Egy új spamszűrő módszer

Sass Bálint

MTA Nyelvtudományi Intézet, 1068 Budapest, Benczúr u. 33.
joker@nytud.hu

Kulcsszavak spamszűrés, szövegosztályozás, naív bayesi osztályozó

A kéretlen levelek (*spam*ek) jelensége mára az internet egyik legégetőbb problémájává vált. A spamellenes küzdelem egyik formája a szűrés, melynek során a beérkező leveleket két csoportra osztjuk: tartalmuk alapján spamnek vagy rendes levélnek jelöljük meg őket. A spamszűrést így tekinthetjük szövegosztályozási problémának. Bevált szövegosztályozási módszer az ún. naív bayesi osztályozó (NBC): az egyes kategóriákba sorolt példák (tanulókorpusz) alapján felépített nyelvi modell segítségével állapítjuk meg, hogy adott dokumentum melyik kategóriába tartozik. A nyelvi modell itt az egyes kategóriákhoz tartozó szógyakorisági listákat jelenti.

NBC képezi az alapját *Paul Graham* 2002-ben publikált spamszűrő eljárásának [2]. Ennek lényegi többlete, hogy figyelembe veszi a spamszűrés aszimmetrikusságát: egy spam átengedése sokkal kisebb baj, mint egy rendes levél elvesztése.

A módszer előnyei: (1) nagyon jó szűrési teljesítményt biztosít, (2) a szűrő felépítése spam és rendes levelekből álló tanulókorpusz alapján automatikus, (3) időről időre újra betanítható, így adaptálódik, (4) a tanulókorpusz megadásával mindenki maga definiálhatja, hogy mit tart spamnek.

Implementáltam az algoritmust és az elmúlt hat hónapban teszteltem a saját beérkező leveleimen. A pontosság 98.6%, a lefedettség 94.1% volt.

Látjuk, hogy jelen esetben a nyelvi feldolgozás mindössze az emailek tokenizálását és a szóalakok gyakorisági listáinak elkészítését jelentette. Próbálkoztak lemmatizálással vagy a nagyon gyakori szavak elhagyásával, de ez nem hozott lényeges teljesítményjavulást [1]. Úgy tűnik, hogy egy efféle viszonylag egyszerű szövegosztályozási feladat megoldásában a nyelvi feldolgozás szempontjából minimalista hozzáállás célravezető. A kapott algoritmus nyelvfüggetlen, azaz bármilyen nyelvű emailek szűrésére alkalmas.

Hivatkozások

1. Androutsopoulos, I. et al.: An Evaluation of Naïve Bayesian Anti-Spam Filtering. In proceedings of the 11th European Conference on Machine Learning. Workshop on Machine Learning in the New Information Age. (2000) 9–17
http://arxiv.org/PS_cache/cs/pdf/0006/0006013.pdf
2. Graham, P.: A Plan for Spam. (2002)
http://www.paulgraham.com/spam.html