```
L4 -- Jaccard Similarity + Shingling
[Jeff Phillips - Utah - Data Mining]

Many datasets "text documents"
 - homework assignments -> detect plagiarism
 - webpages (news articles/blog entries) -> index for search (avoid
duplicates)
      {same source duplicates, mirrors}
      {financial industry -> company doing good or bad?}
 - emails -> place advertising

How do we compare?
  exactly the same is easy  (similar is hard)
  ->  abstract space
      {R^d  ,  sets}

---------------
Distance:  d(A,B) :=  - small if close
                      - large if far
                      - 0 if the same
                      - in [0,infty]
Similarity: s(A,B):=  - large if close
                      - small if far
                      - 1 if the same
                      - in [0,1]
Often can set d(A,B) = 1 - s(A,B)
                in [0,1]
-------

Jaccard Similarity
  A = {0,1,2,5,6}
  B = {0,2,3,5,7,9}
How similar are A,B?

JS(A,B) = |A cap B| / |A cup B|
        = |{0,2,5}|/|{0,1,2,3,5,6,7,9}|
        = 3/8

Add clustering:
  C1 = {0,1,2},  C2 = {3,4},  C3 = {5,6},  C4 = {7,8,9}
similar movies get similar clusters

A-clu = {C1,C3}
B-clu = {C1,C2,C3,C4}

JS-clust(A,B) = JS(A-clu,B-clu)
              = |{C1,C3}|/|{C1,C2,C3,C4}|
```

= 2/4 = 1/2

How do we apply this to text?

All words in a document?   "bag of words" (little context)

Singling:
a "k-shingle" is a set of k consecutive items in a sequence.
items = {words, characters}

I am Sam
Sam I am
I do not like green eggs and ham.
I do not like them, Sam I am.

k=1
[I] [am] [Sam] [do] [not] [like] [green] [eggs] [and] [ham] [them]

k=2
[I am] [am Sam] [Sam Sam] [Sam I] [am I] [I do] [do not] [not like] [like green] [green eggs] [eggs and] [and ham] [like them] [them Sam]

Size := O(k + n)
k-shingle , n words
Space := O(k*n)

I am Sam
Sam I am

k-shingles on characters:

k=3:
[iam] [ams] [msa] [sam] [ami] [mia]

k=4:
[iams] [amsa] [msam] [sams] [sami] [amia] [miam]

How big to make k? characters of words? white space? punctuation? capitalization?

white space: "plane has touch down" "threw a touchdown"

punctuation: may be indication of education,
                  dialects of English (India v. US)
                  news article, blog, twitter
character v. words: similar distinctions?
  characters works surprisingly well!

How large should k be?
  * k should be large enough so probability of (almost all) shingles in any documents in corpus is low.
  emails : k = 2 or 3   (small documents)
  research articles : k = 3 or 4   (large documents)
  news articles, blog posts (in between)

26 characters + whitespace = 27
  27^5 = 14 million possible shingles
really about
  20^5 possible shingles since "z,q,x" used rarely

With news articles:
 "stop words" : {a you for the to and that it is ...}
 k = 3 where first is a stop word

Jaccard w/ shingles:

A: I am Sam.
B: Sam I am.
C: I do not like green eggs and ham.
D: I do not like them, Sam I am.

k=2, words
[I am] [am Sam] [Sam Sam] [Sam I] [am I] [I do] [do not] [not like] [like green] [green eggs] [eggs and] [and ham] [like them] [them Sam]

A = {[I am] [am Sam]}
B = {[Sam I] [I am]}
C = {[I do] [do not] [not like] [like green] [green eggs] [eggs and] [and ham]}
D = {[I do] [do not] [not like] [like them] [them Sam] [Sam I] [I am]}

Jac(A,B) = 1/3
Jac(A,C) = 0
Jac(A,D) = 1/8
Jac(B,C) = 0
Jac(B,D) = 2/7
Jac(C,D) = 3/11