K-GRAM INDEX

USED TO APPROXIMATE EDIT DISTANCE (WHICH IS TOO EXPENSIVE) IN THE CASE OF MORE THAN ONE SPELL ERROR (WHICH THE 1 ERROR CORRECTION APPROACH DON'T COVER, OFC).

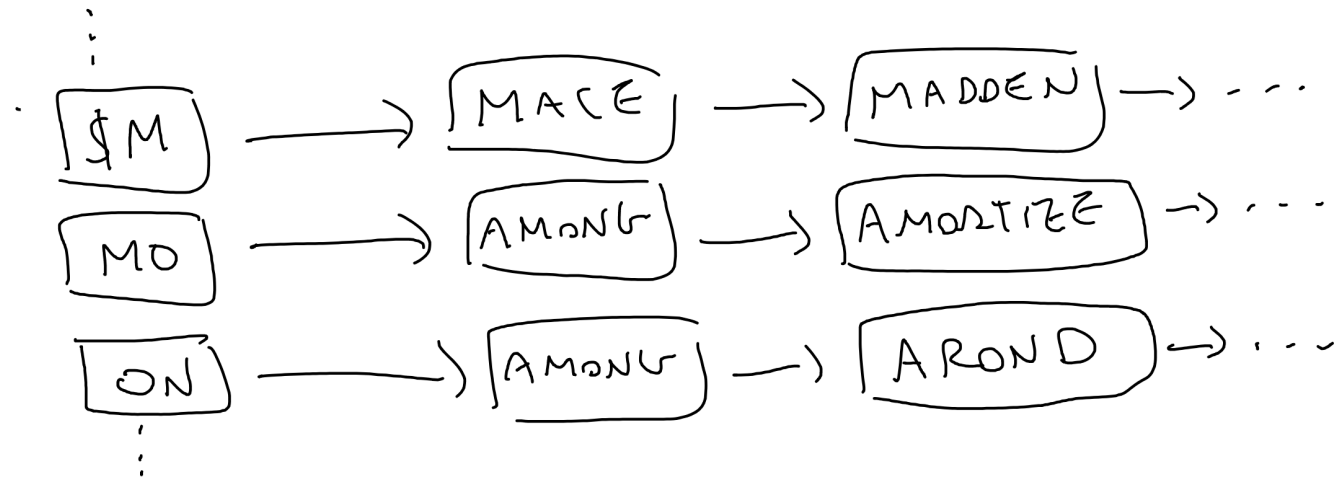
THE K-GRAM INDEX IS AN INVERTED INDEX CONTAINING, FOR EACH K-GRAM, ALL TERMS INCLUDING THAT K-GRAM.

Let's see how to build it and use it for our problem

(1) First off, we have to build the K-gram index using the terms from our lexicon. Let's assume that each term is anticipated by K-1 special char $, in order to ensure that the number of K-grams is equal to the length of the string generating them.

GENERATE ALL POSSIBLE OVERLAPPING KGRAM FOR EACH TERM AND THEN USE THEM TO BUILD THE INVERTED LISTS

EXAMPLE OF K-GRAM INDEX (K=2)



NOTE: I SHOWN ONLY THE K-GRAMS OF THE TERM "MON", BUT THE K-GRAM $M IS GENERATED BY ALL STRINGS STARTING WITH M

NOTE: GIVEN A TERM OF SIZE $|t|$, THE NUMBER OF RESULTING K-GRAMS IS $|t|$

$CASA $\Rightarrow$ $C, CA, AS, SA

(2) NOW WE CAN CHECK FOR ERRORS IN A QUERY USING THE K-GRAM INDEX.

ENUMERATE ALL OVERLAPPING K-GRAMS IN A QUERY AND SEARCH THEM IN THE INDEX, THEN KEEP TRACK OF HOW MANY K-GRAMS ARE IN COMMON BETWEEN Q AND THE TERMS OF THE LEXICON

(3) SELECT THE POSSIBLE CORRECT TERMS BY THRESHOLDING ON THE K-GRAMS MATCHING THE MOST

- SINCE OUR QUERY TERM IS OF LENGHT $|Q|$
- GIVEN e THE MAXIMUM AMOUNT OF ALLOWED ERRORS, $e*k$ K-GRAMS OF OUR QUERY TERM MIGHT BE DIFFERENT FROM THE K-GRAMS OF A TERM IN THE LEXICON.

THEREFORE, AT LEAST $|Q| - e*k$ K-GRAMS OF THE QUERY TERM MATCH WITH THOSE OF A LEXICON'S TERM.

$$\boxed{|Q| - e*k \geq \#\text{ matching k-grams with a term}}$$

WARNING: THIS IS ONLY AN APPROXIMATION AND WE MIGHT WANT TO DO THE ED ANYWAY

EXAMPLE OF WARNING

GIVEN $K = 3$, COMPARE $\$\$$NOVEMBER (FROM LEXICON) WITH $\$\$$DECEMBER (φ)

$\$\$$NOVEMBER $\Rightarrow$ $\$\$N$, $\$NO$, NOV, OVE, VEM, EMB, MBE, BER

$\$\$$DECEMBER $\Rightarrow$ $\$\$D$, $\$DE$, DEC, ECE, CEM, EMB, MBE, BER

if $e = 1$:

$$|\varphi| - e^* K = 8 - 3 = 5 \quad \text{NO}$$

if $e = 2$:

$$|\varphi| - e^* K = 8 - 6 = 2 \quad \text{OK}$$

BUT WE NEED E.D. TO SAY THAT THE DISTANCE IS ACTUALLY 3

EXAMPLE

GIVEN $S = \{PITOM, DAD, DADDY, ZOOM\}$

1. BUILD THE CORRESPONDING 2GRAM INDEX
2. SHOW THE EXECUTION OF THE 1-EDIT ERROR SEARCH ON S USING THE INDEX, GIVEN P = ATOM

(1) $PITOM$, DAD, $DADDY$, $ZOOM$
1 2 3 4

$P → 1
PI → 1
IT → 1
TO → 1
OM → 1, 4
$D → 2, 3
DA → 2, 3
AD → 2, 3
DD → 3
DY → 3
$Z → 4
ZO → 4
OO → 4

SORT FOR BETTER I/OS

$D
$P
$Z

(2) $ATOM

Kgrams = $A, AT, TO, OM

Let's check the matches with the kgram index

$A → ∅

AT → ∅

TO → 1

OM → 1, 4

2 CANDIDATES = PITOM, ZOOM

THE CORRECT ONE IS THE ONE SUCH THAT

$$|Q| - e * K \geq \#\text{matching Kgrams}$$

4 — length of ATOM

1 — 1-edit error

2 — 2-grams

4 - 2 = 2

the correct candidate is PITOM