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

- Problem motivation
 - PDF content extraction
 - Rule-based approaches
- Proposed solution
 - Sequence labeling problem
 - Implementation
 - Advantages
 - Besults

🕽 Final remarks

Table of Contents

- Problem motivation
 - PDF content extraction
 - Rule-based approaches
- Proposed solution
 - Sequence labeling problem
 - Implementation
 - Advantages
- B Results

Final remarks

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]

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

- 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...

- 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:

- 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...

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

- **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

8 DX 2 DRI M DNMS

Conse on to Charden to Chinall, And Caardy pass to Chi Shafii Chy Charle companiilles al Che Mayder, and conseilly resilts to Che Che Mistere anna. Als a second mariley, Als

officer s questions were more pointed, focusing on Mr. Khadr s relationship with al-Qaeda chief Osama bin Laden. Another creamy white or reddish silk called Eri is produced from larvae that feed on castor bean plants (Ricinus communis) of the Euphorbiaceae. We supply cranshshaft, cylinder block, cylinder head assembly, pt fuel pump assembly, and injector assembly suitable for all models of Cummins engines from Chemai. Cruise Cheap.com - Cruise Only travel agency that specializes in discount cruise rates to Alaska, Hawali, the Caribbean, Bahamas and Europe. At the U.N., Soviet

Figure: How textual content is stored on a character level

8 DXPDRIMDNIS

Cense og le Glandforfar di Chitteli, And Casedy gene le Gla Statik Rieg Graefi er menentiker di Che Magder, and enverdig resilte in Gla Che Michae anen. Als a er resilter, Gla

officer s questions were more pointed, focusing on Mr. Tkhadr s relationship with al-Qaeda chief Osama bin Laden. Another creamy white or reddish silk called Eri is produced from larvae that feed on castor bean plants (Ricinus communis) of the Euphorbiaceae. We supply crankslaft, cylinder block, cylinder head assembly, pt fuel pump assembly, and injector assembly suitable for all models of Cummins engines from Chemai. Cruise Cheap.com - Cruise Only travel agency that specializes in discount cruise rates to Alaska, Hawaii, the Caribbean, Bahamas and Europe. At the U.N., Soviet

Figure: How textual content is stored on a character level

EXPERIMENTS

I grew up in the suburbs of Detroit, lived twenty years in the South Bay beach communities of Los Angeles, and currently reside in the Des Moines area. At a second meeting, the officer s questions were more pointed, focusing on Mr. Khadr s relationship with al-Qaeda chief Osama bin Laden. Another creany white or reddish silk called Eri is produced from larvae that feed on castor bean plants (Ricinus communis) of the Euphorbiaceae. We supply crankshaft, cylinder block, cylinder head assembly, pt fuel pump assembly, and injector assembly suitable for all models of Cummins engines from Chennai. Cruise Cheap.com - Cruise Only travel agency that specializes in discount cruise rates to Alaska, Hawaii, the Caribbean, Bahamas and Europe. At the U.N., Soviet

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

6

6

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

- 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)

Markus Näther

Georges Köhler Allee 051

Probute 79110

University of Freiburg

Recursive X-Y cut 2/2

Efficient Generation of Geographically Accurate Transit Maps

Hannah Bast University of Freiburg Georeen Köhler Allee 051 Freihure 79110

Patrick Brosi University of Freiburg Georges Köhler Allee 051 Freiburg 79110

University of Freiburg Georges Köhler Allee 051 Freiburg 79110

Elmar Haussmann 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 a arrival in Queen Charlotte s Sound, he went himself, at daybreak, to look for scurvy-grass, celery, and other vegetables; 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

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 lastone s annual event coinciding with San Francisco & Alternative Press Expo featuring the awarding of the Isotone Award For Excellence in Mini-Comics. It has a population of 39 million people, more or less equal sey. 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 Janiper pushed RSVP at the standard when MPLS first came around. Yet Rowling a 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 down 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

