Generowanie własnych zbiorów

W celu wygenerowania zbiorów z klasami obcymi (artykułami nienależącymi do żadnej z kategorii) stworzone zostały skrypty add_random.py i create_random.py.

Pierwszym etapem tworzenia zbiorów jest stworzenie bazy danych losowych artykułów. Odpowiedzialny jest za to skrypt create_random.py. Jego parametry wywołania to kolejno liczba artykułów do pozyskania i plik, do którego zostaną zapisane. W przypadku podania nazwy pliku zawierającego już jakieś dane, nowe artykuły zostaną do niego dopisane. Pozwala to na przerwanie i wznowienie pobierania.

Artykuły pobierane są z angielskiej wersji Wikipedii, a następnie poddawane operacji usunięcia znaków niebędących literami, zamianie liter na małe, usunięciu słów krótszych niż 3 znaki, usunięciu „Stop Words” (słów nie wpływających na sens zdania) i zastosowaniu algorytmu Porter Stemmer, obcinającego końcówki słów.

W kolejnym etapie pozyskane artykuły dodawane są do zbioru danych przez skrypt add_random.py. Jego parametrami wywołania są kolejno: plik zawierający losowe artykuły, zbiór treningowy i testowy (na podstawie których wygenerowane będą nowe zbiory), liczba losowych artykułów w zbiorze treningowym i testowym, nazwy nowych zbiorów. Plik z losowymi artykułami musi zawierać co najmniej tyle wpisów, ile wynosi suma dodawanych danych treningowych i testowych.

Zbiory danych

W tej sekcji opisane zostały użyte zbiory danych. Każdy ze zbiorów podzielony jest na zbiór uczący i zbiór testowy. Wszystkie zbiory zapisane są w wersji Stemmed – mają usunięte słowa krótsze niż 3 znaki, usunięte „Stop Words” (słowa nie wpływające na sens zdania) i mają obcięte końcówki (przy użyciu Porter's Stemmer). Zbiory 20 Newsgroups i Reuters dostępne są dodatkowo w wersjach niezmodyfikowanych.

20 Newsgroups

Zbiór wpisów z grup dyskusyjnych, podzielony na 20 kategorii.

Klasa	Zbiory treningowe	Zbiory testowe	Razem
alt.atheism	480	319	799
comp.graphics	584	389	973
comp.os.ms-windows.misc	572	394	966
comp.sys.ibm.pc.hardware	590	392	982
comp.sys.mac.hardware	578	385	963
comp.windows.x	593	392	985
misc.forsale	585	390	975
rec.autos	594	395	989
rec.motorcycles	598	398	996

rec.sport.baseball	597	397	994
rec.sport.hockey	600	399	999
sci.crypt	595	396	991
sci.electronics	591	393	984
sci.med	594	396	990
sci.space	593	394	987
soc.religion.christian	598	398	996
talk.politics.guns	545	364	909
talk.politics.mideast	564	376	940
talk.politics.misc	465	310	775
talk.religion.misc	377	251	628
Suma	**11293**	**7528**	**118821**

Wielkość zbioru treningowego: 9,5 MB

Wielkość zbioru testowego: 6,2 MB

Wielkość bazy danych: 36 MB

Wielkość zbioru treningowego (niezmodyfikowany): 16 MB

Wielkość zbioru testowego (niezmodyfikowany): 11 MB

Wielkość bazy danych (niezmodyfikowany): 50 MB

Reuters 21578 R8

Zbiór komunikatów opublikowanych przez agencję prasową Reuters w 1987 roku. Zbiór został podzielony na 8 ogólnych kategorii.

Klasa	Zbiory treningowe	Zbiory testowe	Razem
acq	1596	696	2292
crude	253	121	374
earn	2840	1083	3923
grain	41	10	51
interest	190	81	271
money-fx	206	87	293
ship	108	36	144
trade	251	75	326
Suma	**5485**	**2189**	**7674**

Wielkość zbioru treningowego: 2,2 MB

Wielkość zbioru testowego: 0,8 MB

Wielkość bazy danych: 3,9 MB

Wielkość zbioru treningowego (niezmodyfikowany): 3,2 MB

Wielkość zbioru testowego (niezmodyfikowany): 1,2 MB

Wielkość bazy danych (niezmodyfikowany): 5,5 MB

Reuters 21578 R52

Zbiór komunikatów opublikowanych przez agencję prasową Reuters w 1987 roku. Zbiór został podzielony na 52 szczegółowe kategorie.

