LIN 353C: Introduction to Computational Linguistics, Spring 2017, Erk

Homework 4: Word meaning

Due: Thursday March 2, 2017

This homework comes with two files, `lastname_firstname_hw4.txt` and `lastname_firstname_hw4prob1.csv`. These are (basically empty) files, called a *stub files*. Please record your answers for the "what sense is this" part of problem 1 in the CSV file. See more instructions on this below. Please record your answers for all other problems in the appropriate places of the .txt file. Please make sure that you save the file in *plain text*, not something like the Word format.

For submission, please rename the files such that they reflect your name. That is, if your name was Haskell Curry, you would rename them to

> `curry_haskell_hw4.txt` and
>
> `curry_haskell_hw4prob1.csv`

Please also remember to put your name and EID in the file `lastname_firstname_hw4.txt`

Please submit the homework electronically on Canvas.

If any of these instructions do not make sense to you, please get in touch with the instructor right away.

A perfect solution to this homework will be worth *100* points.

1. **Doing word sense assignment by hand (60 points)**

   This problem asks you to do some manual word sense annotation, and to comment on your observations. Please record your annotation in the file `lastname_firstname_hw4prob1.csv`. Please record your comments in the file `lastname_firstname_hw4.txt`.

   Data. Below, you find sentences from the SemCor corpus, all featuring the verb *leave*. Look up the WordNet senses of the verb *leave* at

   `http://wordnet.princeton.edu/`

   using the option "Use WordNet online" in the menu on the left. (Use the senses for the verb *leave*. Do not use senses listed for the noun *leave*.) For each of the sentences below, choose *one* best fitting WordNet sense for the verb *leave* in that sentence, and record your choice. On the "Use WordNet online" webpage, click on *Display Options* and choose *Show Sense Numbers* to see the sense numbers that you can use to record your annotation choice.

   There are 14 senses, `leave#1` through `leave#14`. Please only record the sense number in the CSV file, omitting the `leave#`.

   (a) But questions with which committee members taunted bankers appearing as witnesses **left** little doubt that they will recommend passage of it .

   (b) The departure of the Giants and the Dodgers to California **left** New York with only the Yankees .

   (c) After the coach listed all the boy 's faults , Hartweger said , " Coach before I **leave** here , you 'll get to like me " .

   (d) R. H. S. Crossman , M.P. , writing in The Manchester Guardian , states that departures from West Berlin are now running at the rate not of 700 , but of 1700 a week , and applications to **leave** have risen to 1900 a week .

   (e) The house has been swept so clean that contemporary man has been **left** with no means , or at best with wholly inadequate means , for dealing with his experience of spirit .

   (f) A second and also good practice is to shear off the tops , **leaving** an inch high stub with just a leaf or two on each branch .

(g) No doubt some experiences vanish so completely as to **leave** no trace on the sleeper 's mind .

(h) He is a widower , his three children are dead , he has no one **left** on earth ; also he is a drunk , and has lost his job on that account .

(i) Piepsam tries to stop him by force , receives a push in the chest from " Life " , and is **left** standing in impotent and growing rage , while a crowd begins to gather .

(j) The audience **leaves** the play under a spell , It is the kind of spell which the exposure to spirit in its living active manifestation always evokes .

Reflect on your annotation work (and safe your comments in `lastname_firstname_hw4.txt`):

- Which sentences were hard to annotate, and why?
- Which pairs (or groups) of WordNet senses did you find hard to distinguish, and what criteria did you use to distinguish between them?

Recording your answers. The file `lastname_firstname_hw4prob1.csv` is a "comma-separated values" file. You can open it in Excel to record your answers. Or you can open it as a plain-text file in your favorite text editor. If you do the latter, please make sure to save the file in plain-text mode. If you open the file as a plain-text file, you will see that it consists of comma-separated values (as the name CSV suggests):

```
Sentence,Sense no. (1-14)
a,
b,
d,
...
```

For each sentence number, record the sense for that occurrence of the target word behind the comma. Please make sure to retain the formatting, as we will read your answers automatically.

Incidentally, CSV files are a very nice data exchange format, as Excel can read and write this format, and Python can, too. See `https://docs.python.org/3.6/library/csv.html`.

2. **Word sense disambiguation using Naive Bayes (40 points)**

The file `wsd_train.txt` distributed with this homework contains word sense annotation data for the lemma *feel*. There is data for two WordNet senses of feel:

- "experience": to undergo an emotional sensation or be in a particular state of mind. As in: "She felt resentful"; "He felt regret"
- "think": to come to believe on the basis of emotion, intuitions, or indefinite grounds. As in: "I feel that he doesn't like me"

Each line starts with a word that is the sense label, either EXPERIENCE or THINK. After that comes the sentence. Here is the first line as an example:

```
EXPERIENCE  I began to *feel* some guilt
```

This says that the sense label for this item is EXPERIENCE, and that this item contains the features *I*, *began*, *to*, *some*, *guilt*. The target word is encased in stars so you can omit it from feature counts.

For this problem, you will compute probabilities for a Naive Bayes classifier. For counts of a feature with a sense, $P(f|s)$, count in *how many occurrences of the sense* that feature appears, not how many times that feature appears. That is, if a feature appears twice in the same sentence, only count it once. (This does not mean that you should always do this for this task; it is just simpler for the purpose fo this homework.)

(a) Use the training data in `wsd_train.txt` to determine the probability of the sense THINK: Estimate its probability $P(THINK)$ based on corpus counts in `wsd_train.txt`. You can use Python, or do the calculation by hand. In either case, show your work.

(b) One of the features occurring in the training file `wsd_train.txt` is *I*. For this feature, compute its probability given the sense *EXPERIENCE*, that is, compute $P(I|EXPERIENCE)$, based on corpus counts in `wsd_train.txt`. You can use Python, or do the calculation by hand. In either case, show your work.

(c) The file `wsd_test` distributed with this homework contains a test item in the same format as the training items in `wsd_train`, except that the sense is ?. The item is

```
? I *felt* that there was a pain in my leg
```

Use Naive Bayes to compute the most likely sense for this test item. Remember that in Naive Bayes, the most likely sense s is the one for which

$$P(s) \prod_f P(f|s)$$

is largest. (Notation: $\prod_f P(f|s)$ is the product of the conditional probabilities $P(f|s)$ for all features f.) You can use Python, or do the calculation by hand. In either case, show your work.