DALHOUSIE UNIVERSITY
Inspiring Minds

Fall 2021 (Sep7-Dec7)
Faculty of Computer Science

CSCI 4152/6509 — Natural Language Processing

Assignment 3

Due: *Friday, Nov 19, 2021 by midnight*
Worth: 155 marks (= 18 + 25 + 27 + 25 + 35 + 25)
Instructor: Vlado Keselj, CS bldg 432, 902.494.2893, vlado@cs.dal.ca vlado@dnlp.ca

---

Assignment Instructions:

The submission process for Assignment 3 is mostly based on the `submit-nlp` command on `timberlea` as discussed in the lab, or in the equvalent way by using the course web site, where you need to follow 'Login' and then the 'File Submission' menu option.

Important: You must make sure that your course files on timberlea are **not** readable by other users. For example, if you keep your files in the directory `csci6509` or `csci4152` you can check its permission using the command:

```
ls -ld csci6509
```

or

```
ls -ld csci4152
```

and the output must start with `drwx------`. If it does not, for example if it starts with `drwxr-xr-x` or similar, then the permissions should fixed using the command:

```
chmod 700 csci6509
```

or

```
chmod 700 csci4152
```

1) (18 marks) Complete the Lab 5 (Python NLTK 1) as instructed, and submit files using the command `submit-nlp` as instructed. In particular, you will need to properly:

a) (5 marks) Submit the file '`lab5-list_merge.py`' as instructed.

b) (3 marks) Submit the file '`lab5-stop_word_removal.py`' as instructed.

c) (5 marks) Submit the file '`lab5-explore_corpus.py`' as instructed.

d) (5 marks) Submit the file '`lab5-movie_rev_classifier.py`' as instructed.

2) (25 marks) Complete the Lab 6 as instructed. In particular, you will need to properly:

a) (5 marks) Submit the file '`hmm_tagger.py`' as instructed.

b) (5 marks) Submit the file '`crf_tagger.py`' as instructed.

c) (5 marks) Submit the file '`brill_demo.py`' as instructed.

d) (5 marks) Submit the file '`ne_chunker_exercise.py`' as instructed.

e) (5 marks) Submit the file '`first_notebook.ipynb`' as instructed.

3) (27 marks) Complete the Lab 7 as instructed. In particular, you will need to properly:

a) (6 marks) Submit the files '`lab7-twitter-profiler.py`' and '`lab7-readme.txt`' as instructed.

b) (5 marks) Submit the program '`lab7-tweets.py`' as instructed.

c) (6 marks) Submit the files '`lab7-tweets-csv.py`' and '`lab7-tweets.csv`' as instructed.

d) (10 marks) Submit the file '`lab7-hashtags.py`' as instructed.

4) (25 marks) Submit your answer using the command `nlp-submit` or the equivalent page on the web site either as file `a3q4.txt` `a3q4.pdf`, or `a3q4.jpg`. If you need to submit several pictures, you can use names like `a3q4-1.jpg`, `a3q4-2.jpg`, etc.

Consider the phrase 'to be or not to be'. We will consider an alphabet of 26 lowercase English letters, and the space character used to separate the words, which makes the total vocabulary of 27 characters. Our main task it to create a uni-gram character model; i.e., a Markov Chain model, of this sentence, or in other words, to calculate probability estimates for each character.

a) (7 marks) What are probabilities of the 27 characters based on the given phrase if we do not use smoothing?

b) (8 marks) What are the probabilities if we use Laplace add-one smoothing method?

c) (10 marks) What are the probabilities if we use Witten-Bell smoothing method?

5) (35 marks) Submit your solution as a file named `a3q5.txt` or `a3q5.pdf` using the command `submit-nlp`.

Let us assume that you work on a problem of classifying news articles into finance and other classes. After analyzing a set of articles you found that three features D (for 'dollar'), E (for 'economy'), and F (for 'finance') of the articles are particularly useful in your classification task.

You decided to work on creating a small Naïve Bayes classifier to classify articles into two classes. In summary, your model uses the following variables:

- The variable $D \in \{t, f\}$ is set to 't' (true) if the word 'dollar' is present in the article, otherwise it is set to 'f' (false).

- The variable $E \in \{t, f\}$ is set to 't' (true) if the word 'economy' is present in the article, otherwise it is set to 'f' (false).
- The variable $F \in \{t, f\}$ is set to 't' (true) if the word 'finance' is present in the article, otherwise it is set to 'f' (false).
- The class variable $C \in \{t, f\}$ is set to 't' (true) if the article is in the financial domain, and 'f' (false) otherwise.

The training data is presented in the following table:

articles	D	E	F	C
10	t	t	t	t
1	t	t	t	f
27	t	t	f	t
8	t	t	f	f
13	t	f	t	t
1	t	f	t	f
9	t	f	f	t
33	t	f	f	f
1	f	t	t	t
3	f	t	f	t
14	f	t	f	f
5	f	f	t	t
6	f	f	t	f
1	f	f	f	t
68	f	f	f	f

a) (15 marks) Calculate the conditional probability tables (CPTs) for the Naïve Bayes model.

b) (5 marks) Calculate $P(C=t \mid D=f, E=t, F=t)$ using the Naïve Bayes model and briefly describe what this conditional probability represents.

c) (5 marks) What is the most likely value of the class variable C for the partial configuration $(D=f, E=t, F=t)$ according to the Naïve Bayes model discussed in a) and b)?

d) (5 marks) What is $\mathrm{P}(C=t \mid D=f, E=t, F=t)$ if we use the Joint Distribution Model?

e) (5 marks) What is $P(C=t \mid D=f, E=t, F=t)$ if we use the Fully Independent Model?

Note: In assignments, always include intermediate results and sufficient details about the way the results are obtained.

6) (25 marks, programming) Write and submit a program written in Perl, Python, C, C++, or Java, named `a3q6.pl`, `a3q6.py`, `a3q6.c`, `a3q6.cc`, or `a3q6.java`, which calculates the edit distance between two strings as done in class.

The program must read the standard input, calculate edit distance between each two lines, and print each such result on a separate line of the output. The program will stop when it reads the first line and it contains just the word '`END`' exactly, and a new-line character at the end.

For each two lines that the program reads, the first line will contain the 'source' string and the second line will contain the 'target' string, as discussed in class. Your program must remove the trailing new-line character of each line and any trailing white-space characters. It should then apply the edit distance algorithm, as shown in class, the one with costs 1 for insert, delete, and substitution of different characters. Finally, the program should print the edit distance by itself on a line in the output.

The following is an example of sample input:

```
there
ythre
aaaa
bbb
A simple sentence.
A sentence simple.
Sunday
Saturday
sunday
Saturday
SUNDAY
saturday
intention
execution
END
```

and the following is the sample output:

```
2
4
12
3
4
8
5
```

These files are given in the assignment folder on the course web site as the files `a3q6-test1.in` and `a3q6-test1.out`, and on `timberlea` in the directory: `~prof6509/public/a3`