Headlines corpora with automatically extracted news values scores

This dataset consists of two corpora comprising of unique headline identifiers (to enable recreating the corpus by querying the relevant API) and automatically extracted news values scores.

Recreating the headlines corpus:

To obtain the headline text and associated metadata, use the "article_id" column and query the relevant API using the "id" parameter (refer to the latest API documentation for parameter name):

- The Guardian: http://www.theguardian.com/open-platform

- New York Times: http://developer.nytimes.com/docs

You will need to apply for an API key first.

Column headings

Column name	**Description**	**News value**
article_id	Unique identifier	N/A
num_entities	Number of entities	NV1: Prominence
current_burst_size	Wikipedia current burst size	NV1: Prominence
burstiness	Wikipedia burstiness	NV1: Prominence
wiki_long_term	Wikipedia long-term prominence	NV1: Prominence
wiki_day_before	Wikipedia day-before prominence	NV1: Prominence
news_recent	News source recent prominence	NV1: Prominence
sentiment	Sentiment	NV2: Sentiment
polarity	Polarity	NV2: Sentiment
connotations	Connotations	NV2: Sentiment
bias	Bias	NV2: Sentiment
comparative_superlative	Comparative/superlative	NV3: Superlativeness
intensifiers	Intensifiers	NV3: Superlativeness
downtoners	Downtoners	NV3: Superlativeness
proximity	Proximity	NV4: Proximity
surprise	Surprise	NV5: Surprise
head_unique	Uniqueness	NV6: Uniqueness

Related publications:

Piotrkowicz, A, Dimitrova, VG and Markert, K (2016) Automatic Extraction of News Values from Headline Text. In: Proceedings of EACL. European Chapter of the Association for Computational Linguistics, 03-07 Apr 2017, Valencia, Spain.