**Narrative Dataset - Test**

**Sources**

|  |  |  |  |
| --- | --- | --- | --- |
| **Type** | **Description** | **From** | **Link/Paper** |
| World Folktales | Collection of folktales from different cultures | Mostly https://fairytalez.com/ | [Folktales](https://www.kaggle.com/datasets/thedevastator/folk-tales) |
| Books | Plot summaries of books | CMU Book Summary Dataset (extracted from Wikipedia) | Bamman and Smith, 2013  (New Alignment Methods for Discriminative Book Summarization) |
| Reddit | Responses to real-life story prompts on Reddit | r/AskReddit | Ouyang et al, 2015 (Modeling Reportable Events as Turning Points in Narrative) |
| Global News | News articles from October-November 2023 | Various sources (collected via NewsAPI) | [Global News](https://www.kaggle.com/datasets/everydaycodings/global-news-dataset/data) |
| Movies/TV | Plot summaries of movies and TV episodes | Wikipedia, IMDb, fandom.com, thetvdb.com, themoviedb.org | Zhao et al, 2022 (NARRASUM: A Large-Scale Dataset for Abstractive Narrative Summarization) |

Cleaning:

* Length requirement:
  + Min: 200 words, Max: 2000 words
* Duplicates removed
* Folktales
  + Explanatory and promotional notes are removed
  + Formatting fixes
* News
  + “Politics” articles only
  + News agency names are all replaced by the generic word “News Agency”

**General**

Number of documents: 112

|  |  |
| --- | --- |
|  | **Number of Words** |
| **mean** | 798 |
| **std** | 492 |
| **min** | 200 |
| **25%** | 326 |
| **50%** | 661 |
| **75%** | 1145 |

**Genre Breakdown**

Number of documents

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Type** | **Book** | **Folktale** | **Movies-TV** | **News** | **Reddit** |
| **count** | 16 | 32 | 16 | 32 | 16 |

Distribution on number of words

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Type** | **Book** | **Folktale** | **Movies-TV** | **News** | **Reddit** |
| **mean** | 630 | 952 | 709 | 992 | 361 |
| **std** | 391 | 553 | 354 | 472 | 223 |
| **min** | 236 | 219 | 200 | 262 | 202 |
| **25%** | 292 | 416 | 424 | 608 | 229 |
| **50%** | 566 | 1054 | 680 | 998 | 274 |
| **75%** | 704 | 1383 | 990 | 1285 | 362 |
| **max** | 1531 | 1943 | 1354 | 1927 | 1032 |

**Folktales and News**

Folktales:

Western: German, French

Eastern: Indian, Japanese

News:

Western: CNN

Eastern: Al Jezeera English

Number of documents

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Subtype** | **Folktale-East** | **Folktale-West** | **News-East** | **News-West** |
| **count** | 16 | 16 | 16 | 16 |

Distribution on number of words

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **subtype** | **Folktale-East** | **Folktale-West** | **News-East** | **News-West** |
| **mean** | 996 | 909 | 972 | 1011 |
| **std** | 371 | 701 | 450 | 507 |
| **min** | 429 | 219 | 270 | 262 |
| **25%** | 703 | 255 | 538 | 640 |
| **50%** | 1094 | 703 | 1046 | 880 |
| **75%** | 1191 | 1482 | 1288 | 1281 |
| **max** | 1721 | 1943 | 1830 | 1927 |