

RED dataset description

Marilena Oita

February 19, 2014

The RED dataset has a source-centered structure. By source, we mean an active channel of a Web site.

RED has been constructed using Web feeds, semi-automatically collected. The dataset provides around 90 Web sources, and over 1000 manually annotated Web pages.

Each source information has been grouped in a folder, and named after the host name of the Web channel that defined that source.

Each high-level folder contains the representative Web feed for that source, fetched at a certain point in time from the Web. Through Web feed parsing, item URLs have been discovered and used to obtain sample pages.

Each item has then an associated folder, named using a hash id and containing the raw HTML contents of the sample page, and the annotations that have been manually performed for that page, to represent types of main content that we targeted.

We have done two types of annotations: `gold-std.txt` contains the text that is used in our research as the main annotation based on which all results have been computed; `full.txt` is meant to be an alternative main content interpretation.