**Prezentarea setului de date**

[SNAP: Web data: Amazon reviews (stanford.edu)](http://snap.stanford.edu/data/web-Amazon.html)

Xiang Zhang, Junbo Zhao, Yann LeCun. [Character-level Convolutional Networks for Text Classification](https://arxiv.org/abs/1509.01626). Advances in Neural Information Processing Systems 28 (NIPS 2015).

**PREZENTARE GENERALĂ**

Setul de date constă in recenzi de pe Amazon. Datele se întind pe o perioadă de 18 ani intre 1995 si 2013, incluzand aproximativ 35 de milioane de entitati. In total setul inglobeaza aproximativ 2,84 miliarde de cuvinte cu o medie de 82 cuvinte pe review.

**DESCRIERE**

Setul de date contine 34,686,770 de recenzii de pe Amazon de la 6,643,669 de utilizatori pentru 2,441,053 de produse. Recenziile includ informații despre produse și utilizatori, evaluări (valori de la 1 la 5) și o recenzie în text.

Subsetul de date utilizat a fost construit luând scorurile 1 și 2 ca fiind negative iar scorurile 4 și 5 ca pozitive. Recenziile cu scorul 3 sunt ignorate. În setul de date, clasa 0 reprezinta recenziile negative iar clasa 2 cele pozitive. Fiecare clasă are 2.000.000 de mostre impartite in 1.600.000 pentru subsetul de antrenament 200.000 pentru subsetul de validare si 200.000 pentru subsetul de testare insumand 4.000.000 de recenzii.

**Uniformitatea datelor :**

Deoarece avem o problema de clasificare iar cele doua clase(pozitiv si negativ) contin un nivel egal de informatie este important sa folosim un set de date echilibrat pentru a nu introduce in model prejudecati(biases).

**Setul de antremanet** este folosit pentru a antrena și a face modelul să învețe. În fiecare epocă datele de antrenament sunt transmise rețelei neuronale în mod repetat, iar modelul continuă să învețe caracteristicile acestora.

**Setul de validare** este utilizat pentru a verifica performanța modelului în timpul antrenamentului. La fiecare epocă modelul este antrenat pe setul de antrenament si evaluat cu setul de validare, acest proces oferă informații care ne ajută să reglam parametrii modelului.

**Setul de testare** este utilizat pentru a testa modelul după finalizarea antrenarii. Acesta trebuie sa urmeaze aceeași distribuție a claselor ca și setul de date de antrenament pentru a putea oferii o imagine de incredere a performantelor modelului.

Setul de antrenament:

Setul de antrenament este constituit din 3.200.000 de recenzii cu o lungime maxima de 255 cuvinte si un numar mediu de 74.16 cuvinte pe recenzie.

Daca impartim subsetul de antrenament in cele 2 clase se poate sesiza o diferenata de 8.3% intre lungimea medie a recenziilor positive(71,18 cuvinte) si lungimea medie a recenziilor negative(77.15 cuvinte)

|  |  |  |  |
| --- | --- | --- | --- |
|  | Numar de cuvinte | Lungime maxima | Lungime medie |
| Total | 237,338,790 | 255 | 74.16 |
| Subset pozitiv | 113,897,346 | 214 | 71.18 |
| Subset negativ | 123,441,444 | 255 | 77.15 |

Setul de validare:

Setul de validare este constituit din 400.000 de recenzii cu o lungime maxima de 219 cuvinte si un numar mediu de 74.17 cuvinte pe recenzie. Fata de setul de antrenament lungimea maxima a scazut cu 14.12% insa cea medie a ramas constanta, indicand posibilitatea prezentei unor recenzii “outliers” in setul de antrenare.

Distributia dinte cele 2 clase a ramas relative identica, subsetul negativ avand o lungime medie cu 8.21% mai mare decat subsetul pozitiv.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Numar de cuvinte | Lungime maxima | Lungime medie |
| Total | 29,668,690 | 219 | 74.17 |
| Subset pozitiv | 14,249,477 | 219 | 71.24 |
| Subset negativ | 15,419,213 | 214 | 77.09 |

Setul de test:

Setul de test contine 400.000 de recenzii cu o lungime maxima de 217 cuvinte si un numar mediu de 74.10 cuvinte pe recenzie. Lungimea maxima este in scadere cu 14.91% fata de setul de antrenament, avand o valoare aproximativ egala cu setul de validare, confirmand prezenta unor “outliers” in setul de antrenare.

