Introduction to NLTK
Lab Session II

Dr. JASMEET SINGH

ASSISTANT PROFESSOR, CSED

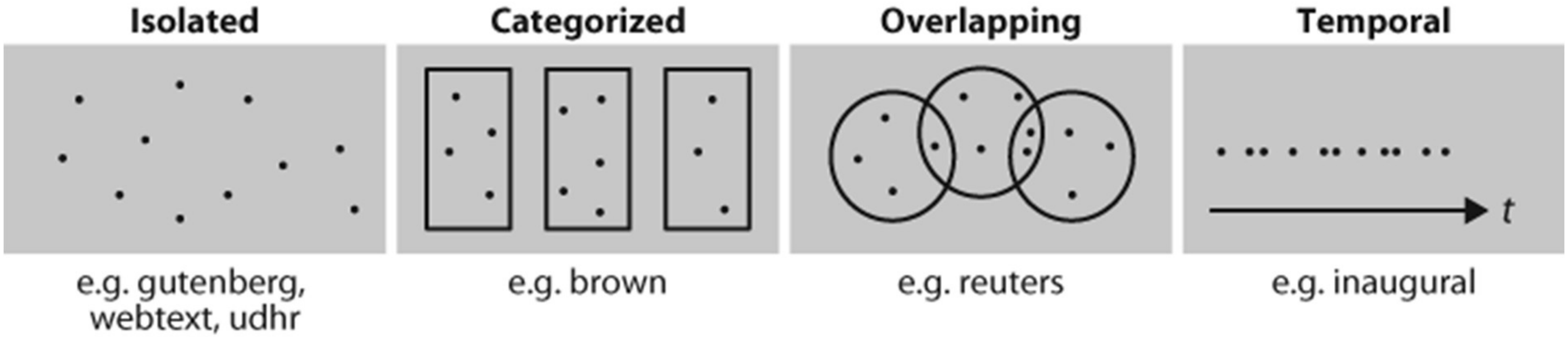
TIET, PATIALA

Introduction to Textual Data

- Textual Machine Learning models typically uses large bodies of linguistic data, or **corpora.**

- A **computer corpus** is a large body of machine-readable texts.

Introduction to Textual Data Contd....



The simplest kind of corpus is a collection of isolated texts with no particular organization;

some corpora are structured into categories, such as genre (Brown Corpus);

some categorizations overlap, such as topic categories (Reuters Corpus);

Other corpora represent language use over time (Inaugural Address Corpus)

Python NLTK

- NLTK was originally created in 2001 as part of a computational linguistics course in the Department of Computer and Information Science at the University of Pennsylvania.
- Since then it has been developed and expanded with the help of dozens of contributors.
- NLTK includes extensive software, data, and documentation, all freely downloadable from http://www.nltk.org/.
- Installing NLTK: pip install nltk (on terminal)
- Importing NLTK in python: import nltk

In-built Corpus in NLTK

- NLTK provides variety of in built corpus and lexical resources.
  - Isolated: gutenberg, Web and chat corpus
  - Categorical: brown
  - Overlapping: reuters
  - Temporal: inaugural
  - Lexical Resources: stopwords, wordlists, names, wordnet
- All these resources can be imported in python by installing *corpus* library in nltk using nltk.download().
- After installing corpus package, we can import it as:
  - from nltk.corpus import * (OR) from nltk.corpus import gutenberg

In-built functions with Corpus

- corpus_name.fileids()- gives file identifiers
- corpus_name.raw()- the contents of the file without any linguistic processing.
- corpus_name.words()-divides the text up into its words
- corpus_name.sents()-divides the text up into its sentences where each sentence is a list of words.
- raw(fileids=[f1,f2,f3]), words(fileids=[f1,f2,f3]), sents(fileids=[f1,f2,f3]) gives the text, words, and sentences respectively in the specified file ids.