Artificial Intelligence-Tutorial

CSN – 371

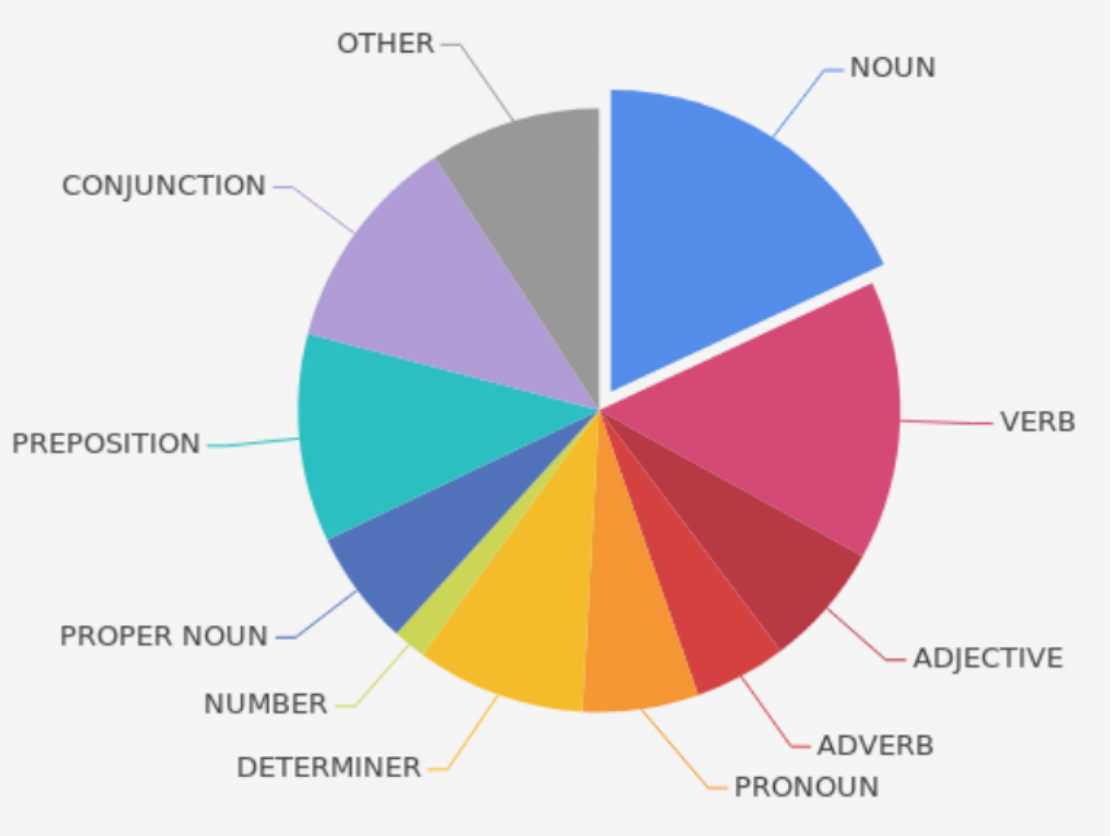
Instructor: Raksha Sharma

BNC Corpus

- The British National Corpus (BNC) was originally created by Oxford University press in the 1980s - early 1990s, and it contains 100 million words of texts from a wide range of genres (e.g. spoken, fiction, magazines, newspapers, and academic).
- The corpus covers British English with the intention that it be a representative sample of spoken and written British English of that time.

BNC Corpus: Tag Distribution

Noun: 18%
Verb: 15%
Adj: 6%
Adv: 5%
Pronoun: 6.22%
Det: 10%
Number: 1.76%
ProperNoun: 6.2%
Prepositon: 11%
Conjunction: 11%
Other: 9.20%



BNC Corpus: Tag Distribution

- A tagset is a list of part-of-speech tags (POS tags for short), i.e. labels used to indicate the part of speech and sometimes also other grammatical categories (case, tense etc.) of each token in a text corpus.

- When creating user corpora, the recommended tagset is always preselected.

- BNC corpus is tagged with Penn tagset (File is available with the assignment).

Assignment

- Bring the corpus into required format. (preprocessing)

word1_tag word2_tag ..... (week-1)

- Create a dictionary having entry for unique word+tag combination with it's frequency count in the corpus. (week-2)

- Report top 10 frequently used words and 10 frequently used tags. Provide your analysis of the word and tag distribution in the corpus. (week-3)

- For each word, compute probabilities of word associations with tags. Program should be able to display probability of each word given the tag for the training corpus. (week-4)

- Predict the new tags for the words in the test corpus. (week-5)

- Generate confusion matrix for the word-tag pair. (week-6)

- Implement Baysian Network to predict POS tag. (week-7-8-9)

- Report accuracy over test data. (week-10).