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 narticular 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 Evaneton 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 ianissary garrison and Muslim shonkeeners. He s the fellow were analyzed, they were found to match the same natterns Sheldrake had identified! For instance, Seam offers a wide 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 pins 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

Markus Näther

University of Freiburg

Georges Köhler Allee 051

Probute 79110

Recursive X-Y cut 2/2

Efficient Generation of Geographically Accurate Transit Maps

Hannah Bast University of Freiburg Georges Köhler Allee 051 Freihure 79110

Patrick Brosi University of Freiburg Georges Köhler Allee 051 Freiburg 79110

University of Freiburg Georges Köhler Allee 051 Freiburg 79110

Elmar Haussmann 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; 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 150 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

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 lastone s annual event coinciding with San Francisco & Alternative Press Expo featuring the awarding of the Isotone Award For Excellence in Mini-Comics. It has a population of 39 million people, more or less equal sey. 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 Janiper pushed RSVP at the standard when MPLS first came around. Yet Rowling a 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 reof Control. The project also provides coverage in four down 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

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 narticular 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 Evaneton 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 ianissary garrison and Muslim shonkeeners. He s the fellow were analyzed, they were found to match the same natterns Sheldrake had identified! For instance, Seam offers a wide 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 pins 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

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 Freiberg 79110 Elmar Haussmann University of Freiburg Georges Köhler Allee 051 Freiburg 79110

1 ABSTRACT

In January this year. Rosselli refused to narticinate in the election of officers of the 650,000 member SEIU California State Council or to run for reelection as its remident accusing Stern of rigging the process to guarantee the success of his arrival in Oneen Charlotte a Sound, he went himself, at daybreak, to look for sourvy-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 Laurente Wangari Maathai to create sustainable change in Africa provided the same inspiration to well over 1500 women across North America. The Java Serviet and PHP session mechanisms both use this method if cookies are not enabled. When Stahroek News was at Marlhorough in the Pomeroon River on Tuesday, five families were prepari to make an eight-hour and 64-mile journey up the Akiwini

While I have done Linux and Windows 2000, I have not seen the Solaris 7 and Windows 2000 combination. Let a 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 lastone 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 nonulation of 39 million neople, more or less equal sey. In the presidential election of 1999, opposition leader Abdoulase 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 sublisher. Scholastic, had the good sense to low-key her tour. teaming with the independent booksellers association.

2 INTRODUCTION

Surbasive R. Kettamanu, MD, of Saginaw, Mch., was erouthy appointed to the Contral Michigan University Board of Control. The project also provide overage in four downtoon South parks. Coclestical, Prevaw, Westklae and Victor Statisticosek, as well as the City Hall isbly area. Formary 242 2009 06:01 Missiway through assuma 2008, I' run an review of the top 10 handhallers in the lengue HBRE. The voters is California and Artenna have recognized this at the halids

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 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 a research on net telenathy, but when Wiseman a results Sheldrake had identified! For instance, Seam offers a wide variety of ways to express these restrictions through roles. 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 Iarael and the pigs of CCC. Power output is 28 BHP in standard form with 4.1.3 Nor of teopes and tented under Euro 3 conditions the engine returns just under 80 mpg, which means an impressive tank range of over 200 miles before writching to reserve. As Levy sees it, after the collapse

Figure: Second (vertical) cut of the PDF page

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 bandnicked choices. The morning after Cantain Cook s. arrival in Oueen Charlotte a Sound, he went himself, at day, break, to look for scurvy-grass, celery, and other vegetables: and he had the good fortune to return with a boatload, in a very short snace of time. This encompassed Pell's lands 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 1900 women across North America. The Java Servict and PHP session mechanisms both use this method if cookies are not enabled. When Stabroek News was at Marlborough in the to make an eight-hour and 64-mile journey up the Akiwini

