CSL772 Assignment 1

Swapnil Palash 2012cs10257

August 19, 2014

1 Broad Idea

Tokenization using Regex with python

2 Brief description

- The first step was tokenization using just the whitespaces in the input, that gave a baseline F1 of 53%
- Using regex, additional whitespaces were added around punctuation marks, whereever deemed necessary, pushing the F1 to 80% with just 3-4 regex replaces.
- Using the diff utility on the gold file and the produced output, additional regex were added and that pushed the F1 to around 90%.
- Some clitics were manually replaced and the F1 jumped to around 93.
- Now, time handling was implemented,again via Regex and followed by Date handling, achieving a final F1 of 95.1098%.

3 Why This?

Testing on the gold-output provided gave the best F-Score for the current implementation.