CPEG 657 - Search & Data Mining 2/22

Search Results - Ranking Preferred

```
* Relevance Modeling
*
```

Search Before Google

- Information gathering done at library
- Search engines were based on "keyword" matching
  Spammers took advantage of this

Google Search

- Page Rank - Link authority based

Inverse Document Frequency(IDF)

- Higher weight to less common terms

$$IDF(t) = 1 + log(n/k)$$

Similarity

$$Similarity(D, Q) = \sum_{i=1}^{n}$$

Similarity (Pivoted Normaliation Formula)

- Give priority to first occurrance and less to repeated

$$S(D, Q) = \sum_{t=Q,D} \frac{1 + ln(1 + ln(c(t, D)}{(1 - s) + s\frac{|D|}{avdl}} * c(t, Q) * ln\frac{N + 1}{df(t)}$$