```
# I think ‘gauch’ occurs only in 1 document;
# so it did not pass the threshold.
$ ./retrieve.py -p out/ -n 5 -q "gauch"
gauch None
$ ./retrieve.py -p out/ -n 5 -q "Gauch"
gauch None

$ ./retrieve.py -p out/ -n 5 -q "description"
description ('description', 47, 238751)
n70.html    (1193, 5433)
n200.html   (911, 5377)
n5.html     (1170, 5377)
n122.html   (824, 5321)
n107.html   (807, 5267)

$ ./retrieve.py -p out/ -n 5 -q "infringe"
infringe None

# This URL doesn’t appear in the collection.
$ ./retrieve.py -p out/ -n 5 -q "http://m.eonline.com/news/2014_emmys"
httpmeonlinecomnewsemmys None

$ ./retrieve.py -p out/ -n 5 -q "abcdef"
abcdef None

# Notice how there is none ‘None’ output for ‘the’.
# This is because the tokenizer ignores the stopwords.
$ ./retrieve.py -p out/ -n 5 -q "the"

# The tokenizer completely ignores numbers
# and punctuations.
$ ./retrieve.py -p out/ -n 5 -q "20"
$ ./retrieve.py -p out/ -n 5 -q "20.07"
$ ./retrieve.py -p out/ -n 5 -q "123-456-7890"

$ ./retrieve.py -p out/ -n 5 -q "sgauch@uark.edu"
sgauchuarkedu None

$ ./retrieve.py -p out/ -n 5 -q "dominant"
dominant ('dominant', 16, 611780)
f240.html   (453, 3060)
s267.html   (1411, 2149)
s208.html   (1346, 1544)
```

```
f416.html   (648, 876)
s221.html   (1361, 875)

$ ./retrieve.py -p out/ -n 5 -q "upturn"
upturn None

# Notice how the results are the same as just “dominant”
# This is because “upturn” never appreared.
$ ./retrieve.py -p out/ -n 5 -q "dominant upturn"
upturn None
dominant ('dominant', 16, 611780)
f240.html   (453, 3060)
s267.html   (1411, 2149)
s208.html   (1346, 1544)
f416.html   (648, 876)
s221.html   (1361, 875)

$ ./retrieve.py -p out/ -n 5 -q "Tulsa"
tulsa ('tulsa', 37, 463277)
s357.html   (1511, 29823)
s445.html   (1608, 14359)
s320.html   (1471, 11125)
s455.html   (1619, 4477)
s301.html   (1450, 4077)

# Notice how the results are the same as just “tulsa”
# This is because “tulsa” documents have really high weights.
$ ./retrieve.py -p out/ -n 5 -q "dominant upturn Tulsa"
tulsa ('tulsa', 37, 463277)
upturn None
dominant ('dominant', 16, 611780)
s357.html   (1511, 29823)
s445.html   (1608, 14359)
s320.html   (1471, 11125)
s455.html   (1619, 4477)
s301.html   (1450, 4387)

$ ./retrieve.py -p out/ -n 10 -q "arkansas"
arkansas ('arkansas', 403, 509859)
n230.html   (944, 10958)
n188.html   (896, 10605)
n183.html   (891, 10380)
n258.html   (974, 10310)
```

```
n155.html   (860, 9412)
n242.html   (957, 9409)
n279.html   (997, 9146)
n144.html   (848, 8812)
n138.html   (841, 8704)
n31.html    (1032, 8360)

$ ./retrieve.py -p out/ -n 10 -q "razorbacks"
razorbacks ('razorbacks', 41, 11680)
n279.html   (997, 10804)
n138.html   (841, 5220)
n325.html   (1049, 4962)
n102.html   (802, 4962)
n336.html   (1061, 4962)
n344.html   (1070, 4962)
n248.html   (963, 4962)
n35.html    (1076, 4950)
n402.html   (1135, 4950)
n106.html   (806, 3514)

