Week 2 Review

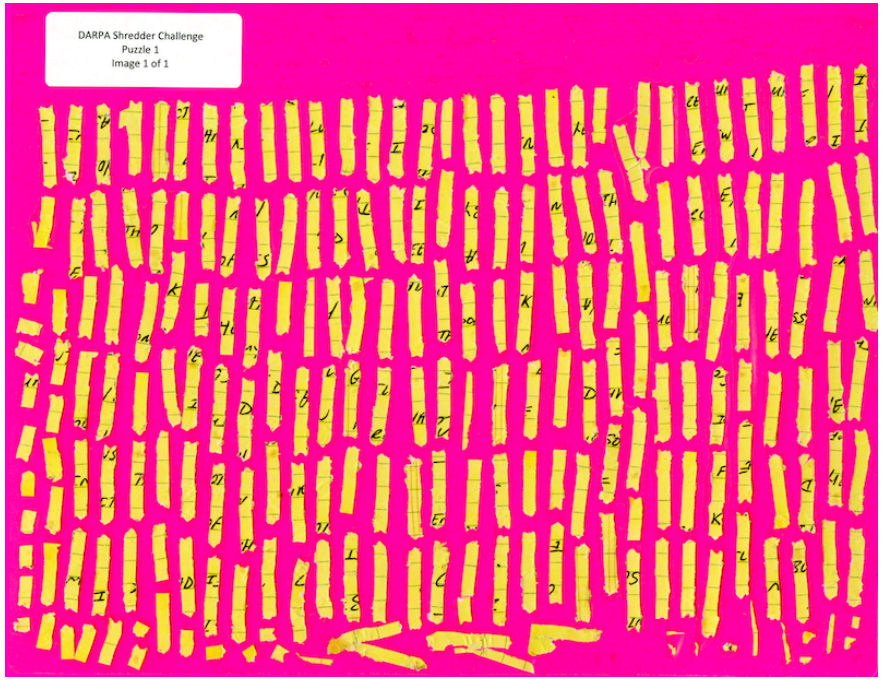
OCR-Aided Cross-Cut Shredded Document Reconstruction

Student: Zachary Daniels

Mentor: Haroon Idrees

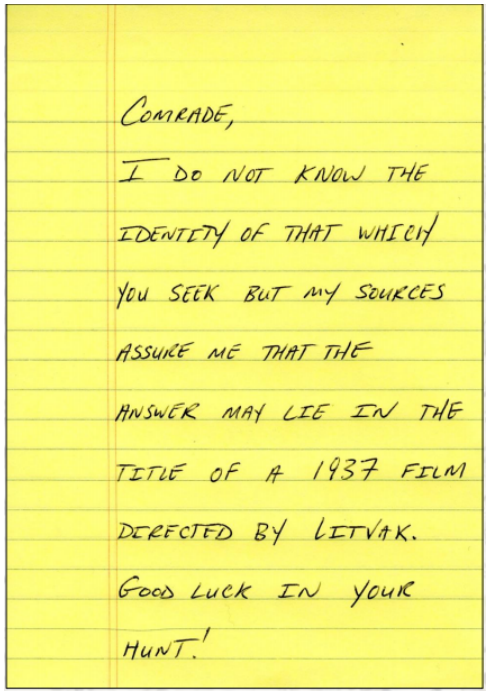
Problem

- Given:
  - A cross cut document



Problem Cont.

- Output:
  - An image of the original document reconstructed from the shreds



Approaches

- Edge-Based Matching
- Content-Based Matching
  - e.g. brightness, color, etc.
- Semantics-Based Matching
  - e.g. OCR in combination with algorithms and techniques borrowed from NLP and information retrieval
    - This is the focus of the REU project

