Python ZH

Határidő dec 14, 19:20 **Pont** 8 **Kérdések** 1
Elérhető dec 14, 17:45 - dec 14, 19:20 körülbelül 2 óra **Időkorlát** 100 perc
Engedélyezett próbálkozások Korlátlan

Instrukciók

A megoldásaidat töltsd fel HTML formátumba (File -> Save and Export Notebook as -> HTML).
Ha nem sikerül a HTML export, akkor .ipynb formátum kell.

Ponthatárok: 1-3.9 (elégtelen), 4-4.9 (elégséges), 5-5.9 (közepes), 6-6.9 (jó), 7-8 (kiváló)

Ezt a kvízt ekkor zárolták: dec 14, 19:20 .

Próbálkozások naplója

	Próbálkozás	Idő	Eredmény
LEGUTOLSÓ	1. próbálkozás	87 perc	0 az összesen elérhető 8 pontból *

* Néhány kérdés még nem lett értékelve

Ezen próbálkozás eredménye: **0** az összesen elérhető 8 pontból *
Beadva ekkor: dec 14, 19:12
Ez a próbálkozás ennyi időt vett igénybe: 87 perc

1. kérdés

Még nincs értékelve / 8 pont

Adatleírás:

Az adathalmaz az IMDB.com weboldalról begyűjtött filmes értékeléseket tartalmaz. A 'review' oszlop tartalmazza az értékelést szöveges formában, a 'sentiment' oszlop pedig azt adja meg, hogy az értékelés pozitív vagy negatív hangvételű-e. A feladat során olyan osztályozó modelleket kell készíteni, amely egy értékeléshez pozitív vagy negatív címkét rendel (forrás: https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews).

Az adatokat a következő linken találod: https://vargadaniel.web.elte.hu/datasets/

Feladatok:

1. Olvasd be az adathalmazt: imdb_2k.csv

2. Előfeldolgozás (1.5 pont)

- Szűrd ki azokat a sorokat, amelyek hiányzó (NULL) értékeket tartalmaznak. (0.5 pont)
- Készíts egy új oszlopot 'text_len' néven, amely az értékelés hosszát tartalmazza. (0.5 pont)
- Szűrd ki azokat a sorokat, ahol az értékelés hossza kevesebb, mint 10 karakterből áll. (0.5 pont)

3. Naive-Bayes osztályozás (2 pont)

- Olyan osztályozó modelleket akarunk készíteni, amelyek egy értékelésről megmondják, hogy az poztív vagy negatív.
- Az osztályozást a 'review' oszlop alapján végezd, az osztálycímkét a 'sentiment' oszlop tartalmazza.
- Bontsd fel az adathalmazt tanító és tesztelő halmazokra 90%-10% arányban. (0.5 pont)
- Készíts két Naive-Bayes osztályozó modellt. Az egyik esetén a szöveges adatot CountVectorizer-el, a másiknál TfidfVectorizer-el alakítsd át. (1 pont)
- Értékeld ki mindkét modellt a teszt halmazon, és add meg a pontosságukat (accuracy). (0.5 pont)

4. KNN osztályozás (2 pont)

- Készíts KNN (k-legközelebbi szomszéd) osztályozókat is TfidfVectorizer használatával.
- Nézd meg a KNN osztályozó pontosságát különböző K paraméterekkel: 1-től 23-ig a páratlan számokat vedd. (2 pont)

5. Kiértékelés (1 pont)

- Add meg, hogy a három modell közül (két Naive-Bayes, legjobb KNN) melyik a legjobb a pontosság alapján. (0.5 pont)
- Használd a legjobb modellt arra, hogy általad megadott szöveget osztályozza. Mutass egy olyan példát, ahol a szöveget pozitívnak és egy olyat ahol negatívnak osztályozza. (0.5 pont)

6. Vizualizáció (1.5 pont)

- A legjobb modellhez készíts Confusion Matrix-ot. (0.5 pont)
- Mutasd meg ábárval hogyan változik a KNN modellek pontossága a K növekedésével (horizontális tengely - K, vertikális tengely - accuracy) (1 pont)

zh(1).html (https://canvas.elte.hu/files/2611413/download)

Kvízeredmény: **0** az összesen elérhető 8 pontból