REPORT

Group Members –
Soutik Gayen – 19EC10081
Subhajyoti Ghosh – 19EC10061
Rajat Rathi – 19IE10041
Sumeet Bohra – 19CS10059
Vishesh Anand – 19IE10042

Dataset generation code –

- “sentence_generation.py” is used to generate the “.sent” and “.pointer” files from thexml files for both Hindi and Bengali.
- The xml trees are parsed using “xml.etree.ElementTree” library.
- Each line corresponds to a “<P>” tag in the xml tree.
- All <W> tags in each <P> tag constitute the sentences in “.sent” file.
- There are two <unknown> tags in the beginning of each sentence to account foreventswith no event argument.
- The xml tags with no “<LINK>” tags in the subtree are the “EVENT”s
- The rest of the xml tags (except ‘<W>’ of course) are the “EVENT ARGUMENTS”.
- The Event arguments are linked to the Events through the “EVENT_ARG” attribute ofthe arguments tag and the “ID” attribute of the Event tag .
- The sentences with no Events have a blank line in the “.pointer” file.

Github link - https://github.com/rajatrathi31/Event-Extractor