Week 3

Zachary Daniels

Mentor: Haroon Idrees

Updated Synthetic Dataset

Document taken from scanned Google book

first mount my wall.

The very next day after this wall was finished, I had almost had all my labour overthrown at once, and myself killed. The case was thus: As I was busy in the inside of it, behind my tent, just in the entrance into my cave, I was terribly frighted with a most dreadful surprising thing indeed; for, all on a sudden I found the earth come crumbling down from the roof of my cave, and from the edge of the hill over my head, and two of the posts I had set up in the cave cracked in a frightful manner. I was heartily scared, but thought nothing of what really was the cause, only thinking that the top of my cave was falling in, as some of it had done before. And for fear I should be buried in it, I ran forward to my ladder, and not thinking myself safe there neither, I got over my wall for fear of the pieces of the hill which I expected might roll down upon me. I was no sooner stepped down upon the firm ground but I plainly saw it was a terrible earthquake; for the ground I stood on shook three times at about eight minutes distance, with three such shocks as would have overturned the strongest building that could be supposed to have stood on the earth, and a great piece of the top of a rock which stood about half a mile from me, next the sea, fell down with such a terrible noise as I never heard in all my life. I perceived also the very sea was put into a violent motion by it, and I believe the shocks were stronger under the water than on the island.

I was so amazed with the thing itself, having never felt the like, or discoursed with any one that had, that I was like one dead or

“Shredded” with displacement

first mount my wall.

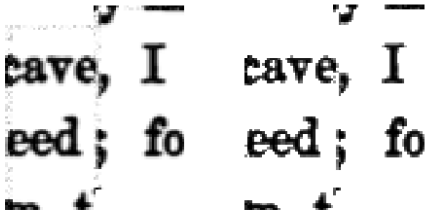
The very next day after this wall was finished, I had almost had all my labour overthrown at once, and myself killed. The case was thus: As I was busy in the inside of it, behind my tent, just in the entrance into my cave, I was terribly frighted with a most dreadful surprising thing indeed; for, all on a sudden I found the earth come crumbling down from the roof of my cave, and from the edge of the hill over my head, and two of the posts I had set up in the cave cracked in a frightful manner. I was heartily scared, but thought nothing of what really was the cause, only thinking that the top of my cave was falling in, as some of it had done before. And for fear I should be buried in it, I ran forward to my ladder, and not thinking myself safe there neither, I got over my wall for fear of the pieces of the hill which I expected might roll down upon me. I was no sooner stepped down upon the firm ground but I plainly saw it was a terrible earthquake; for the ground I stood on shook three times at about eight minutes distance, with three such shocks as would have overturned the strongest building that could be supposed to have stood on the earth, and a great piece of the top of a rock which stood about half a mile from me, next the sea, fell down with such a terrible noise as I never heard in all my life. I perceived also the very sea was put into a violent motion by it, and I believe the shocks were stronger under the water than on the island.

I was so amazed with the thing itself, having never felt the like, or discoursed with any one that had, that I was like one dead or

How does Tesseract perform?

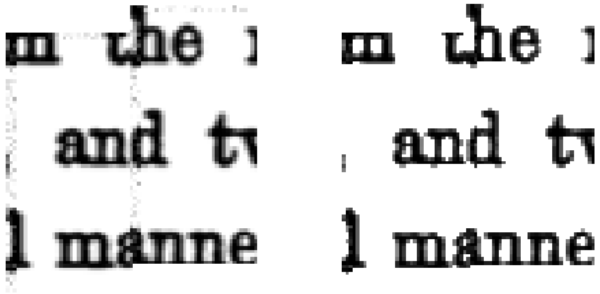
Examples

- ‘cave, I’
- 'eed ; !o’



