SE4475 Introduction to NLP – Assignment
Lecturer: Assoc. Prof. Dr. Mete EMİNAĞAOĞLU

Submission Deadline: 29th of November, Friday at 22:00
(If you don't deliver your assignment on time, you will get 0)

You must deliver your report, and all of the source code of your program in a zipped file by uploading to this course's Moodle web portal's assignment section.

Assignment Materials, Methods & Rules:

You will use the data named as "Assignment-data.rar" for multi-class text classification by using the methods and constraints explained in the following sections.

- In this .rar file, there is a "Raw_texts" folder and there are 3 different folders under this folder. There are a total of 3000 different tweet documents belonging to 3 different classes. These tweets are collected from the Internet. The tweets are written in Turkish, but may contain words of foreign origin, as well as abbreviations not defined in Turkish grammar rules, etc.
- Text classification will be done for 3 different classes where the tweet documents of each class are kept in a separate folder and the relevant tweet classes are as follows:
  1- Positive Tweets (a total of 756 records)
  2- Negative Tweets (a total of 1287 records)
  3- Neutral Tweets (a total of 957 records)

- ○ You must do some tokenization, stemming (Zemberek, for instance), conversion to lower-case operations on these tweets. Not a must, but you can also discard the Turkish stop-words.
- ○ Then, you must convert the remaining words / terms to their tf-idf values. You must also include a report in your assignment that shows these tf-idf values (.csv format). An example of the report is given as below. **If you don't deliver this report, you will be punished with a -20 from this assignment's grade.**

	a1	a2	a3	...	...	an	Class
Doc1	0	0	2.68	0	0	0	1
Doc2	0	1.24	0	0	3.567	0.88	3
Doc3	1.78	0	1.12	4.77	0	0	1
…	…	…	…	…	…	…	…
Doc3000	0	1.78	0	0	0	0	2

- ○ You must use k-NN (k-nearest neighbors) for classification. You must use Cosine Similarity within k-NN. In addition, if you like, you can also try other similarity metrics like Euclidean distance, Pearson cc, etc.
- ○ **You cannot use any other algorithm / method / model other than tf-idf and k-NN.**
- You will use **stratified 10-folds cross-validation** to measure the performance of the algorithm.
- You should also try k-NN for different values (k=1,2,3…n) in order to obtain and observe different performance measures.
- The results of the **best performance** measures must also be delivered in another report as follows:

Best results of k-NN obtained by:	k = …, similarity metric: ….				
	Class 1	Class 2	Class 3	MACRO Average	Micro Average
Precision					
Recall					
F-Score					
Total no. of True Positive records					
Total no. of False Positive records					
Total no. of False Negative records					

If you don’t deliver this report, you will be punished with a -25 from this assignment’s grade. Attention: You cannot use any built-in library, function, etc. for this part (calculation of performance measures). If you use such, you will be punished with a -30 from this assignment’s grade. *However, you can use any built-in library, function, etc. for all of the other parts of this assignment.*

Students can use any of the programming languages given below (but you cannot use anything else except the ones given below).
Python, Java, C, C++, C#, .Net

Mandatory Deliverables & Report:

1-All of your source code.
2-Necessary short notes & explanations in the source code as comment lines.
3-Report of performance scores.
4-A comma separated output text file with the tf-idf values.