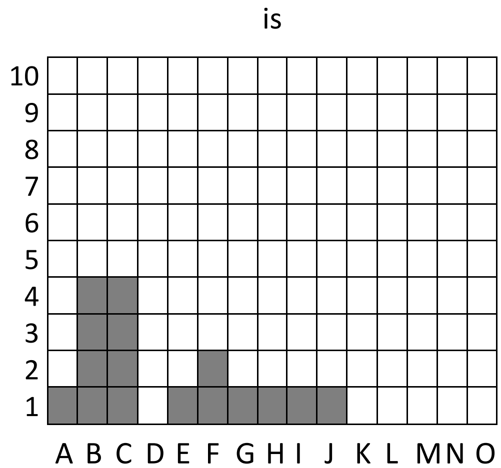
(K) Kings, Queens, and Counts (1/2)

K1.



The graph is filled in by looking at a window of two words on either side of each occurrence of is, giving the graph above. For example, whether occurs once within two words of is, the occurs four times within two words of is, etc.

K2.

a. antismartnessesquely	mystery word #10
b. aunt	mystery word #11
c. big	mystery word #5
d. can	mystery word #4
e. cats	mystery word #2
f. Kenya	mystery word #3
g. Kenyan	mystery word #9:
h. meow	mystery word #7
i. strange	mystery word #1
j. strangest	mystery word #8
k. the	mystery word #6

The key insight for this part is that analogies between words can be expressed by adding and subtracting graphs. For example, the analogy "*queen* is to *king* as *woman* is to *man*" is reflected in the fact that the difference between the graphs of *queen* and *king* is roughly equal to the difference between the graphs of *woman* and *man* (e.g., *queen* has 6 fewer co-occurrences with A than *king*; 2 more co-occurrences with C than *king*; 7 more co-occurrences with E than *king*; 1 fewer co-occurrence with H than *king*; 3 more co-occurrences with M than *king*; and 3 fewer co-occurrences with N than *king*. *Woman* and *man* have roughly the same differences in co-occurrences). I say "roughly" because there is a margin of error of plus or minus one in all cases to reflect the fact that the addition and subtraction of distributional vectors is by no means exact. Thus, for example, *aunt* can be identified as 11 because the difference between *uncle* and graph 11 is similar to the difference between *king* and *queen* or the difference between *man* and *woman*. Similarly, *cats* and *meow* can be identified as the pair that satisfies:

cats - meow = horse + (queens - queen) - neigh

(K) Kings, Queens, and Counts (2/2)

and *Kenya* and *Kenyan* are the pair such that:

Kenya - *Kenyan* = *India* - (*rupee* - (*ariary* - (*Antananarivo* - (*Berlin* - (*Merkel* - (*Roussef* - *Brazilian*))))))

Lastly, *the* and *antismartnessesquely* can be identified as the words that occur with other words extremely frequently and not at all, respectively.

K3. *Mystery word #4* is *can*. You might expect *can* to have the graph labeled *expected graph for mystery word #4* because that graph reflects the analogy "*king* is to *kings* as *can* is to *cans*" or "*queen* is to *queens* as *can* is to *cans*." However, in addition to being the singular form of *cans* as in "a **can** of soup," *can* also is an auxiliary verb as in "Nothing **can** stop me now!" Thus, when the graph is formed for *can*, it will include counts for the noun *can* but also the (much more common) auxiliary verb *can* (plus the verb *can*, as in "I love to **can** vegetables") which muddies the waters even further). That is why the actual graph for *can* has much higher counts than the *expected graph for mystery word #4*.

(I know the solution provided above is kind of hypocritical for being more than the requested maximum of two sentences. An actual answer could just be something like "Mystery word #4 is *can*, which has multiple meanings." Anything that mentions how *can* has more than one meaning will get full points.)