Klasa	Zbiory treningowe	Zbiory testowe	Razem
acq	1596	696	2292
alum	31	19	50
bop	22	9	31
carcass	6	5	11
cocoa	46	15	61
coffee	90	22	112
copper	31	13	44
cotton	15	9	24
cpi	54	17	71
cpu	3	1	4
crude	253	121	374
dlr	3	3	6
earn	2840	1083	3923
fuel	4	7	11
gas	10	8	18
gnp	58	15	73
gold	70	20	90
grain	41	10	51
heat	6	4	10
housing	15	2	17
income	7	4	11
instal-debt	5	1	6
interest	190	81	271
ipi	33	11	44
iron-steel	26	12	38

jet	2	1	3
jobs	37	12	49
lead	4	4	8
lei	11	3	14
livestock	13	5	18
lumber	7	4	11
meal-feed	6	1	7
money-fx	206	87	293
money-supply	123	28	151
nat-gas	24	12	36
nickel	3	1	4
orange	13	9	22
pet-chem	13	6	19
platinum	1	2	3
potato	2	3	5
reserves	37	12	49
retail	19	1	20
rubber	31	9	40
ship	108	36	144
strategic-metal	9	6	15
sugar	97	25	122
tea	2	3	5
tin	17	10	27
trade	251	75	326
veg-oil	19	11	30
wpi	14	9	23
zinc	8	5	13
Suma	**6532**	**2568**	**9100**

Wielkość zbioru treningowego: 2,8 MB

Wielkość zbioru testowego: 1 MB

Wielkość bazy danych: 27 MB

Wielkość zbioru treningowego (niezmodyfikowany): 4,1 MB

Wielkość zbioru testowego (niezmodyfikowany): 1,5 MB

Wielkość bazy danych (niezmodyfikowany): 38 MB

Cade

Zawartość brazylijskich stron internetowych pobranych z CADÊ Web Directory.

Klasa	Zbiory treningowe	Zbiory testowe	Razem
01--servicos	5627	2846	8473
02--sociedade	4935	2428	7363
03--lazer	3698	1892	5590
04--informatica	2983	1536	4519
05--saude	2118	1053	3171
06--educacao	1912	944	2856
07--internet	1585	796	2381
08--cultura	1494	643	2137
09--esportes	1277	630	1907
10--noticias	701	381	1082
11--ciencias	569	310	879
12--compras-online	423	202	625
Suma	**27355**	**13661**	**40983**

Wielkość zbioru treningowego: 25 MB

Wielkość zbioru testowego: 12 MB

Wielkość bazy danych: 66 MB

WebKB

Zbiór stron internetowych zgromadzony przez World Wide Knowledge Base.

Klasa	Zbiory treningowe	Zbiory testowe	Razem
project	336	168	504
course	620	310	930
faculty	750	374	1124
student	1097	544	1641
Suma	**2803**	**1396**	**4199**

Wielkość zbioru treningowego: 2,5 MB

Wielkość zbioru testowego: 1,3 MB

Wielkość bazy danych: 1 MB

Własne zbiory

Zbiory zostały wygenerowane na podstawie Reuters 21578 R8 i losowych artykułów pobranych z

Wikipedii. Dostępne są wersje, w których losowe artykuły stanowią kolejno 10%, 20% i 50% artykułów z początkowego zbioru.

Zbiór	Dodane zbiory treningowe	Dodane zbiory testowe	Razem dodanych
10%	539	210	749
20%	1079	419	1498
50%	2695	996	3691

Wielkość zbioru treningowego (10%): 2,4 MB

Wielkość zbioru testowego (10%): 0,9 MB

Wielkość bazy danych (10%): 5,9 MB

Wielkość zbioru treningowego (20%): 2,6 MB

Wielkość zbioru testowego (20%): 1 MB

Wielkość bazy danych (20%): 7,6 MB

Wielkość zbioru treningowego (50%): 3,1 MB

Wielkość zbioru testowego (50%): 1,1 MB

Wielkość bazy danych (50%): 11 MB

Wyniki

20 Newsgroups

Przetestowane zostały jedynie zbiory w wersji Stemmed.

Poprawność zbioru treningowego: 95,29%

Poprawność zbioru testowego: 81,03%

Klasa	Poprawność – treningowy	Poprawność – testowy
alt.atheism	97,08%	74,92%
comp.graphics	92,47%	80,98%
comp.os.ms-windows.misc	92,31%	61,93%
comp.sys.ibm.pc.hardware	91,53%	73,47%
comp.sys.mac.hardware	94,98%	76,10%
comp.windows.x	95,28%	78,83%

misc.forsale	89,59%	66,15%
rec.autos	96,97%	93,42%
rec.motorcycles	97,49%	93,47%
rec.sport.baseball	98,32%	91,44%
rec.sport.hockey	98,67%	97,49%
sci.crypt	98,82%	95,20%
sci.electronics	92,89%	67,68%
sci.med	99,16%	85,35%
sci.space	99,33%	89,34%
soc.religion.christian	98,33%	94,47%
talk.politics.guns	99,08%	92,31%
talk.politics.mideast	98,40%	89,89%
talk.politics.misc	97,20%	58,06%
talk.religion.misc	78,51%	39,04%

