DATA MINING WEB PAGES

BY JI WOO KIM

OVERVIEW

- Using different libraries to extract text from a web page
- Naïve sentence detection based on periods and part of speech tagging
- Harvesting blog data
- Showing blog with important words bolded

IMPORTANT LIBRARIES USED

- Boilerpipe and Feedparser for extracting text
- NLTK for sentence detection and part of speech tagging
- IPython.display and IPython.core.display to display the final result