Универзитет „Св. Кирил и Методиј“ во Скопје
ФАКУЛТЕТ ЗА ИНФОРМАТИЧКИ НАУКИ И КОМПЈУТЕРСКО ИНЖЕНЕРСТВО

06.11.2018

Предмет: **Вовед во препознавање на облици**

Домашна работа 1 – **Bayes classifier**

Задача:

Да се развие едноставен Баесов класификатор за филтрирање на SMS SPAM. Податоците за обука и тестирање може да се преземат од овде (и се придружени и на овој документ) и претставуваат само реорганизација на оригиналните преземени од SMS Spam Collection Data Set. Дадени се како 2 датотеки (SMSSpamTrain.txt и SMSSpamTest.txt) со ист формат, првата намената за обука, втората за тестирање. Колекцијата од SMS пораки е организирана како текстуална датотека во која во секој ред првиот збор ја означува едната од двете класи (“ham” = легална порака или “spam”) по кој следи табулатор (‘\t’) и пораката до крајот на редот. Некои од пораките се и подолги од 140 знаци.

Пример:

ham What you doing?how are you?
ham Ok lar... Joking wif u oni...
ham dun say so early hor... U c already then say...
ham MY NO. IN LUTON 0125698789 RING ME IF UR AROUND! H*
ham Siva is in hostel aha:-.
ham Cos i was out shopping wif darren jus now n i called him 2 ask wat present he wan lor. Then he started guessing who i was wif n he finally guessed darren lor.
spam FreeMsg: Txt: CALL to No: 86888 & claim your reward of 3 hours talk time to use from your phone now! ubscribe6GBP/ mnth inc 3hrs 16 stop?txtStop
spam Sunshine Quiz! Win a super Sony DVD recorder if you canname the capital of Australia? Text MQUIZ to 82277. B
spam URGENT! Your Mobile No 07808726822 was awarded a L2,000 Bonus Caller Prize on 02/09/03! This is our 2nd attempt to contact YOU! Call 0871-872-9758 BOX95QU

Да се развие решение во некој општонаменски програмски јазик (C/C++, Java, C#, …) без употреба на специјални библиотеки за машинско учење или статистика (дозволена употреба на библиотеки за генерици/податочни структури) кое ќе ги чита двете датотеки, ќе го обучи баесовиот класификатор на податоците од правата, а ќе ги класификува пораките од втората (користејќи ги ознаките на класите при тестирањето за одредување на TruePositives, TrueNegatives, FalsePositives и FalseNegatives кои треба да се пријават на крајот.

Изворниот код на решението (C/C++, C#, Java или Python без непотребни датотеки) да биде сместен во директориум со име source. Кус опис на решението со назнаки за начинот на обработка на влезот, на креираните евентуални класи, употребените податочни структури, функции и нивните одговорности, начин на функционирање и употреба, евентуални дополнителни упатства за користената платформа, развојна околина, и сл., како и добиените резултати на тест датотеката да се наведат во посебен документ со име Homework1 или како коментари во изворниот код. Доколку изведете различни експерименти (на пр. со игнорирање на мало/големи букви, игнорирање на одредени знаци, ...) како и различен праг за одлучување дали пораката е SPAM наведете ги сите експерименти и измени. Решението на задачата со изворниот код (во посебен директориум) и документот да бидат ставени во директориум со име Homework1_<indeks>_<Prezime>_<Ime> и спакуван (zip или rar) се поставуваат на Moolde сајтот на курсот не подоцна од назначеното време. Задоцнета домашна може да се испратат на e-mail dejan.gjorgjevikj@finki.ukim.mk не подоцна од 30.12.2018 НО ЌЕ НОСИ ПОМАЛКУ ПОЕНИ.

Рок за испраќање: 18.11.2018 23:54