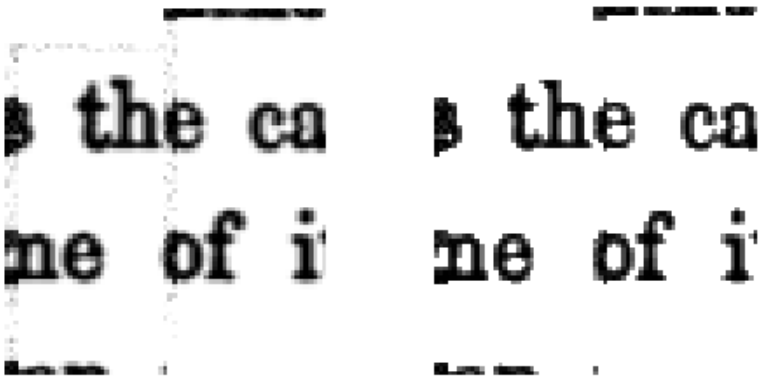
Examples

- 'mvhe l’
- “
- 'lnmnne'



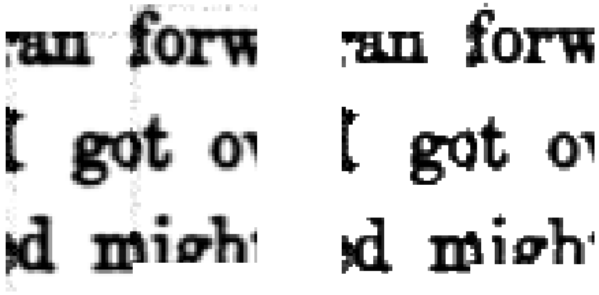
Examples

- 'ILhaea'



Examples

- ' 'm fonu’
- 'I got ov’
- 'I d minh '



Accuracy is Down

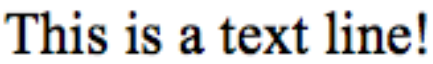
- Likely Reasons
  - Tesseract might not be trained on specific font face
  - Little spacing between characters might make character segmentation more difficult

Potential Solutions

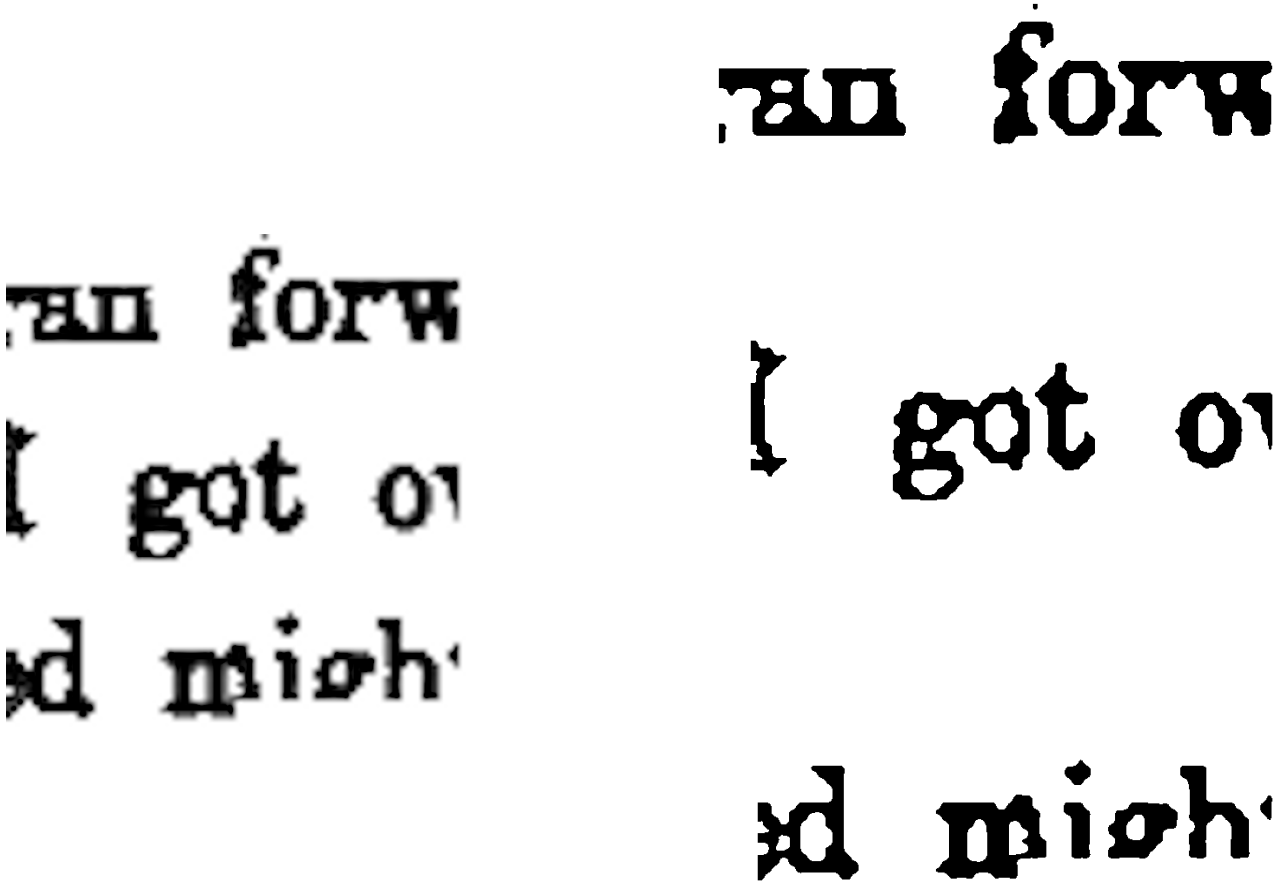
- Train on different page from same book?
  - Want a generalized system at this point in time
- Use different noise?
- Try different OCR software?
  - OCRopus (also Google-funded)

