Hadoop Project Results

Alexandra Berke and Matt Patenaude

1) a) What is the most common bigram of all time? b) What is the most common bigram in 1987? c) How about 1953?

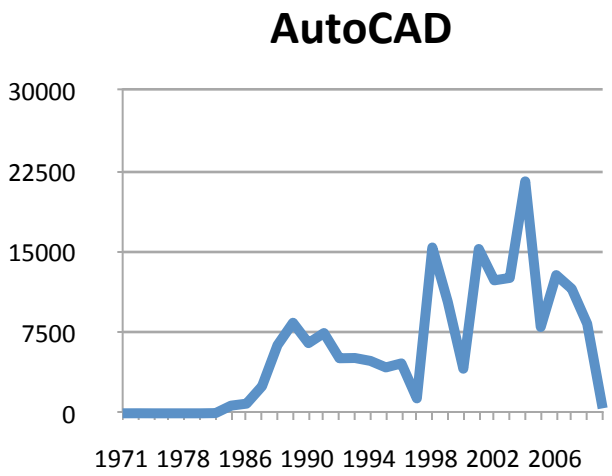
The most common bigram of all time is "of the," followed by "in the" and "to the." The most common bigram in 1987 is also "of the" (similarly followed by "in the" and "to the"), as is the most common bigram in 1953.

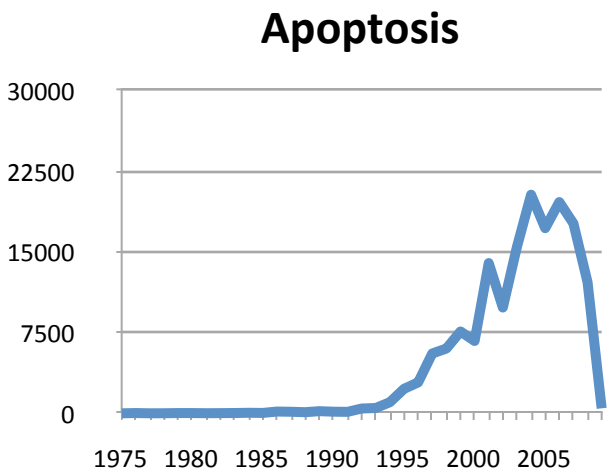
2) Identify a few words that were coined after 1970. These should be words that never appear before 1970 but begin to appear (hopefully with some non-negligible frequency) in later decades. You might illustrate the increasing usage of the new term by plotting its frequency in the corpus on the *y*-axis against time on the *x*-axis.

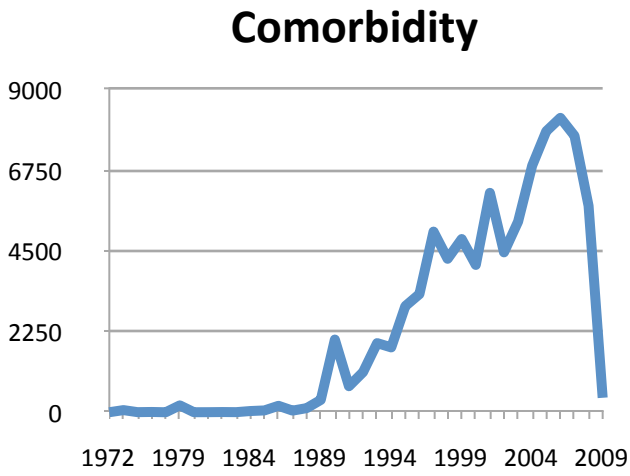
We selected the 5 most frequently-occurring words that only occurred after 1970: AutoCAD, apoptosis, comorbidity, comorbid, and ActionScript. Their frequency graphs (from time of first occurrence) are shown on the next page.

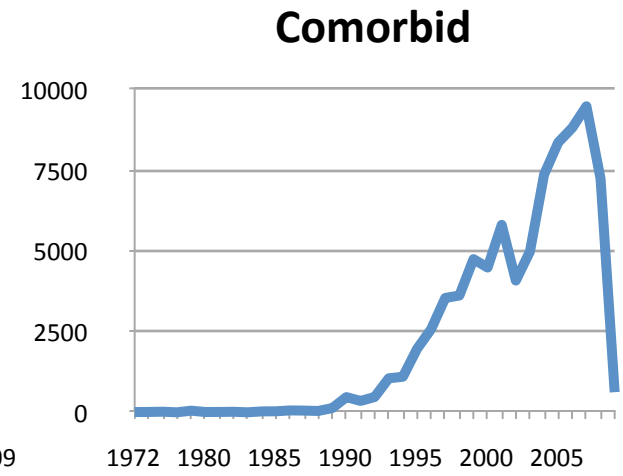
3) Suppose we're interested in finding trigrams which tend to appear many times on the same page or many times in the same book. a) What is the trigram that appears at least 10 times with the lowest ratio of page count to total count? b) What is the trigram that appears at least 10 times with the lowest ratio of *book count* to total count?

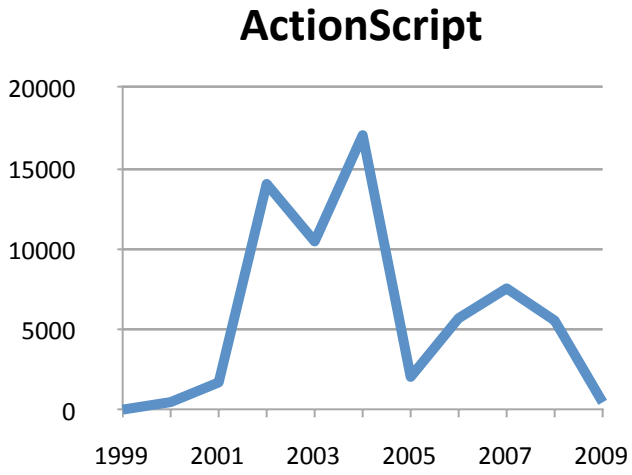
The trigram that appears at least 10 times with the lowest ratio of page count to total count is "let go let." The trigram that appears at least 10 times with the lowest ratio of *book count* to total count is "- dlb -" (the lowest one with actual "words" is "viii high sch").


AutoCAD
30000
22500
15000
7500
0
1971 1978 1986 1990 1994 1998 2002 2006


Apoptosis
30000
22500
15000
7500
0
1975 1980 1985 1990 1995 2000 2005


Comorbidity
9000
6750
4500
2250
0
1972 1979 1984 1989 1994 1999 2004 2009


Comorbid
10000
7500
5000
2500
0
1972 1980 1985 1990 1995 2000 2005


ActionScript
20000
15000
10000
5000
0
1999 2001 2003 2005 2007 2009