While I have done Linux and Windows 2000, I have not seen the Solaris 7 and Windows 2000 combination. Let a 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 a annual event coinciding awarding of the Isotope Award For Excellence in Mini-Comics. It has a population of 39 million people, more or less equal sey. In the presidential election of 1999, opposition leader Abdoulave 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 Ciaco pushed LDP and Juniper pushed RSVP at the standard when MPLS first came around. Yet Rowling a 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

Sanhasira R. Kottamuu, MD, of Sagiane, Wich, was recently appointed to the Contral Michigan University Board of Control. The project also provide coverage in four downtown South particle. Covidential, Ference, Wentikla and Visior Steinbrenck, as well as the City Hall lobby area. February 24 2000 0501 Molecy through second 2006, I can a neriew of College 10 hand-failed in the Isages HTML: The voture of College 10 hand-failed in the Isages HTML: The voture of College and Printees have recognized this as the Isabi bear.

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 truly loved God, and that loving heart was given to them by the God who elected them to nut their faith in Jesus and to do those good works for his own glory. He also encouraged Northwestern students who had Ouaker ties to 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 Repert 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 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, lazal and the pige of GCC. Power output is 28 BHP in standard from with 4.3. No it foreps and tested under Euro 3 conditions the engine returns just under 80 mpg, which means an impremire tank range of over 200 miles before witching to reserve. As Levy sees 81, after the collapse

Figure: Third (horizontal) cut of the PDF page

Elmar Haussmann University of Freiburg Georges Köhler Allee 051 Engluere 70110

2 INTRODUCTION

Sanhaaria R. Kottamanu, MD, of Saginaw, Mich, was recently appointed to the Central Michigan University Board of Control. The project also provide coverage in four downtown South particle. Covidential, Ference, Wentikle and Visior Steinforced, as well as the City Hall holdy area. February 24 (2003) 4010 Miching through sources (2004), for an arrive and 2003 1010 Miching through sources (2004). The antial micro stein and the stein stein and the stein and the bar.

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 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 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 Repert 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 declarative rules, and access control lists (ACLs). Holland 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, Iszał and the pige of GCC. Power output is 28 BHP in standard from with 4.1.3 Nm of teropes and tested under Euro 3 conditions the engine returns just under 80 mpg, which means an impremire tank range of over 200 miles before witching to reserve. As Levy sees 83, after the collapse

Figure: Fourth (vertical) cut of the PDF page

2 INTRODUCTION

Sanhaaris R. Kottamun, MD, of Sagiane, Wich, was recently appointed to the Central Michigan University Board of Control. The project also provide coverage in four downtown South particle. Covidential, Ference, Wentikle and Vision Steinforced, as well as the City Hall holdy area. February 24 (2003) 4010 Miching through sources (2004), for an arrive and 2003 1010 Miching through sources (2004). The analysis of the second stein stein and the second stein stein and a second stein and Articona have recognized this at the holds how.

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 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 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 Repert 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 declarative rules, and access control lists (ACLs). Holland 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, Iszał and the pige of GCC. Power output is 28 BHP in standard from with 4.1.3 Nm of teropes and tested under Euro 3 conditions the engine returns just under 80 mpg, which means an impremire tank range of over 200 miles before witching to reserve. As Levy sees 81, after the collapse

Figure: Final (vertical) cut of the PDF page

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

3 Approach

She also less cate seperative Window, her lifetite shorthar, and Abdu. A bate Com muck, and plasters with her separation from the provide are nulabout join over an tellum delines in grants in support, which, densing HW, address, support drive no morphics, here plasters quark, the histoger on and con. The union is organizing a plobal campaign against Rio Thoto s and here one. The union is organizing a plobal campaign against Rio Thoto s and here instants of support from instem in Grandar and beyond, and if Wolfand Gay Farrell. The grasp have also relationed for inform.



