Princípy dátovej vedy - domáca úloha

Vašou úlohou bude vybudovať predikčný model, ktorý na základe textu tweetu predikuje, či sa daný tweet týka nejakej katastrofy.

Dáta: https://www.kaggle.com/competitions/nlp-getting-started/data
Budeme využívať iba stĺpec text a target (location a keyword ignorujte).

Metrika: https://www.kaggle.com/competitions/nlp-getting-started/overview/evaluation

Do classroomu odovzdajte kód a stručný komentár k riešeniu. Takisto vaše predikcie (z príslušných podčastí) odovzdajte na stránke Kaggle (dajte si pozor, máte iba 5 submitov denne).
Výsledná presnosť modelu nie je veľmi dôležitá, podstatný je korektný postup.

Krok 1: Validačná množina

Súbor train.csv si vhodne rozdeľte na validačnú množinu pre vaše lokálne testovanie a trénovaciu množinu.

Krok 2: Jednoduchý model

Využite CountVectorizer z balíku sklearn, ktorý text premení na vektor v ktorom spočíta počet výskytov každého slova. Dajte si pozor, aby ste slovník vyrábali iba na trénovacej množine. Následne natrénujte jednoduchý predikčný model pomocou logistickej regresie, potom urobte predikcie na validačnej množine a odmerajte F1 skóre na vašej validačnej množine.Nakoniec urobte predikcie na testovacej množine z Kaggle.com a submitnite.

Krok 3: Ladenie modelu - threshold

V kroku 2 sme vyrobili model, ktorej mal veľa zaujímavých vecí prednastavených. Teraz ich postupne upravíme, aby sme dosiahli lepšie výsledky.

Začneme thresholdom predikcie. Logistická regresia predikuje pravdepodobnosť, že výsledok bude 1. Ale keďže predikcie majú byť 0 alebo 1, tak všetko vačšie ako nejaký threshold (štandardne 0.5) sa nastaví na 1 a všetko menšie na 0. Všimnite si, že F1 skóre je pomerne citlivé na posun thresholdu.

Teraz dostante z modelu predikcie pravdepodobnosti (pomocou funkcie predict_proba) a následne nakreslite tieto grafy (na validačnej množine):

- Precision-recall curve (https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_recall_curve.html)
- Závislosť F1 skóre od thresholdu

Na základe týchto grafov vyberte najlepší threshold a urobte nové predikcie na testovacej množine a submitnite.

Krok 4: Ladenie modelu - hyperparametre

CountVectorizer aj LogisticRegression majú hyperparametre, ktoré sa dajú nastaviť.
Budeme nastovať nasledovné:

- CountVectorizer - ngram range. Môžeme počítať nielen výskyt jednotlivých slov, ale aj dvojíc, trojíc slov…
- CountVectorizer - max_features - ako veľký slovník budeme mať
- LogisticRegression - C - tento parameter ovplyvňuje regularizáciu. Čím menšie C, tým viac model preferuje malé váhy a menej sa zameriava na optimalizáciu chyby.

Vašou úlohou je pomocou validačnej množiny vhodne nastaviť tieto parametre (spôsob nechávame na vás).

Následne zoberte najlepšie parametre, odpredikujte na testovacej množine a submitnite.

Bonus: Lepší model

Tu to nechávame na vás. Skúste vyrobiť lepší model na predikciu.