# This is a good example of a combined weight.
$ ./retrieve.py -p out/ -n 10 -q "Arkansas razorbacks"
arkansas ('arkansas', 403, 509859)
razorbacks ('razorbacks', 41, 11680)
n279.html   (997, 19950)
n138.html   (841, 13924)
n258.html   (974, 12371)
n230.html   (944, 10958)
n106.html   (806, 10951)
n188.html   (896, 10605)
n183.html   (891, 10380)
n198.html   (907, 9736)
n325.html   (1049, 9735)
n336.html   (1061, 9735)

$ ./retrieve.py -p out/ -n 5 -q "dog"
dog ('dog', 35, 341157)
n37.html    (1098, 9700)
n163.html   (869, 8216)
n406.html   (1139, 6625)
n329.html   (1053, 4265)
s341.html   (1494, 2594)
```

```
$ ./retrieve.py -p out/ -n 5 -q "cat"
cat ('cat', 29, 330716)
n62.html    (1184, 7998)
n19.html    (898, 2996)
n192.html   (901, 2857)
n98.html    (1223, 2844)
e141.html   (43, 2776)

$ ./retrieve.py -p out/ -n 5 -q "dog cat"
cat ('cat', 29, 330716)
dog ('dog', 35, 341157)
n37.html    (1098, 11734)
n406.html   (1139, 8940)
n163.html   (869, 8216)
n62.html    (1184, 7998)
n329.html   (1053, 4265)

# Each unique token is processed only once
# And the weights are multiplied. (more about this in the report)
$ ./retrieve.py -p out/ -n 5 -q "dog cat dog cat dog"
dog ('dog', 35, 341157)
cat ('cat', 29, 330716)
n37.html    (1098, 33168)
n163.html   (869, 24648)
n406.html   (1139, 24505)
n62.html    (1184, 15996)
n329.html   (1053, 12795)

# This is what happens when the tokenizer
# ignores numbers and punctuation completely.
# In this case, we don’t know the difference between
# mtv.com and m.4029tv.com
$ ./retrieve.py -p out/ -n 5 -q "http://m.4029tv.com/14430994"
httpmtvcom ('httpmtvcom', 24, 417317)
n393.html   (1124, 10481)
n392.html   (1123, 8121)
n353.html   (1080, 8121)
n239.html   (953, 8121)
n186.html   (894, 8121)

$ ./retrieve.py -p out/ -n 5 -q
"http://scores.espn.go.com/ncf/recap?gameId="
```

```
httpscoresespngocomncfrecapgameid ('httpscoresespngocomncfrecapgameid', 15,
197789)
s127.html   (1256, 29147)
s73.html    (1653, 23120)
s376.html   (1532, 3622)
s397.html   (1555, 3067)
s43.html    (1591, 3046)

# Very long token.
# Notice how the token only the first 20 and
# last 20 characters are stored in the dict file.
$ ./retrieve.py -p out/ -n 5 -q
"http://espn.go.com/los-angeles/college-football/story/_/id/11486164/usc-tr
ojans-coach-steve-sarkisian-regrets-decision-summon-ad-pat-haden-sideline"
httpespngocomlosangelescollegefootballstoryidusctrojanscoachstevesarkisianr
egretsdecisionsummonadpathadensideline
('httpespngocomlosangeonadpathadensideline', 29, 205977)
s343.html   (1496, 4184)
s430.html   (1592, 4125)
s456.html   (1620, 4097)
s133.html   (1263, 4097)
s4.html     (1558, 4068)

$ ./retrieve.py -p out/ -n 10 -q "barack"
barack ('barack', 32, 41557)
n6.html     (1181, 5244)
n330.html   (1055, 4585)
f236.html   (448, 3309)
n160.html   (866, 2464)
e16.html    (63, 1753)
e295.html   (209, 1711)
n216.html   (928, 1657)
e186.html   (91, 1443)
n371.html   (1100, 1422)
f478.html   (716, 1418)

$ ./retrieve.py -p out/ -n 10 -q "obama"
obama ('obama', 85, 215454)
n330.html   (1055, 14972)
n39.html    (1120, 10859)
f236.html   (448, 10597)
n78.html    (1201, 10274)
n6.html     (1181, 9879)
```

```
n216.html   (928, 8739)
n160.html   (866, 6498)
f418.html   (650, 5397)
f478.html   (716, 5344)
n365.html   (1093, 4861)

$ ./retrieve.py -p out/ -n 10 -q "whitehouse"
whitehouse None

$ ./retrieve.py -p out/ -n 10 -q "white house"
white ('white', 233, 431588)
house ('house', 133, 440144)
s177.html   (1311, 11585)
s242.html   (1384, 11430)
s298.html   (1445, 10832)
s216.html   (1355, 8515)
s421.html   (1582, 7632)
e192.html   (98, 6780)
s265.html   (1409, 6524)
e215.html   (123, 6339)
e177.html   (81, 6089)
s119.html   (1247, 6026)

$ ./retrieve.py -p out/ -n 10 -q "barack obama in the white house"
white ('white', 233, 431588)
barack ('barack', 32, 41557)
obama ('obama', 85, 215454)
house ('house', 133, 440144)
n330.html   (1055, 21298)
n6.html     (1181, 16431)
f236.html   (448, 15790)
n78.html    (1201, 14009)
s177.html   (1311, 11585)
s242.html   (1384, 11430)
n39.html    (1120, 10859)
s298.html   (1445, 10832)
n216.html   (928, 10396)
n160.html   (866, 9576)
```