Figure 1: The SCI tree for our example sentence after anaphora resolution. The band 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 totts/rheumatic dissusses. Four dises are of flow from my Native Land, the US

Figure: Randomly generated PDF with a Manhattan layout

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

3 Approach

She also less cate seperative Window, her lifetite shorthar, and Abdu. A bate Com muck, and plasters with her separation from the provide are nulabout join over an tellum delines in grants in support, which, densing HW, address, support drive no morphics, here plasters quark, the histoger on and con. The union is organizing a plobal campaign against Rio Thoto s and here one. The union is organizing a plobal campaign against Rio Thoto s and here instants of support from instem in Grandar and beyond, and if Wolfand Gay Farrell. The grasp have also relationed for inform.



Figure 1: The SCI tree for our example sentence after anaphora resolution. The band 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 totts/rheumatic dissusses. Four dises are of flow from my Native Land, the US

Figure: Randomly generated PDF with a Manhattan layout

- 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 Ratch and Merala Kenhar. Now prespectives on gender and particle denied by Yaussin K. Kadi, Carteri Hoster, J. Di Domer, and Joseffer Y. Sus Jonak review, Mani di has shalared part of the roots in through Pitry Channel and Antone and State and State and State and State and State and Antone and State and State and State and State and State and Mathematical Antone and State and State and State and State Mathematical Antone and State and State and State and State Mathematical Antone and State and

See also ions cats especially Winton, her littish sherihar, and Ablu, a Mana Coto made, and jaktir with her grandfalfers. They use can real about joint over its littise datase is provide the strengther with the strengther previous previous concentration, RelA. Mana strength on the strengther and the strengther and the strengther on. The union is organizing a global campeing adaret files Thus s and her inter or singorie romations in Gaussian aboved, and Gaussian down first or disapper from matters in Gaussian ab bored, and Gaussian files for a strengther and the strengther and the strengther adaret files that is non-streng relies that ad A player is 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 Panamie (Nobra), Chumathang and Puga of Changthang, which are famous for early euring of joints/rheumatic diseases. Four days ago I flew from my Native Land, the US

Figure: Randomly generated PDF with a Manhattan layout

- 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 as War, 1030-1045 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 data measures 7.2 Were Maxwell, due to the source of the data of the source of t

3 Experiments

CHABOT Raymond Joseph Emile of Danielson, Con-		lt hurt	son series 7.5. Presentation at the annual meeting of the
necticut and of St Peters- burg, Florida, USA, third	L	because	National Council for Occur pational Education, Scotts
in descent from Joseph Chabot, born in Canada. If	it	mat-	dale, AZ. On 1 Octobe 1938, the 28th PS was re
the head-to-head tiebreaker comes into play because the		tered.	assigned to the 3rd PG and sent to Lanchou to train or
losse has the WC to fall back of			Que .