Reuters 21578 R8

Poprawność zbioru treningowego: 97,70%

Poprawność zbioru testowego: 96,07%

Poprawność zbioru treningowego (niezmodyfikowany): 97,19%

Poprawność zbioru testowego (niezmodyfikowany): 95,39%

Klasa	Poprawność treningowy (stemmed)	Poprawność testowy (stemmed)	Poprawność treningowy (niezmodyfikowany)	Poprawność testowy (niezmodyfikowany)
acq	98,87%	98,28%	99,25%	99,14%
crude	97,63%	96,69%	98,02%	97,52%
earn	98,17%	98,34%	97,01%	97,51%
grain	80,49%	60,00%	70,73%	30,00%
interest	93,16%	69,14%	93,16%	65,43%
money-fx	93,20%	94,25%	94,17%	90,80%
ship	94,44%	61,11%	92,59%	47,22%
trade	96,41%	94,67%	97,21%	96,00%

Reuters 21578 R52

Przetestowane zostały jedynie zbiory w wersji Stemmed.

Poprawność zbioru treningowego: 93,19%

Poprawność zbioru testowego: 86,92%

Klasa	Poprawność – treningowy	Poprawność – testowy
acq	98,93%	98,42%
alum	74,19%	21,05%
bop	18,18%	11,11%
carcass	16,67%	0,00%
cocoa	95,65%	73,33%
coffee	98,89%	90,91%
copper	87,10%	38,46%
cotton	60,00%	11,11%
cpi	88,89%	58,82%
cpu	0,00%	0,00%
crude	97,63%	96,69%
dlr	0,00%	0,00%
earn	98,24%	98,34%
fuel	0,00%	0,00%
gas	0,00%	0,00%
gnp	96,55%	26,67%
gold	94,29%	55,00%
grain	65,85%	10,00%
heat	0,00%	0,00%
housing	40,00%	50,00%
income	0,00%	0,00%
instal-debt	0,00%	0,00%
interest	91,58%	66,67%
ipi	75,76%	45,45%
iron-steel	30,77%	8,33%
jet	0,00%	0,00%
jobs	72,97%	66,67%
lead	0,00%	0,00%
lei	0,00%	0,00%
livestock	92,31%	20,00%
lumber	0,00%	0,00%
meal-feed	0,00%	0,00%

money-fx	93,20%	94,25%
money-supply	93,50%	75,00%
nat-gas	45,83%	0,00%
nickel	0,00%	0,00%
orange	38,46%	44,44%
pet-chem	30,77%	16,67%
platinum	0,00%	0,00%
potato	0,00%	0,00%
reserves	18,92%	8,33%
retail	31,58%	0,00%
rubber	90,32%	33,33%
ship	92,59%	61,11%
strategic-metal	11,11%	0,00%
sugar	100,00%	88,00%
tea	50,00%	0,00%
tin	82,35%	0,00%
trade	96,02%	94,67%
veg-oil	15,79%	0,00%
wpi	0,00%	0,00%
zinc	0,00%	0,00%

Cade

Zbiór nie został przetestowany z powodu długiego czasu obliczeń, wynikającego z jego rozmiaru. Podane wyniki są wynikami znalezionymi w opracowaniach zbiorów danych.

Poprawność zbioru testowego: 57,27%

WebKB

Poprawność zbioru treningowego: 90,94%

Poprawność zbioru testowego: 83,52%

Klasa	Poprawność – treningowy	Poprawność – testowy
project	92,86%	75,00%
course	96,77%	92,58%
faculty	79,20%	73,80%
student	95,08%	87,68%

Własny 10%

Poprawność zbioru treningowego: 97,68%

Poprawność zbioru testowego: 95,96%

Klasa	Poprawność – treningowy	Poprawność – testowy
acq	73,17%	98,42%
crude	96,84%	95,87%
earn	98,24%	98,61%
grain	73,17%	50,00%
interest	91,05%	60,49%
money-fx	93,69%	94,25%
ship	91,67%	61,11%
trade	96,41%	94,67%
other	98,70%	97,14%

Własny 20%

Poprawność zbioru treningowego: 97,68%

Poprawność zbioru testowego: 95,63%

Klasa	Poprawność – treningowy	Poprawność – testowy
acq	98,93%	98,42%
crude	96,44%	95,04%
earn	98,38%	98,61%
grain	60,98%	40,00%
interest	88,42%	55,56%
money-fx	93,69%	90,80%
ship	89,81%	50,00%
trade	96,41%	94,67%
other	99,17%	97,61%

Własny 50%

Poprawność zbioru treningowego: 97,76%

Poprawność zbioru testowego: 96,11%

Klasa	Poprawność – treningowy	Poprawność – testowy
acq	99,00%	98,13%
crude	96,05%	94,21%
earn	98,38%	98,61%
grain	48,78%	40,00%
interest	81,05%	53,09%
money-fx	93,20%	86,21%
ship	84,26%	44,44%
trade	95,62%	94,67%
other	99,55%	99,10%