TF is freq of word / max word freq in doc, idf is $\log \frac{\text{num doc total}}{\text{num docs in}}$

Dot product: sim (item in doc, query) → sum((item · query))

Eluciden Distance: $\sqrt{\text{sum}((\text{doc item} - \text{query})^2}$ → large diff length

Cosin Angles: 0-90° map 0-1, 1 means same
0 totaly different

Cosine Similarity: $\frac{\text{Sum of (doc item} \times \text{query)}}{\sqrt{\text{sum of square items in doc}}\sqrt{\text{sum of square item in query}}}$

Precision = $\frac{\text{\# Relavant items retrieved}}{\text{\# retrieved items}} = \frac{TP}{TP+FP}$

Recall = $\frac{\text{\# relavant items retrived}}{\text{total \# relavant items}} = \frac{TP}{TP+FN}$

Accuracy = $\frac{TP+TN}{TP+FP+FN+TN} = \frac{\text{\# true}}{\text{total}}$

Query = (num times item appears / max times item appears) x tf idf

	Relavant	Nonrelavant	total
retrieved	TP	FP	TP+FP
Not retrieved	FN	TN	FN+TN
total	TP+FN	FP+TN	total

→ total # retrieved

↑ # relavant

↑ # nonrelavant

↑ total #

$$F_{\text{measure}} = \frac{1}{\alpha \frac{1}{P} + (1-\alpha)\frac{1}{R}} = \frac{(\beta^2+1)PR}{\beta^2 P + R}$$

Reciprical Rank = 1/pos relevant

$$F_1 = \frac{2PR}{P+R}$$ → this is when $\alpha = \frac{1}{2}$ or $\beta = 1$

$$\text{Precision @}k = \frac{TP@k}{TP@k + FP@k}$$

Average Precision

$$\frac{\text{Precision @ relevant}}{\text{\# of relevant total}}$$

$$\text{Recall @}k = \frac{TP@k}{TP + FN}$$

Initial Rank

1/ # items

New page rank is

$$\text{Sum}\left[\frac{\text{PR of input}}{\text{num outputs of item}}\right]$$

$$MMR = \frac{1}{\text{total \# query}}\left(\sum_{i=1}^{\#\text{queries}} \frac{1}{\text{rank}_i}\right)$$

If a sink you distribute page rank equally on all other items, after new calc think iteration i and i'

Robots.txt specifies rules of a webpage → get by /robots.txt

VSM assumes independence, needs to be precise

Crawl BFS, DFS

MAP is average precision on multiple queries

One pagerank can't tell how important a page is

web crawler like librarian → places items index for later

Freshness is accuracy of copy, Age is how outdated a copy is

Uniform policy → revisit all pages with the same frequency

Porportional Policy → revisit pages with more changes more frequently

Web crawling is search, web scraping is extracting data

avoid duplication

↳ create index for search

Canonical form is http://www.xyz.com/xyz/