Figure: Randomly generated PDF with a non-Manhattan layout

- 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

Evolutione 1, 3 or / days in Oslo, Norwar/ See also Albania at War, 1939-1945 M				
Const Horkey. Not says allow the laws one I would importe since addits owns the				
The late because where definite second line size are second to be her to be the second				
bit the forme.				
One of the older structures, T/1 West Marwoll, dates from the 1800s and a				
me of only four buildings left in this part of Chicago from the period before the				
sreat Hire. They have every kind of style and everything is so exputste we drive a				
177 hours out every year to pack out out Christinas empiricas each year and they				
to have acrosous Santa hats. One of the proneering figures in the actual Latin				
Wortheline movement, Brano Garda Talka, Servent Gardial will Grine his converted				
colliting and default sound to leativals and clubs across the US and Canada, Ulins				
or a score in Descrip that most illustrates this floor in mind that I m working from				
memory and have at read it in the work! look new the horizoine when lowerhad				
Under is visibles the Louise a cards and is set upon by the sometry women. Hereine				
in some Verserer is not related to destructure of these sides and respect the				
the second s				
WHAT AND THE OF PAPER OF ASSAULT BRAN STUD WHAT BRAND WAS STUD A				
A mone was about beauty dealanter. It will average your people range me of				
un mightadi. Tane 1-210 Spir Unaru 1 495 S / Notumiti Virgina				
3 Experiments				
The Incorporation by reference of an Exchange Act report, such as a Form 30-K, that				
The incorporation by relevance of an Exchange Accreport, such as a 1-orm IU-R, that populate a restrigation scattering. For purposes of Section 10 (a) (3) of the Securities				
The incorporation for relevance of an Exchange Act report, such as a horm 30 K, that potential a representation statement for purpose of Section 30 (a) (a) of the Sectified Part Is not an emergence for purpose of the Question 3. In 1996, R.C., Nardin				
The incorporation by relevance of an Exchange Accreport, such as a from 10-K, they gotaxies a redistration statement for purposes of Section 10 (a) (b) of the Securitise Pact is not an amendment, for purposes of this Question 3, in 1955, K-L, Nature For Uniform to Arr Mocore them: Composed at this Duesda at the one to be control cubru.				
The incorporation by reference of an Exchange Accregory, such as a Form VER, what potence a reference consistence for purpose of Sectors 10 Fe1133 of the SectorsRep ACL is force an elementeric to propose of the Sectors 31 Fe1205, RCL, Resetting Part Americans Ber Att Monten Element source peaked as the Sector Sector for dearmone brinds that a feature Longer Accession 31 Fe1205 and the dearmone brinds that a feature Longer Accession 31 Fe1205 and a dearmone brinds that a feature Longer Accession.				
The homoparticle is a constrained and behavior developments are a set on the UK-BANA Constrained and the set of the Set of the set of the Set of the set of the Set of the set of the Set of the set of the Set of the set of the set of the set of the set of the set of the set of the set of the set of the set of the set of the set of the set of the set of the set of the set of the set of the set o				
The incomparison is protocol of a 1 belongs Arc support, and a 1 second biol. We operate a response to instruction of a proposed in second to 1 at 1 at 1 at 1 at 2 at 2 at 2 at 2 at				
The decomposition is a decomposition of a first hard and express which are been detected by the composition of the decomposition of th				

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

- 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

trio of Matsuda Haruka, Asada Rina and Ishimir	ne Kanako, 663-639 to advance
to the final while the second Korean trio knocke	ed out the second Japanese trio
of Harigaya Junko, Sasaki Natsuki and Totsuka R	ie, 640-628 to assure themselves
either a gold or silver. The Republican Party is fri	ghtened by Ron Paul.
	the Braves currently lead the sea-
CHABOT Raymond Joseph	son series 7-5. Presentation
Emile of Danie lson, Con-	at the annual meeting of the
hecticut and of St Peters-	National Council for Occu-
burg, Florida, USA, third	pational Education, Scotts-
n descent from Joseph	dale, AZ. On 1 October
Chabot, born in Canada. If 112 III Chabot, born in Canada. If	1938, the 28th PS was re-
the head-to-head tiebreaker	assigned to the 3rd PG and
comes into play because the Lerea.	sent to Lanchou to train on
oser has the WC to fall back on.	-15bis

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

Table of Contents

Problem motivation

- PDF content extraction
- Rule-based approaches
- Proposed solution
 - Sequence labeling problem
 - Implementation
 - Advantages

Results

Final remarks

- 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:

- 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)

- 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:
 - the text of the character
 - the distance to the next character
 - I "font features" of the character

- 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:
 - the text of the character
 - 2 the distance to the next character
 - I "font features" of the character
 - Feed line (sequence) of characters to model to make predictions, if each character is:
 - at the end of a word
 - at the end of a column
 - Inone of the above
• Line grouping - groups the characters of a PDF page to processable sequences

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

- 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**

- 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

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

Hānnah 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: Snippet from randomly generated PDF document

Broccoli: Semantic Full-Text Search at your Fingertips

Bjärn Buchhold Nüklas Schälle Uaiversity of Freiburg University of Freiburg buchkelder, uni-fresburg, de Hännah Bäst University of Freiburg bartes, uni-freiburg