Relevant Reading

- Deever, Aaron, and Andrew Gallagher. "Semi-automatic assembly of real cross-cut shredded documents." Image Processing (ICIP), 2012 19th IEEE International Conference on. IEEE, 2012.
- 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.
- Zhang, Haoqi, John K. Lai, and Moritz Bächer. "Hallucination: A Mixed-Initiative Approach for Efficient Document Reconstruction." Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence. 2012.
- Smith, Nathan. Automatic Character Identification for Shredded Document Reconstruction. Diss. Rochester Institute of Technology, 2013.
- Richter, Fabian, et al. "Learning to Reassemble Shredded Documents." (2013): 1-1.
- Perl, Johannes, et al. "Strip shredded document reconstruction using optical character recognition." Imaging for Crime Detection and Prevention 2011 (ICDP 2011), 4th International Conference on. IET, 2011.
- Lin, Huei-Yung, and Wen-Cheng Fan-Chiang. "Reconstruction of shredded document based on image feature matching." Expert Systems with Applications 39.3 (2012): 3324-3332.
- Harding, Stephen M., W. Bruce Croft, and C. Weir. "Probabilistic retrieval of OCR degraded text using n-grams." Research and advanced technology for digital libraries. Springer Berlin Heidelberg, 1997. 345-359.
- Smith, Ray. "An overview of the Tesseract OCR engine." Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on. Vol. 2. IEEE, 2007.
- Smith, Ray W. "History of the Tesseract OCR engine: what worked and what didn't." IS&T/SPIE Electronic Imaging. International Society for Optics and Photonics, 2013.
- Rakshit, Sandip, and Subhadip Basu. "Development of a multi-user handwriting recognition system using Tesseract open source OCR engine." arXiv preprint arXiv:1003.5886 (2010).

Tasks

1. Construct synthetic dataset for testing
2. Apply optical character recognition
3. Build algorithm for restricting space of potential solutions using OCR output
4. Integrate OCR-restriction algorithm into existing de-shredding software
5. Test on DARPA challenge dataset

Synthetic Shredding 1

- Microsoft Word
- Machine-printed typeface
  - Arial, Times New Roman, handwriting-mimicking font
- Natural language text from classic novel
- Broken into grid
- Erosion and Gaussian noise applied along edges

Synthetic Shredding 1 Cont.

After this he pressed me earnestly, and in the most affectionate manner, not to play the young man, nor to precipitate myself into miseries which nature, and the station of life I was born in, seemed to have provided against; that I was under no necessity of seeking my bread; that he would do well for me, and endeavour to enter me fairly into the station of life which he had just been recommending to me; and that if I was not very easy and happy in the world, it must be my mere fate or fault that must hinder it; and that he should have nothing to answer for, having thus discharged his duty in warning me against measures which he knew would be to my hurt; in a word, that as he would do very kind things for me if I would stay and settle at home as he directed, so he would not have so much hand in my misfortunes as to give me any encouragement to go away; and to close all, he told me I had my elder brother for an example, to whom he had used the same earnest persuasions to keep him from going into the Low Country wars, but could not prevail, his young desires prompting him

Synthetic Shredding 2

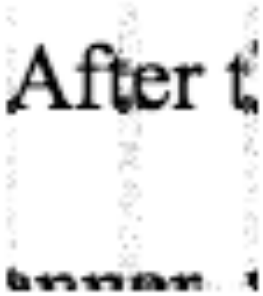
- Actual scan of printed document
- Machine-printed typeface
- Natural language
- Grid with offsets, so shreds don't align perfectly
- Experiment different deformations applied to edge

Optical Character Recognition

- Apply open-source, freely available software
  - Google's Tesseract v. 3.02.02
    - Designed for reading passages of machine-printed text
    - Ultimately it will be applied to short sequences of characters of noisy, handwritten text

Preliminary Results

- Output: 'Afher L ’



