T-61.5020 Statistical Natural Language Processing

Exercises 7 — Word sense disabiguation
Version 1.0

1. According to Bayes' theorem, the probability of the sense s_k, if we know the context c, is

$$P(s_k|c) = \frac{P(c|s_k)P(s_k)}{P(c)}$$

Naive Bayes method is based on the assumption that the probabilities of the words in a context are independent. Deduce equation for Naive Bayes classifier.

Note: The next two problems require some knowledge of Finnish.

2. You have been given two groups of sentences. Use those to train a Naive Bayes classifier that separates the two senses of word "sataa" ("rain" or "hundred") from the test set. You have a tool that changes all the other numbers to form "num" — but of course cannot do that for the abiguous word "sataa".
Hint: It is worth to use additive smoothing for the estimates. Apply a prior that all words are equally probable in every context ($P(w_j|s_k) = \frac{1}{N}$). Use a made-up coefficient λ that represents the strength of the prior assumption and add it to the ML estimate. You can assume that the model recognizes only the words in training and test sets (total $N = 85$).
Hint 2: Calculate only those parameters that are needed to recognize the test set.

Group 1:

Tuulet puhaltavat edelleen noin 100 kilometrin tuntivauhtia ja paikoin **sataa** rankasti

Washingtonin alueella **sataa** lunta tai keli on muuten huono

Meille **sataa** mannaa taivaasta

Jos kilpailun aikana **sataa** vettä

Joulun aikana täytyy **sataa** lunta 20-30 senttiä

Joulunpyhinä **sataa** runsaammin lunta, Lounais-Suomessa mahdollisesti räntää sekä vettä

Group 2:

Kapinoivat vangit pitivät edelleen maanantaina noin **sataa** vanginvartijaa panttivankeinaa

Kaikkiaan paikalla oli noin pari **sataa** virkavallan edustajaa

Poliisi pidätti pari **sataa** ihmistä

Räjähdyksissä kuoli noin kaksi **sataa** ihmistä

Lusin pohjoispuolella saa ajaa **sataa** viiden kilometrin matkalla

Talvimyrsky irrotti Stora-Enson Kotkan sahan kattoa monta **sataa** neliömetriä

Test set:

> Koirasusitarhassa vieraili pari **sataa** ihmistä.
>
> Pohjois-Suomessa räntää tai lunta **sataa** keskimäärin joka kolmantena vappuna
>
> Itse tapahtumaan odotetaan noin kuutta **sataa** vierasta
>
> Pommeja **sataa** kaikkialla eikä kaduilla ole ketään

3. You have the following quotations for a Finnish dictionary:

   > ammunta: (1) Tilanne, jossa harjoitellaan aseiden käyttöä. Esim.
   > *Joukkue harjoitteli konepistoolilla sarjatulen ammuntaa.*
   > *Ammunta on aiheuttanut varusmiehille kuulovaurioita*
   >
   > ammunta: (2) Nautakarjasta lähtevä äännähtely. Esim.
   > *Niityltä oli kuulunut lehmän ammuntaa.*
   > *Ammunnan hälyttämä naapurin isäntä löysi nälkiintyneen vasikan.*
   >
   > varusmies: asepalvelusta suorittava kansalainen
   >
   > loukkaantua (1): pahastua
   > *Hän pahastua kovasti siitä, mitä kuuli.*
   >
   > loukkaantua (2): satuttaa itsensä
   > *Hän loukkaantui törmäyksessä.*
   >
   > kivääri: ampuma-ase, jota usein käytetään isomman riistan, esim. hirven metsästyksessä. *rynnäkkö˜*, sotaa varten kehitetty versio, joka pystyy ampumaan sarjatulta.
   >
   > harjoitella: toistaa usein jotain harrastusta oppiakseen paremmin suoriutumaan siitä.
   > *Hän harjoitteli joka päivä pianon soittoa.*

   Let's consider the following sentence: "Varusmies loukkaantui harjoitellessaan kiväärillä ammuntaa niityllä." Use the information from the dictionary to disambiguate the word "ammunta" in the sentence with Lesk's algorithm[1]. Assume that you have a tool for stemming the words.

Note: The next three problems require usage of a computer.

4. You have English material (e.g. Google, http://www.google.com) available and want to know

   a) the meaning of the word *kallistua* in the sentence *"Hinnat kallistuivat"*. From a dictionary you have the facts given in Table 1.

   b) whether the words *potkia, sorkkia, maksaa* and *kärsiä* in the sentence *"Haluatko potkia, sorkkia, maksaa vai kärsiä ?"* are verbs or products of the butcher's.

[1]Michael Lesk, 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In Proceedings of the 1986 SIGDOC Conference, pp. 24–26, New York, Association for Computing Machinery. `http://portal.acm.org/citation.cfm?id=318728&dl=GUIDE,`

Word	Translation
hinta	price
kallistua	s_1 : slant, lean, lurch s_2 : go up
haluta	want, like, desire, covet
potkia	kick
potka	shin
sorkkia	poke, prod, fiddle
sorkka	hoof
maksaa	cost, pay
maksa	liver
vai	or
kärsiä	suffer, ache, sustain
kärsä	snout

Table 1: *Samples from a dictionary*

5. Word "kuusi" has occurred in the contexts given in the list below. We know that it has two meanings ("six" and "spruce"). Classify the contexts to two groups according to in which sense "kuusi" was, using the expectation-maximization (EM) algorithm.

    yksi kaksi kolme (one two three)

    kaksi kolme neljä (two three four)

    neljä viisi seitsemän (four five seven)

    mänty leppä haapa (pine alder aspen)

    leppä haapa koivu (alder aspen birch)

    haapa koivu kataja (aspen birch juniper)

    koivu kataja leppä (birch juniper alder)

    yksi mänty kaksi haapa leppä (one pine two aspen alder)

    yksi haapa seitsemän kahdeksan (one aspen seven eight)

    kaksi haapa (two aspen)

6. Find a text corpus and make a program that does unsupervised word sense disambiguation. You can either use a real ambiguous word, or make up your own pseudoword (e.g. change all words 'rain' and 'commission' in the corpus to word 'raincommission' and study its ambiguity).