September 2017

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

Hānnah 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

Bjårn Buchhold Nüklas Schälle University of Freihung University of Freihung buchholdfes, uni-freihung, de Hännah Båst University of Freihung hasteter, uni-freihung, de Spottembler 2017

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

Broccdi: Semantic Full-Text Search at your Fingertips

Bjàrn Buchhold Nüklas Schàlle University of Freiburg University of Freiburg buchhold@cs.uni-freiburg.de Hännah Bäst

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

September 2017

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

Hānnah 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 s

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

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

September 2017

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

Hānnah 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

Bjärn Buchhold Nüklas Schälle University of Predung University of Predung buchholdes, uni-freshure, de Hännah Bäst University of Predung bastes, uni-freshurg, de September 2017

Broccoli: Semantic Full-Text Search at your Broccoli: Semantic Full-Text Search at your Broccoli: Semantic Full-Text Search at your Fingertips Fingertips Fingertips Bjàrn Buchhold Nüklas Schàlle Bjårn Buchhold Nüklas Schàlle Bjàrn Buchhold Nüklas Schàlle University of Freiburg buchhold@cs.uni-freiburg.de schnelle@cs.uni-freiburg.de buchhold@cs.uni-freiburg.de schnelle@cs.uni-freiburg.de buchhold@cs.uni-freiburg.de schnelle@cs.uni-freiburg.de Hānnah Bàst Hännah Bàst Hānnah Bàst University of Freiburg University of Freiburg University of Freiburg bast@cs.uni-freiburg.de bast@cs.uni-freiburg.de bast@cs.uni-freiburg.de September 2017 September 2017 September 2017 Figure: Snippet from randomly

generated PDF document

Figure: Building the first line of characters

Broccoli: Semantic Full-Text Search at your Broccoli: Semantic Full-Text Search at your Broccoli: Semantic Full-Text Search at your Fingertips Fingertips Fingertips Nüklas Schàlle Bjàrn Buchhold Nüklas Schàlle Bjårn Buchhold Nüklas Schàlle Bjarn Buchhold University of Freiburg buchhold@cs.uni-freiburg.de schnelle@cs.uni-freiburg.de buchhold@cs.uni-freiburg.de schnelle@cs.uni-freiburg.de buchhold@cs.uni-freiburg.de schnelle@cs.uni-freiburg.de Hānnah Bàst Hännah Bàst Hānnah Bàst University of Freiburg University of Freiburg University of Freiburg bast@cs.uni-freiburg.de bast@cs.uni-freiburg.de bast@cs.uni-freiburg.de September 2017 September 2017 September 2017 Figure: Snippet from randomly

generated PDF document

Figure: Building the first line of characters

Broccoli: Semantic Full-Text Search at your Broccoli: Semantic Full-Text Search at your Broccoli: Semantic Full-Text Search at your Fingertips Fingertips Fingertips Bjarn Buchhold Nüklas Schàlle Bjàrn Buchhold Nüklas Schàlle Bjårn Buchhold Nüklas Schàlle University of Freiburg buchhold@cs.uni-freiburg.de schnelle@cs.uni-freiburg.de buchhold@cs.uni-freiburg.de schnelle@cs.uni-freiburg.de buchhold@cs.uni-freiburg.de schnelle@cs.uni-freiburg.de Hānnah Bàst Hännah Bàst Hānnah Bàst University of Freiburg University of Freiburg University of Freiburg bast@cs.uni-freiburg.de bast@cs.uni-freiburg.de bast@cs.uni-freiburg.de September 2017 September 2017 September 2017 Figure: Snippet from randomly