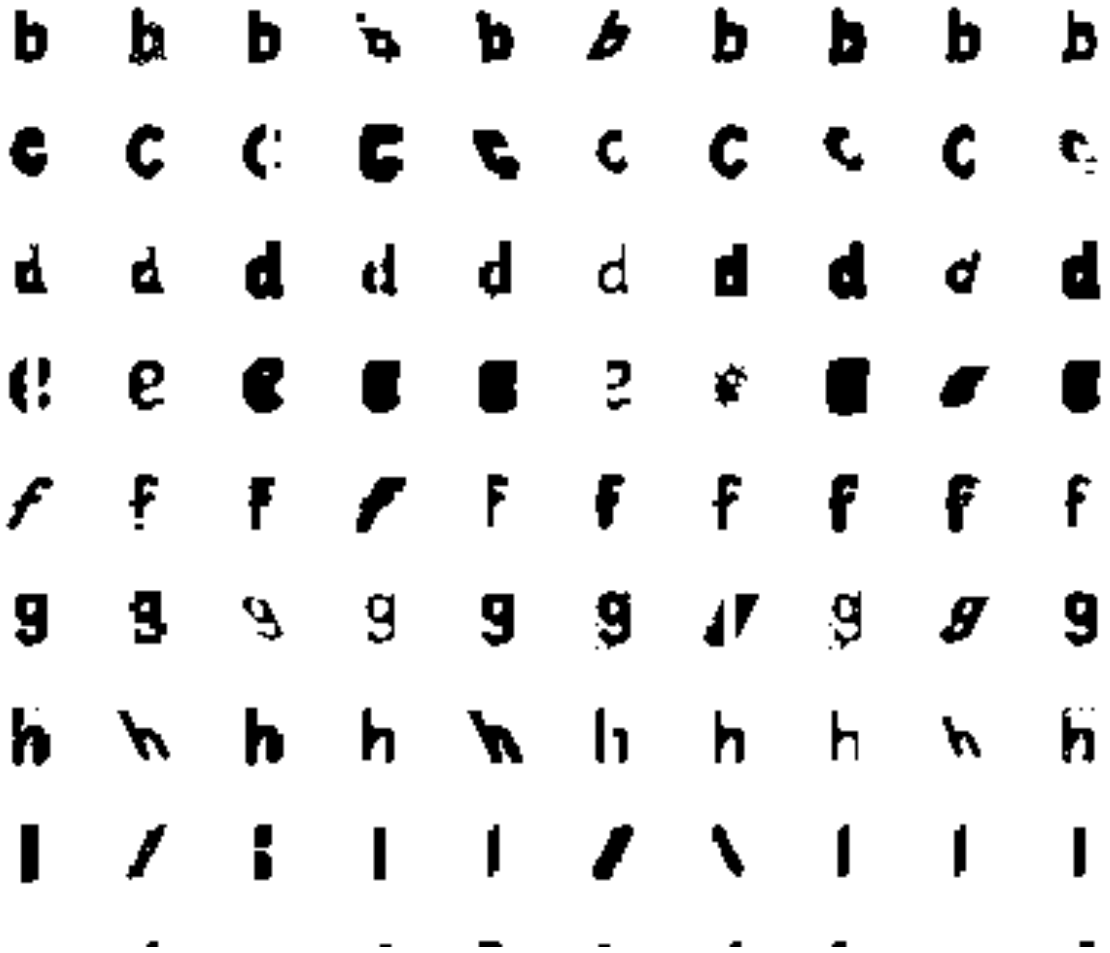
Noise and Deformities

- Shredded documents have more noise and deformities than typical documents
- How to account for noise?
  1. Train OCR software on noisy data
  2. Clean image

Training on Noisy Data

- Build character set (62 characters)
- 23 transformations
  - Morphology
  - Noise
  - Cuts
  - Etc.
- Apply combinations of 3 transformations to characters
- Train Tesseract

Noisy Character Subset

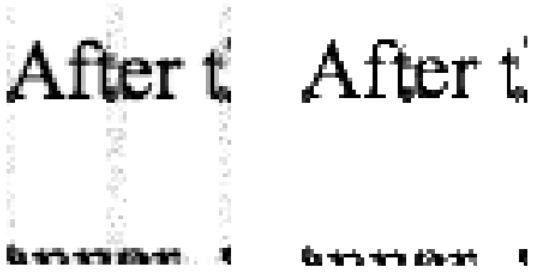


Training on Noisy Data Cont.

- Advantages
  - Should (in theory) increase robustness
- Disadvantages
  - Very time consuming
  - Very labor intensive
  - Very application-specific
  - Too much variation in characters can cause severe drop in Tesseract's accuracy
  - Tesseract OCR features are designed to be robust to noisy input
    - Only trained on small subset of clean data

Removing Noise

- Clean images with ImageMagick using cleantext script
- Improves results slightly



Why OCR software makes mistakes

1. Cut off characters
   - Most common reason, easily addressable
2. Noise along seam and edges
3. Similar characters

Next Steps

- Building new synthetic dataset
- Researching algorithms for edit distance, and bigram, trigram, word validation
- Determining which software to add OCR component to

Questions?