Headlines data for social media popularity prediction

We present datasets for two news outlets (*The Guardian* and *New York Times*) which consist of headline feature vectors for several prediction models and the corresponding social media popularity measures.

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.

List of files

	The Guardian	New York Times
Feature vectors (our model)	guardian_train.csv guardian_test.csv	nyt_train.csv nyt_test.csv
Feature vectors (baselines)	*Unigrams baseline:* guardian_train_unigrams.csv guardian_test_unigrams.csv *Bandari et al. baseline:* guardian_train_bandari.csv guardian_test_bandari.csv *Arapakis et al. baseline:* guardian_train_arapakis.csv guardian_test_arapakis.csv	*Unigrams baseline:* nyt _train_unigrams.csv nyt _test_unigrams.csv *Bandari et al. baseline:* nyt _train_bandari.csv nyt _test_bandari.csv *Arapakis et al. baseline:* nyt _train_arapakis.csv nyt_test_arapakis.csv
Social media popularity	guardian_train_popularity.csv guardian_test_popularity.csv	nyt_train_popularity.csv nyt_test_popularity.csv

Column headings (feature vectors for our model):

Feature groups: NV = news values, S = linguistic style, M = metadata

Column name	Description	Feature group
article_id	Unique identifier	N/A
number_of_words	Number of words	S: Brevity
number_of_characters	Number of characters	S: Brevity
parse_tree_height	Parse tree height	S: Simplicity
non-terminal_nodes_total	Number of non-terminal nodes in the parse tree	S: Simplicity

entropy	Entropy	S: Simplicity
difficult_words	Proportion of difficult words	S: Simplicity
information_content	Information content	S: Simplicity
word_freq_news	Average word frequency	S: Simplicity
modality	Presence of modal event or a modal relation	S: Unambiguity
median_senses	Median number of senses	S: Unambiguity
exclamation_mark	Presence of exclamation mark	S: Punctuation
question_mark	Presence of question mark	S: Punctuation
quote_marks	Presence of quote marks	S: Punctuation
three_consecutive_nouns	Presence of 'headlinese'	S: Nouns
np_count	Proportion of NPs to other syntactic chunks	S: Nouns
vp_count	Proportion of VPs to other syntactic chunks	S: Verbs
proportion_of_nouns	Proportion of nouns	S: Nouns
proportion_of_verbs	Proportion of verbs	S: Verbs
proportion_of_proper_nouns	Proportion of proper nouns	S: Nouns
proportion_of_adverbs	Proportion of adverbs	S: Adverbs
num_entities	Number of entities	NV: Prominence
current_burst_size	Wikipedia current burst size	NV: Prominence
burstiness	Wikipedia burstiness	NV: Prominence
wiki_long_term	Wikipedia long-term prominence	NV: Prominence
wiki_day_before	Wikipedia day-before prominence	NV: Prominence
news_recent	News source recent prominence	NV: Prominence
sentiment	Sentiment	NV: Sentiment
polarity	Polarity	NV: Sentiment
connotations	Connotations	NV: Sentiment
bias	Bias	NV: Sentiment
comparative_superlative	Comparative/superlative	NV: Magnitude
intensifiers	Intensifiers	NV: Magnitude
downtoners	Downtoners	NV: Magnitude
proximity	Proximity	NV: Proximity
surprise	Surprise	NV: Surprise
head_unique	Uniqueness	NV: Uniqueness
columns 38-68	Day of the month	M: Publication date
columns 69-92	Time of the day	M: Publication time
columns 93-219	Category	M: Category

Column headings (feature vectors for baselines):

Unigrams baseline:

Column name	**Description**
article_id	Unique identifier
columns 2-1001	Unigrams

Bandari et al. baseline:

Column name	Description
article_id	Unique identifier
num_entities	Number of entities
max_prom	Maximum prominence
median_prom	Median prominence
bandari_subj	Subjectivity
T_catscore, F_catscore	Category score (separate for T and F)

Arapakis et al. baseline:

Column name	Description
article_id	Unique identifier
num_entities	Number of entities
sum_prom	Sum of prominence scores
number_of_characters	Number of characters
number_of_words	Number of words
proportion_of_nouns	Proportion of nouns
proportion_of_adverbs	Proportion of adverbs
proportion_of_verbs	Proportion of verbs
arapakis_senti	Sentiment
arapakis_pol	Polarity
columns 10-136	Category
columns 137-167	Day of the month
columns 168-191	Time of the day

Column headings (social media popularity):

Column name	Description
article_id	Unique identifier
T	# tweets and retweets after three days
F	# Facebook likes and shares after three days

References:

Arapakis, I., Cambazoglu, B. B., & Lalmas, M. (2014). On the feasibility of predicting news popularity at cold start. In International Conference on Social Informatics (pp. 290-299). Springer International Publishing.

Bandari, R., Asur, S., & Huberman, B. A. (2012). The Pulse of News in Social Media: Forecasting Popularity. In Sixth International AAAI Conference on Weblogs and Social Media.

Related publications:

- Implementation and evaluation of news values feature extraction:

Piotrkowicz, A., Dimitrova, V., & Markert, K. (2017). *Automatic Extraction of News Values from Headline Text*. In Proceedings of EACL.

- Correlations between news values and linguistic style feature values and social media popularity

Piotrkowicz, A., Dimitrova, V.G., Otterbacher, J., and Markert, K. (2017) *The Impact of News Values and Linguistic Style on the Popularity of Headlines on Twitter and Facebook*. In Proceeding of ICWSM News and Public Opinion Workshop.

- Prediction model using news values and style features:

Piotrkowicz, A., Dimitrova, V.G., Otterbacher, J., and Markert, K. (2017) *Headlines Matter: Using Headlines to Predict the Popularity of News Articles on Twitter and Facebook*. In: Proceedings of ICWSM Short Papers.