OCRopus Simple Recognition

- Simplest test case possible:
- ‘This is a text linel’



Line Segmentation + Cleaning/ Binarization



How does OCRopus perform?

OCRopus Recognition

- ‘rn forw’



OCRopus Recognition Cont.

- ‘I got o’



OCRopus Recognition Cont.

- ‘nd mioh’



Compared to Tesseract

- For the "chad problem" on the new shredded dataset, OCRopus has near perfect accuracy: much better than Tesseract on the same problem

Ideal test set

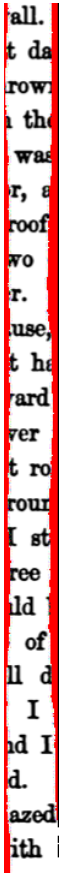
- As similar to real shredded document as possible
  - Real document ✓
  - Shards don't line up perfectly ✓
  - Noisy edges ✓
  - Jagged edges
    - Simulate with Brownian paths

Better Synthetic Shredding

first mount my wall.

The very next day after this wall was finished, I had almost had all my labour overthrown at once, and myself killed. The case was thus: As I was busy in the inside of it, behind my tent, just in the entrance into my cave, I was terribly frighted with a most dreadful surprising thing indeed; for, all on a sudden I found the earth come crumbling down from the roof of my cave, and from the edge of the hill over my head, and two of the posts I had set up in the cave cracked in a frightful manner. I was heartily scared, but thought nothing of what really was the cause, only thinking that the top of my cave was falling in, as some of it had done before. And for fear I should be buried in it, I ran forward to my ladder, and not thinking myself safe there neither, I got over my wall for fear of the pieces of the hill which I expected might roll down upon me. I was no sooner stepped down upon the firm ground but I plainly saw it was a terrible earthquake; for the ground I stood on shook three times at about eight minutes distance, with three such shocks as would have overturned the strongest building that could be supposed to have stood on the earth, and a great piece of the top of a rock which stood about half a mile from me, next the sea, fell down with such a terrible noise as I never heard in all my life. I perceived also the very sea was put into a violent motion by it, and I believe the shocks were stronger under the water than on the island.

I was so amazed with the thing itself, having never felt the like, or discoursed with any one that had, that I was like one dead or









“Deshredder”

- Virginia Tech’s entry into the DARPA Shredder Challenge
  - Butler, Patrick, Prithwish Chakraborty, and Naren Ramakrishan. "The Deshredder: A visual analytic approach to reconstructing shredded documents." Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on. IEEE, 2012.
- Performs edge matching by modeling edges as time-series
- GUI for human validation
- Emailed first and second author to see if we can build on top of their code

Long-Term Goals

- Solve new realistic synthetic dataset
- Solve the DARPA Challenge dataset

Short-Term/Mid-Term Goals

- Get OCRopus functioning with new dataset
- Test OCRopus on false matches and false alignments
- Build an oracle for the synthetic dataset so that matches can be scored automatically
- Integrate OCR software with edge-matching software
- Implement OCR pruning solution