BBM 497: Introduction to NLP Lab.

ASSIGNMENT III

Handed out: 14.04.2017 *Handed in: 05.05.2017*

Please submit your solution (code and a PDF of your report) by 17:00pm on the due date. Please describe your code in a separate report. Your reports should not exceed a page.

Code and data

All files that are necessary to do the assignment are contained in a zip file which you can get from Piazza.

1 Word Similarity/Word Sense Disambiguation

Task 1: Given a word with senses and a small training set of contexts for each of the senses, apply the Naïve Bayes classifier to find the correct sense of a word in the test set.

You can get both training files and test files from Piazza (Files are taken from the SensEval corpus). Training file is given in the following format:

- All files are simple ASCII.
- Instances are separated by a blank line.
- The first line of the instances contains a reference number (a six-digit number)
- Each sentence is on a new line; the default number of sentences of context is two, but sometimes there are more, sometimes just one.
- The word to be disambiguated is preceded by <tag sensenum> and followed by </>
- The sense number is given as a reference number and the word may have 'suffixes'

Few example instances in a training file for different senses of '*wooden*' are given below:

> 800002
> Head for the old part of Lhasa, near the Jokhang Temple.
> Here the streets wind at will, leading you astray through frozen mud and darkened puddles, past <tag "523270">**wooden**</> gateways with their clutter of cloistered courtyards.
>
> 800097
> If there are messages in pictures I have got the message.
> I tell you, Ella, that those flat backdrops like posters &dash. deserted dodgems at the seaside or a <tag "523269">**wooden**</> impassive nurse standing beside a Red Cross van &dash. remind me of the bureaucratic life.

Your task is to tag the given words (e.g. '*wooden*') in different contexts in the given test file by using Naïve Bayes Classifier. (You must implement the Naïve Bayes Classifier)

Test files will be given in the following format:

- All files are simple ASCII.
- Instances are separated by a blank line.
- The first line of the instances contains a reference number (a six-digit number)
- Each sentence is on a new line; the default number of sentences of context is two, but sometimes there are more, sometimes just one.
- The word to be disambiguated is preceded by <tag> and followed by </>

Example instances of the word ‘*wooden*’ to be tagged with one of the senses are given below:

```
700010
Mrs Pat Mitchell, a parish councillor, is raising around #l0,000 to create a new adventure
playground at one end of the playing fields at Launton Sports and Social Club.
She organised the craft fair at the club on Sunday (March 25) and is pictured (left) with one of the
exhibitors Mrs Dawn Chambers, of Chestnut Close, Bicester, who makes painted <tag>wooden</>
door plates and numbers.

700049
But while Mr Dinkins has had death threats made against him for condemning the anti-Semitic
black Muslim leader, Louis Farrakhan, his eloquent defence of Jewish issues appears to be
counting for less and less in the ever tightening race.
Until this week Mr Dinkins was considered a certainty with a 19-point lead over his opponent at
the polls and it seemed that by early next year the first black mayor in the history of New York
would be inaugurated.
All that was before Roger Ailes, the media consultant credited with putting George Bush in the
White House, got down to work improving the rather <tag>wooden</> image of Mr Giuliani.
```

The vocabulary that will be used for building your documents will be given in a file called ‘vocabulary.txt’. For example, for word sense disambiguation of ‘wooden’, the vocabulary file will look like this:

```
door, painted, benches, gateways, courtyards, streets, bench, shaft, fence, chair, chairs, windows,
equipment, rattle, planks, walls, spoon, spoons, shapes, boxes, balls, floor, table, tables, garden,
room, building,  buildings, toy, balconies, material, boards, plank,  barn, stand, gate, arches, gates,
platform,  house, handle, kitchen, impassive, bureaucratic, life, message, pictures, image, figures,
felt, feel, phrase, phrasing, disappointment, excuse, way, frightening, service, character,
guardhouse, mean, dull, emotionless, lifeless, spiritless
```

Use Porter Stemmer for the vocabulary. After Porter Stemmer the vocabulary file will look like:

```
door, paint, bench, gateway, courtyard, street, shaft, fenc, chair, window, equip, rattl, plank, wall,
spoon, shape, box, ball, floor, tabl, garden, room, build, toy, balconi, materi, board, barn, stand,
gate, arch, platform, hous, handl, kitchen, impass, bureaucrat, life, messag, pictur, imag, figur, felt,
feel, phrase, disappoint, excus, way, frighten, servic, charact, guardhous, mean, dull, emotionless,
lifeless, spiritless
```

The documents for each word will consist of the vocabulary words if they exist in the same instance (i.e. the instance given with a reference number). Once you find the right tags for the word tokens in the test set, they will be written in an output file (i.e. output.txt), such as:

```
700010 523270
700049 523269
```

which means the word in the instance 700010 is tagged with the sense id 523270 and the word in the instance 700049 is tagged with the sense id 523269.

Please note that, you will have to use Porter Stemmer (http://tartarus.org/martin/PorterStemmer/) for all the words in all files. Therefore, the word ‘painted’ will be searched as ‘paint’ in the training and in the test file (which are all stemmed as well).

Task: Given two words and a few sentences containing them, your task is to compute their cosine similarity. Use bag-of-words method with a window size of ±3 by assigning frequency values in the feature vector. Please keep in mind that the size of your feature vector will be equal to the number of words in the vocabulary (i.e. all the context words in a window size of 3 for the given target words). Stop words will not be excluded.

Two words are separated by a blank line. The words for which you will calculate the cosine similarity will be delimited by “:”. Only one morphological form of the target words will exist in the file.

An example file is given below:

> cat:
> The **cat** was playing in the garden.
> The owner feed her **cat** every morning.
> You can find **cat** food in the markets.
> The **cat** often eats in the morning.
> They were fighting like a **cat** and a dog.
> How much should I feed my **cat**?
> Her **cat** was always sleeping.
>
> dog:
> The family’s cat and **dog** are playing in the garden.
> Encourage your **dog** to play in the garden.
> **Dog** food is not sold here.
> His **dog** does not eat meat.
> The **dog** was hit by a car.
> I never feed my **dog** raw meat.

You will use Porter Stemmer (http://www.d.umn.edu/~tpederse/data.html) again when creating the feature vector. Therefore, if the words ‘played’ and ‘plays’ both occur within the contexts, only an entry as ‘play’ will be entered in the feature vector.

The cosine similarity between two words will be printed in a file called ‘output.txt’.