Distributia dinte cele 2 clase este consistenta cu cele 2 seturi de date prezentate anterior, subsetul negativ avand o lungime medie cu 8.16% mai mare decat subsetul pozitiv.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Numar de cuvinte | Lungime maxima | Lungime medie |
| Total | 29,642,125 | 217 | 74.10 |
| Subset pozitiv | 14,240,026 | 216 | 71.20 |
| Subset negativ | 15,402,099 | 217 | 77.01 |

TRAIN

Punctuatii: !"#$%&'()\*+,-./:;<=>?@[\]^\_`{|}~

Total: 34,366,892 cuv/punct 6.90

Poz: 16,434,458 cuv/punct 6.93

Neg: 17,932,434 cuv/punct 6.88

StopWords:

Total: 109,307,111 proc: 46.05%

Poz: 52,556,709 proc: 46.14%

Neg: 56,750,402 proc: 45.97%

StopWords neg:

Total: 3,008,576 proc: 1.26 %

Poz: 1,000,979 proc: 0.87%

Neg: 2,007,597 proc: 1.62%

Validare

Punctuatii: !"#$%&'()\*+,-./:;<=>?@[\]^\_`{|}~

Total: 4,295,238 cuv/punct 6.90

Poz: 2,056,272 cuv/punct 6.93

Neg: 2,238,966 cuv/punct 6.88

StopWords:

Total: 13,665,937 proc: 46.06%

Poz: 6,575,294 proc: 46.14%

Neg: 7,090,643 proc: 45.97%

StopWords neg:

Total: 375,440 proc: 1.26 %

Poz: 125,066 proc: 0.87%

Neg: 250,374 proc: 1.62%

TEST

Punctuatii:

Total: 4,292,727 cuv/punct: 6.90

Poz: 2,238,584 cuv/punct: 6.36

Neg: 2,054,143 cuv/punct: 7.49

StopWords:

Total: 13,642,675 proc: 46.02%

Poz: 6,568,011 proc: 46.12%

Neg: 7,074,664 proc: 45.93%

StopWords neg:

Total: 376,163 proc: 1.26%

Poz: 125,503 proc: 0.88%

Neg: 250,660 proc: 1.62%

## **Stopwords**

## “Stopwords” sunt cuvintele care nu oferă nicio informație utilă pentru a determina în ce categorie ar trebui să se încadreze textul. Aceasta se poate datora faptului că nu au sens (prepoziții, conjuncții etc.), sau pentru că sunt prea frecvente în contextul clasificării.

Prin eliminarea acestor cuvinte, am eliminat informațiile de nivel scăzut din text pentru a ne concentra mai mult pe informațiile importante. Cu alte cuvinte, putem spune că ștergerea acestor cuvinte nu va avea niciun impact negativ asupra modelului pe care îl antrenăm.

“Stopwords” definite pentru modelul nostru:

['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've", "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselves', 'he', 'him', 'his', 'himself', 'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'its', 'itself', 'they', 'them', 'their', 'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'that', "that'll", 'these', 'those', 'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having', 'do', 'does', 'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with', 'about', 'against', 'between', 'into', 'through', 'during', 'before', 'after', 'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'over', 'under', 'again', 'further', 'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'both', 'each', 'few', 'more', 'most', 'other', 'some', 'such', 'only', 'own', 'same', 'so', 'than', 'too', 'very', 's', 't', 'can', 'will', 'just', 'should', "should've", 'now', 'd', 'll', 'm', 'o', 're', 've', 'y']

Am decis sa pastrez in text cuvintele cu sens de negatie deoarece in ciuda frecventei lor crescute ele contin un nivel mare de informatie necesara procesului de antrenare a modelului .

“Stopwords” pastrate:

['no', 'nor', 'not', 'ain', 'aren', "aren't", 'couldn', "couldn't", 'didn', "didn't", 'doesn', "doesn't", 'hadn', "hadn't", 'hasn', 'ma', 'mightn', "mightn't", 'mustn', "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn', "shouldn't", 'wasn', "wasn't", 'weren', "weren't", 'won', "won't", 'wouldn', "wouldn't’ , 'don', ‘don't’]

Eliminarea cuvintelor apartinand categoriii “Stopwords” definite mai sus va reduce dimensiunea setului de date, reducând astfel timpul de antrenament datorita numarului mai mic de cuvinte tokenizate implicate in procesul de antrenare.