UNIVERSITÄT WIEN
CSLEARN - EDUCATIONAL TECHNOLOGIES

Natural Language Processing

Exercise Sheet 5

Categorizing and Tagging Words

Exercise 1

Produce a sorted list of tags used in the Brown corpus, removing duplicates. Do the same for the universal part-of-speech tagset.

Exercise 2

Write a program to process the Brown Corpus using the universal part-of-speech tagset to find out which nouns are more common in their plural form than in their singular form. Only consider regular plurals formed with the "-s" suffix. Print an alphabetically sorted list of the nouns together with the frequencies for the singular and plural forms, one per line.

Exercise 3

Find out which word has the greatest number of distinct tags in the Brown corpus using the original tagset. Without using the `most_common` function, print a list of the tags together with the frequencies for the word, sorted by frequency from highest to lowest, one per line.

Exercise 4

Tabulate the frequencies of the universal tags that precede nouns in the Brown Corpus.

Exercise 5

Write a function `ambiguous(tagged_text)` that returns the number of ambiguous word types as well as the number of all word types in a tagged text. A word type is ambiguous if it is tagged with at least two different tags. Use the function to print both values as well as the percentage of ambiguous word types for the Brown Corpus both for the original and the universal tagset.

Exercise 6

Write code to search the Brown Corpus to answer the following questions:

a) produce an alphabetically sorted list of the distinct words tagged as `MD`;

b) identify words that can be plural nouns or third person singular verbs;

c) print an alphabetically sorted list of distinct three-word prepositional phrases of the form `IN+AT+NN`, separated by semicolons.

Exercise 7

Write a function `prec_adv(word, text)` that returns an alphabetically sorted list of distinct adverbs that precede `word` in `text`. Use this function to find out which adverbs precede the words “love”, “like”, and “prefer” in the Brown corpus.