CSC 495/693: Natural Language Processing, Spring 2024 Assignment 1

Assigned Tuesday January 30, due Monday February 12. Max points: 100.

In this assignment, you will practice text prepossessing techniques, explore topic modeling methods, and practice neural network basics.

Programming Requirements

- Dataset: You have been provided with 1000 movie comments as the dataset "comments1k.zip".
- Programming Language: Please use Python3, not Python2.
- Coding Style: define a function for each question.

1. Text Prepossessing (35 points)

Given a NLP dataset, we would like to first analyze it and prepare it for downstream applications through text prepossessing techniques. Use Spacy, NLTK or other related python libraries to finish the following tasks.

1) Split comments into sentences and report the average number of sentences per comment.
2) Do tokenization for the dataset and report the average number of tokens per comment.
3) Without considering punctuation and stop words, how many words are in each comment on average?
4) Try lemmatization and stemming for the database. What are the differences in the results based on your observation?

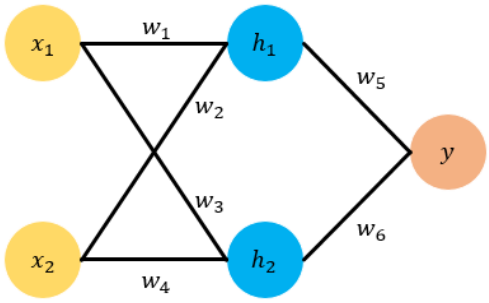
2. Topic Modeling (45 points)

Topic modeling is an unsupervised machine learning technique that's capable of scanning a set of documents, detecting word and phrase patterns within them, and automatically clustering word groups and similar expressions that best characterize a set of documents.

1) Use the Latent Dirichlet Allocation (LDA) method to discover latent topics in the dataset. Try different numbers of topics for LDA. What number of topics do you think is more meaningful?
2) Apply necessary text preprocessing techniques for the dataset. Are the topic modeling results better? If yes, you can use the preprocessed dataset for the following questions.
3) With your topic model, what is the most relevant topic assigned to the document "0_9.txt", "1_7.txt" and "2_9.txt"? Do they make sense? Explain.
4) Any method to better estimate the number of topics? Show it with your experiment.
5) What are the possible limitations of the LDA topic model? (Do not just google it, try to show some of your understanding or explanation.)

3. Machine Learning Basics (20 points)

Neural Network is the most useful machine learning method in the NLP field. The following questions are about how to train a neural network model.



Suppose we designed a neural network with the above structure with x_1 and x_2 as inputs and y as output. h_1 and h_2 are simplified neurons without activation functions (or you can think the activation function is y=x). w_1 to w_6 are parameters. We have:

$$h_1 = w_1 \cdot x_1 + w_2 \cdot x_2$$
$$h_2 = w_3 \cdot x_1 + w_4 \cdot x_2$$
$$y = w_5 \cdot h_1 + w_6 \cdot h_2$$

We use the Backpropagation method to train this network, and let the error $E = \frac{1}{2}(y-t)^2$, where t is the target (or label). If you are given the following dataset with one example:

Data	**x_1**	**x_2**	**t**
Example1	1	0.5	4

and the initialized weights are: $w_1 = 0.5, w_2 = 1.5, w_3 = 2.3, w_4 = 3, w_5 = 1, w_6 = 1$

1) What is the error after one epoch of feed-forward pass?
2) The error is not zero, so we need to update the weights following gradient descent. If we set the learning rate as 0.1, what are the updated weights?
3) With the updated weights, what is the new error? Is the error reduced?

Writeup

Prepare a writeup on your experiments by using any of the following template:

- ACM (https://www.acm.org/publications/proceedings-template/)
- IEEE (https://www.ieee.org/conferences/publishing/templates.html)

Write down any further insights or observations you made while implementing and running the program. Especially interesting insights may be awarded extra points. You may also receive extra points for well-written code with clear comments and runs efficiently. Conversely, poorly written, or not following the ACM/IEEE format, or hard to understand and inefficient code will lose points.

What to turn in

You will turn in:

1. Your writeup, and
2. Your source code. You may include a readme if needed (e.g. if you wish to bring anything to my

attention). Please ensure your code is well documented. **I will not be able to spend a lot of time debugging your code if it crashes during our testing.**

To turn in your code and writeup, use Canvas. Prepare a zip file with all your files and name it <yourname>_assign1.zip. **This zip file should only contain your writeup, source code and readme (if needed) and not executables/object files/data files/unmodified code/anything else, and must be timestamped by the due date to avoid a late penalty.**