

# Bits and Chunks of Information

Ashirbad Sarangi

The main purpose of this document is to keep the record of all the progress done along with date and what information is taken from which place which can be ultimately used in the report writing while submitting the project.

**Date : Mar 19, 2024**

## 1.1 Linguistic Data Consortium - LDC

The Linguistic Data Consortium (LDC) is an open consortium of universities, libraries, corporations and government research laboratories. LDC was formed in 1992 to address the critical data shortage then facing language technology research and development. The Advanced Research Projects Agency provided seed funding for the Consortium and the National Science Foundation provided additional support via Grant IRI-9528587 from the Information and Intelligent Systems division.

Initially, LDC's primary role was as a repository and distribution point for language resources. Since that time, and with the help of its members, LDC has grown into an organization that creates and distributes a wide array of language resources. LDC also supports sponsored research programs and language-based technology evaluations by providing resources and contributing organizational expertise.

LDC is hosted by the University of Pennsylvania and is a center within the University's School of Arts and Sciences. LDC's connection with Penn provides a strong foundation for the Consortium's research and outreach to an active and diverse member community.<sup>1</sup>

## 1.2 Natural Language Toolkit

NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum.

Thanks to a hands-on guide introducing programming fundamentals alongside topics in computational linguistics, plus comprehensive API documentation, NLTK is suitable for linguists, engineers, students, educators, researchers, and industry users alike. NLTK is available for Windows, Mac OS X, and Linux. Best of all, NLTK is a free, open source, community-driven project. //NLTK has been called "a wonderful tool for teaching, and working in, computational linguistics using Python," and "an amazing library to play with natural language."<sup>2</sup>

---

<sup>1</sup><https://www.ldc.upenn.edu/about>

<sup>2</sup><https://www.nltk.org/>