generated PDF document

Figure: Building the first line of characters

Broccoli: Semantic Full-Text Search at your Broccoli: Semantic Full-Text Search at your Broccoli: Semantic Full-Text Search at your Fingertips Fingertips Fingertips Nüklas Schàlle Bjàrn Buchhold Nüklas Schàlle Bjårn Buchhold Nüklas Schàlle Bjàrn Buchhold University of Freiburg buchhold@cs.uni-freiburg.de schnelle@cs.uni-freiburg.de buchhold@cs.uni-freiburg.de schnelle@cs.uni-freiburg.de buchhold@cs.uni-freiburg.de schnelle@cs.uni-freiburg.de Hännah Bàst Hānnah Bàst Hännah Bàst University of Freiburg University of Freiburg University of Freiburg bast@cs.uni-freiburg.de bast@cs.uni-freiburg.de bast@cs.uni-freiburg.de September 2017 September 2017 September 2017

Figure: Snippet from randomly generated PDF document

Figure: Building the first line of characters

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

- Text lines from the previous step sequences to be fed to an Encoder-Decoder model
- Preprocess each data point from the sequences to contain:
 - 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)



Figure: Input features of each data point

lanvu lanev				
Tanvu Tanev	2.00		1.2.5	0.01
		/11		IPV.

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



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

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

- **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



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

• 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:

• 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



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

- 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



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

Advantages

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

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

Problem motivation

- PDF content extraction
- Rule-based approaches
- Proposed solution
 - Sequence labeling problem
 - Implementation
 - Advantages

Besults

Final remarks

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 Manhattan only non-Manhattan only	0.965 0.984 0.923	0.915 0.920 0.903	0.932 0.941 0.912	0.971 0.991 0.929	0.964 0.980 0.931	0.967 0.985 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 Manhattan only	0.965	0.915 0.920	0.932 0.941	0.971 0.991	0.964	0.967
Manhattan with	0.925	0.905	0.912	0.929	0.931	0.930
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 Manhattan only non-Manhattan only	0.965 0.984 0.923	0.915 0.920 0.903	0.932 0.941 0.912	0.971 0.991 0.929	0.964 0.980 0.931	0.967 0.985 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	e extract	ion	Word extract		
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

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

	Line extraction			Word extraction		
Test set	Precision	Recall	F1 Score	Precision	Recall	F1 Score
Manhattan + non-Manhattan Manhattan only	0.913 0.904	0.923 0.911	0.917 0.907	0.916 0.901	0.919 0.904	0.918 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

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

	Line extraction			Word extraction		
Test set	Precision	Recall	F1 Score	Precision	Recall	F1 Score
Manhattan + non-Manhattan Manhattan only	0.913 0.904	0.923 0.911	0.917 0.907	0.916 0.901	0.919 0.904	0.918 0.902
Manhattan with	0.903	0.905	0.904	0.907	0.970	0.908
broken whitespaces	0.961	0.951	0.955	0.920	0.878	0.899

Pdftotext results

	Line extraction Wo			d 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 non-Manhattan only	0.913	0.876**	0.887	0.952	0.959	0.956
Manhattan with	0.070	0.009	0.005	0.921	0.954	0.950
broken whitespaces	0.524	0.529	0.527	0.541	0.520	0.530

Pdftotext results

	Lin	e extracti	on	Wor	d extract	ction	
Test set	Precision	Recall	F1 Score	Precision	Recall	F1 Score	
Manhattan + non-Manhattan Manhattan only non-Manhattan only	0.902 0.913 0.878	0.880** 0.876** 0.889	0.886 0.887 0.883	0.944 0.952 0.927	0.951 0.959 0.934	0.947 0.956 0.930	
Manhattan with broken whitespaces	0.524	0.529	0.527	0.541	0.520	0.530	

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

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 of Contents

Problem motivation

- PDF content extraction
- Rule-based approaches
- Proposed solution
 - Sequence labeling problem
 - Implementation
 - Advantages

Results



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

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

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

- Transform line grouping to a machine learning problem:
 - completely removes all dependencies on custom thresholds
 - $\bullet\,$ 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]

- Transform line grouping to a machine learning problem:
 - completely removes all dependencies on custom thresholds
 - $\bullet\,$ 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

- 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.