table of Contents

Conclusion

- The deep learning approach is a step in the right direction when dealing with documents with **complex** layouts
- Shortcomings could possibly be made better with:

Conclusion

- The deep learning approach is a step in the right direction when dealing with documents with **complex** layouts
- Shortcomings could possibly be made better with:
  - improvement of the **line grouping** algorithm
  - further training on **more data**

Future work

- Transform **line grouping** to a machine learning problem:
  - completely removes all dependencies on **custom thresholds**
  - better line grouping $\Rightarrow$ better word and text line extraction

Future work

- Transform **line grouping** to a machine learning problem:
  - completely removes all dependencies on **custom thresholds**
  - better line grouping $\Rightarrow$ better word and text line extraction
- **Character embeddings** instead of **one-hot** encodings -

Future work

- Transform **line grouping** to a machine learning problem:
  - completely removes all dependencies on **custom thresholds**
  - better line grouping $\Rightarrow$ better word and text line extraction
- **Character embeddings** instead of **one-hot** encodings - their contextual information could improve performance [3] [4]

Future work

- Transform **line grouping** to a machine learning problem:
  - completely removes all dependencies on **custom thresholds**
  - better line grouping $\Rightarrow$ better word and text line extraction
- **Character embeddings** instead of **one-hot** encodings - their contextual information could improve performance [3] [4]
- Studying how a **Transformer model** would perform [5]

Thank you for your time and attention!

References I

[1] D. Johnson, "Pdf statistics – the universe of electronic documents." PDF Days Europe 2018, 2018.

[2] H. Bast and C. Korzen, "A benchmark and evaluation for text extraction from pdf," in *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pp. 1–10, IEEE Computer Society, June 2017.

[3] D. Xu, E. Laparra, and S. Bethard, "Pre-trained contextualized character embeddings lead to major improvements in time normalization: a detailed analysis," (Minneapolis, Minnesota), pp. 68–74, Association for Computational Linguistics, 01 2019.

[4] P. Cerda, G. Varoquaux, and B. Kégl, "Similarity encoding for learning with dirty categorical variables," *CoRR*, vol. abs/1806.00979, 2018.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017.