CS537 Natural Language Processing

Autumn 2015

Assignment

Each student is to build a text corpus of at least 10,000 words in any of the following languages-

1. Assamese 2. Hindi 3. Bengali 4. Thai
5. Any South Asian Language other than English.

The corpus should be in Unicode. While the major portion of the corpus may be collected from various sources in public domain, at least a small part (approx. 500 words) should be typed by the student.

Tag at least a 1000 words segment of the corpus with PoS tags. You may collect an already tagged corpus, but you must tag at least 500 words of previously untagged portion.

Describe the PoS tag-set used. Mention one-or-two other tag-sets that are there and state how is your tag-set different.

[Submit by 26-10-2015].