1. AI4Bharat IndicNLP Corpus (tokenized multilingual Indic text):

* GitHub (main repo, no direct full archive release; download by cloning repo and following scripts to acquire corpora):  
  <https://github.com/AI4Bharat/indicnlp_corpus>
* IndicNLP resources page with dataset references:  
  <https://indicnlp.ai4bharat.org/pages/indicnlp-corpus/>

1. IndicNER Dataset (multilingual NER dataset for 11 Indic languages):

* Hugging Face dataset page with data files, usage instructions, and download:  
  <https://huggingface.co/datasets/ai4bharat/IndicNER>

1. Samanantar Indic-English Parallel Corpus (49.7 million sentence pairs):

* Hugging Face dataset page for download:  
  <https://huggingface.co/datasets/ai4bharat/samanantar>

1. Government eGazette:

* Access and download Gazette notifications (no direct bulk dataset, documents downloadable individually or via API):  
  [https://egazette.gov.in](https://egazette.gov.in/)

1. Digital India Land Records Modernization Programme (DILRMP):

* Dataset page for guidelines and data (no direct bulk link, available by official request or portals):  
  <https://data.gov.in/dataset-group-name/Digital%20India%20Land%20Records%20Modernization%20Programme%20DILRMP>

1. PM-KISAN Scheme Data:

* Open Government Data portal listing PM-KISAN datasets (some direct downloads available):  
  <https://www.data.gov.in/keywords/pm-kisan>

1. Jal Jeevan Mission Data:

* Official site with reports and data references (data download limited, mostly reports and dashboards):  
  [https://jaljeevanmission.gov.in](https://jaljeevanmission.gov.in/)

1. MGNREGA Open Data:

* Public Data Portal with downloadable reports and datasets:  
  <https://mnregaweb4.nic.in/netnrega/dynamic2/dynamicreport_new4.aspx>

1. NLTK Corpora:

* Download via Python NLTK library:  
  <https://www.nltk.org/data.html>