LET'S CONSIDER FOR A MOMENT HOW

SEARCH ENGINES

WORK

A USER TYPES IN A QUERY

THE SEARCH ENGINE RETURNS A LIST OF WEBSITES THAT ARE RELEVANT TO THE SEARCH QUERY

HOW DOES IT DO THIS?

THE SEARCH ENGINE MAINTAINS AN INDEX OF WEBSITES (A DOCUMENT CORPUS)

IT FINDS THE DOCUMENTS FROM THIS INDEX THAT ARE MOST 'SIMILAR' TO THE QUERY DOCUMENT

ANY DOCUMENT CAN BE REPRESENTED AS A POINT IN A HYPERCUBE

HOW?

TAKE THE SET OF ALL WORDS THAT APPEAR IN THE CORPUS

W1, W2,...Wn

ANY DOCUMENT IN THE CORPUS WILL CONTAIN SOME SUBSET OF THESE WORDS M1, M2,..Mi

REPRESENT EACH DOCUMENT AS A TUPLE (X1,X2,X3...Xn)

WHERE A PARTICULAR ELEMENT XJ IS 1 IF WORD J APPEARS IN THE EMAIL, ELSE IS 0

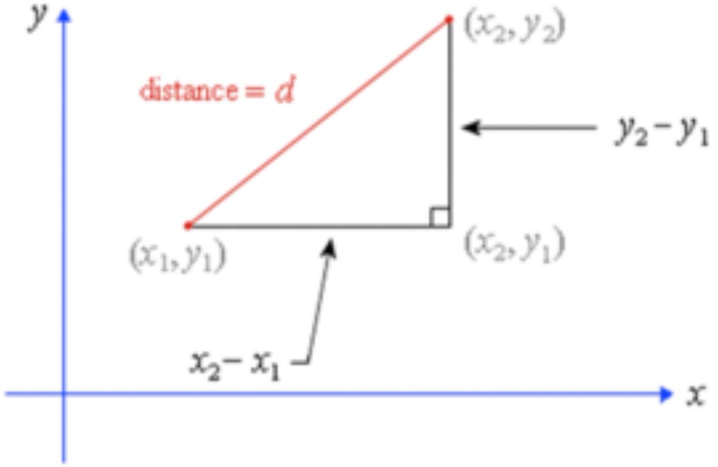
OK, NOW THAT WE HAVE REPRESENTED A DOCUMENT AS A POINT IN A HYPERCUBE - WHAT NEXT?

DO THIS FOR THE SEARCH QUERY THE USER ENTERED WELL AS FOR ALL THE DOCUMENTS IN THE CORPUS

NOW THESE ARE ALL POINTS IN SPACE, SO WE CAN FIND THE DISTANCE BETWEEN THEM

(BUNCH OF WAYS TO CALCULATE DISTANCE BETWEEN 2 POINTS - INCLUDING THE SUPER-SIMPLE EUCLIDEAN DISTANCE FORMULA - MORE ON THIS IN A MINUTE)

THE SIMILARITY OF 2 DOCUMENTS IS THEN A FUNCTION OF THE DISTANCE BETWEEN THEM



THE SETUP ABOVE WON'T YET WORK IF WE WANT TO FIND THE MOST 'RELEVANT' DOCUMENTS TO OUR SEARCH QUERY

EACH POINT JUST REPRESENTS THE PRESENCE OR ABSENCE OF A WORD IN A DOCUMENT

(1) (0)

WITH THIS METHOD WE CAN ONLY ELIMINATE DOCUMENTS THAT DO NOT CONTAIN ANY OF THE WORDS IN OUR QUERY

USUALLY THAT WILL STILL LEAVE US WITH A LOT OF DOCUMENTS

THE PROBLEM IS THAT NO WEIGHTAGE IS GIVEN TO HOW OFTEN THE WORDS IN THE SEARCH QUERY APPEAR IN THE DOCUMENT

THIS IS EXACTLY WHERE

TF-IDF COMES IN

TF-IDF COMES IN

TERM FREQUENCY

TERM FREQUENCY HOW OFTEN THE WORD APPEARS IN OUR DOCUMENT

IF A WORD APPEARS MORE OFTEN IN A DOCUMENT, ITS CONSIDERED MORE IMPORTANT WHILE REPRESENTING THAT DOCUMENT

INVERSE DOCUMENT FREQUENCY

REPRESENTS THE INVERSE OF HOW OFTEN THE WORD APPEARS IN THE ENTIRE CORPUS

SOME WORDS APPEAR IN MOST DOCUMENTS

(THIS, THAT, NEW, ONE, TWO ETC)

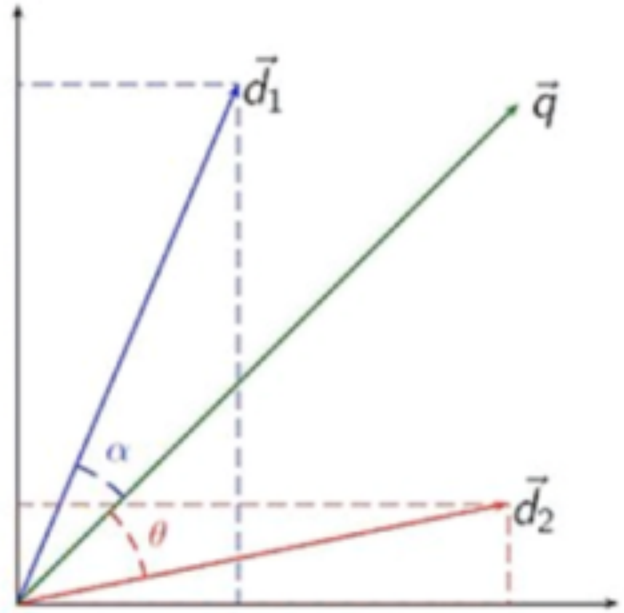
RARE WORDS SHOULD BE GIVEN HIGHER WEIGHTAGE THAN VERY COMMON WORDS

THESE ARE WEIGHTS THAT WE'LL ATTACH TO EACH OF THE WORDS IN OUR DOCUMENT (EACH ELEMENT OF THE TUPLE THAT REPRESENTS IT IS WEIGHTED BY THESE NUMBERS)

COSINE SIMILARITY

WE HAVE ALREADY SAID THAT THE DISTANCE BETWEEN TWO DOCUMENTS (HOW SIMILAR THEY ARE) CAN BE CALCULATED USING EUCLIDEAN DISTANCE

ANOTHER WAY TO COMPUTE SIMILARITY IS TO USE THE ANGLE BETWEEN THE TWO VECTORS THAT REPRESENT THESE DOCUMENTS



INSTEAD OF THE ACTUAL ANGLE, USUALLY THE COSINE OF THE ANGLE IS CALCULATED HENCE THE NAME

$$\cos\theta = \frac{\mathbf{d_2} \cdot \mathbf{q}}{\|\mathbf{d_2}\| \, \|\mathbf{q}\|}$$