Urdu Nastalique Optical Character Recognition System Training Program

April 16, 2012 to 20, April 2012

Session	Task	Time Duration	Trainer
Day 1,Session1	Introduction to DIP 1. Image 2. Color Intensities 3. Masking 4. Connected Components Algorithm	10:30 to 11:15	Qurat ul Ain Akram
Day 1,Session 2	Practice Session 1. 2D Matrix Manipulation a. Segmentation using Vertical and Horizontal Histogram	11:15 to 12:00	
	Lunch Break	12:00 to 12:30	
Day 1, Session 3	b. Connected Components Extraction i. Overlapped Version ii. Isolated Version c. Average Mask Implementation d. Median Mask Implementation	12:30 to 4:30	
Day 1,Session4	• Nastalique and Pre-Processing Papers Discussion • Reading Material #1 (Home Work ☺)	4:30 to 5:00	
Day 2			
Day 2,Session1	2. Image read Write 3. Binarization, and Noise Removal	9:00 to 10:00	
	Tea Break	10:00 to 10:30	
Day 2,Session2	4. Image Read and Write 5. .bmp to .txt File conversion 6. Binarization and noise removal	10:30 to 12:00	
	Lunch Break	12:00 to 12:30	
Day 2,Session3	7. Implementation of (1) using Images	12:30 to 2:30	
Day 2,Session4	Implementation of (1) using Images Discussion on recognizers Reading Material #2 (Home Work ☺)	2:30 to 4:30 4:30 to 5:00	
Day 3			
Day 3, Session1	1. Overview of Machine Learning Techniques a. Neural Networks b. HMMs c. Decision Tree	9:00 to 10:00	Aniqa Dilawari
	Tea Break	10:00 to 10:30	
Day 3, Session 2	Practice Session of one recognizer technique	10:30 to 12:00	

	Lunch Break	12:00 to 12:30	
Day 3, Session 3	2. Practice Session of one recognizer technique 3. Tesseract a. Overview in context of other languages b. Tesseract training Manual for Latin	12:30 to 1:30 1:30 to 2:30	
Day 3, Session4	Practice Session 4. Tesseract Training for Latin 5. Reading Material #3 (Home Work ☺)	2:30 to 5:00	
Day 4			
Day 4,Session1	6. Discussion on Reading Material #3 7. Tesseract Training Manual for Urdu	9:00 to 10:00	Aniqa Dilawari
	Tea Break	10:00 to 10:30	
Day 4,Session 2	Practice Session 8. Tesseract Training for Urdu	10:30 to 12:00	
	Lunch Break	12:00 to 12:30	
Day 4, Session 3	9. Practice Session on Tesseract Training for Urdu 10. Overview of Segmentation-free a. Flow of segmentation-free approach used in the project i. Training ii. Recognition	12:30 to 2:30 2:30 to 4:00	
Day 4, Session4	11. Process Flow Diagram of Urdu Nastalique OCR 12. Reading Material #4 (Home Work ☺)	4:00 to 5:00	
Day 5			
Day 5, Session1	Introduction to Post processing 5. Input 6. Word segmentation 7. Spell Checking 8. POS Tagging 9. Corpus Cleaning Guidelines	9:00 to 10:00	Farah Adeeba
	Tea Break	10:00 to 10:30	
Day 5, Session 2	Practice Session 1. Unicode File Read & Write 2. Urdu Text Normalization	10:30 to 12:00	
	Lunch Break	12:00 to 12:30	
Day 5, Session 3	3. High Frequency wordlist 4. Ligature Splitting	12:30 to 2:30	

Day 5, Session4	5. Ligature List with frequency 6. Top 10 Ligature List	2:30 to 5:00	

Reading Material #1

1. Context Sensitive Shape-Substitution in Nastaliq Writing System: Analysis and Formulation, available at http://www.cle.org.pk/Publication/papers/2006/context_sensitive_shape_substitution.pdf
2. www.LICT4D.asia/fonts/Urdu_Nasta'leeq, available at http://www.cle.org.pk/Publication/papers/2003/www.LICT4D.asia.pdf
3. Layout Analysis of Urdu Document Images, available at http://www.dfki.de/lt/publication_show.php?id=2448
4. Improving Nastalique-Specific Pre-Recognition Process for Urdu OCR, available at http://www.cle.org.pk/Publication/papers/2009/Pre-Recognition-Process-for-Urdu-OCR.pdf

Reading Material #2

1. Segmentation Free Nastalique Urdu OCR
http://cle.org.pk/Publication/papers/2010/segmentation%20free%20nastalique%20Urdu%20OCR.pdf

2. Segmentation-based Urdu OCR
   a. Investigation into a segmentation based OCR for the Nastaleeq Writing System
   http://www.cle.org.pk/Publication/theses/2007/OCRSOBIA.pdf
   b. Urdu Optical Character Recognition System
   http://www.cle.org.pk/Publication/theses/2010/OCRMUAZ.pdf

Reading Material #3

1. An Overview of the Tesseract OCR Engine
http://static.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/en//pubs/archive/33418.pdf

2. Adapting the Tesseract Open Source OCR Engine for Multilingual OCR
http://static.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/en//pubs/archive/35248.pdf

Reading Material #4

1. Corpus Based Urdu Lexicon Development
http://www.cle.org.pk/Publication/papers/2007/corpus_based_urdu_lexicon_development.pdf

2. Analysis and Development of Urdu POS Tagged Corpora
http://www.cle.org.pk/Publication/papers/2009/Analysis-and-Development-of-Urdu-POS-Tagged-Corpora-camera-ready.pdf
3. Urdu Word Segmentation
http://www.cle.org.pk/Publication/papers/2010/ALR812.pdf
4. Normalization

Document provided by Trainers