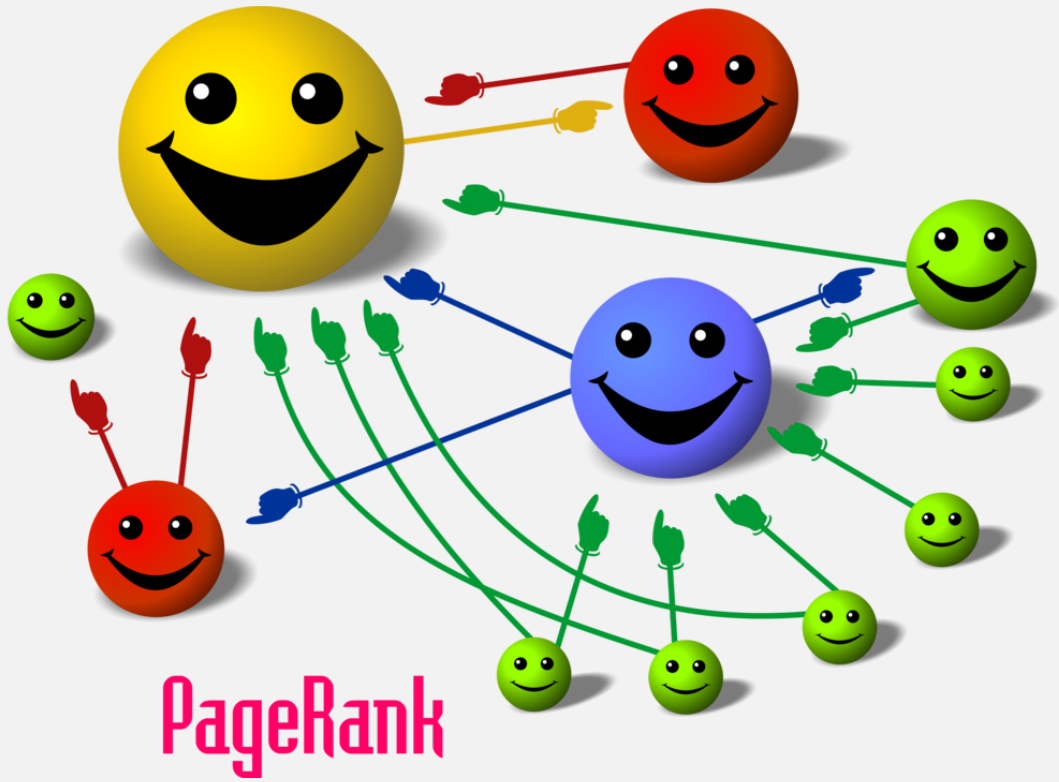
Spark Examples

Spark Examples

- **Wordcount**
- **Case-Analysis**
- **Top N Patents**

PageRank(1)



PageRank(2)

- The PageRank formula:

$$PR(p_i; t+1) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j; t)}{L(p_j)}$$

  where $p_1, p_2, ..., p_N$ are the pages under consideration, $M(p_i)$ is the set of pages that link to the page p_i, $L(p_j)$ is the number of outbound links on page p_j, and N is the total number of pages.
- The initial PageRank is usually set to 1/N (or just 1 for the simpler case). The total page ranks is 1 (or N for the simpler case).
- When calculating PageRank, pages with no outbound links are assumed to link out to all other pages in the collection. Their PageRank scores are therefore divided evenly among all other pages. In other words, to be fair to pages that are not sinks, these random transitions are added to all nodes in the Web.

PageRank(3)

- The computation ends when for some small ϵ,

$$\sum_{j=1...n} \mid PR(p_j; t+1) - PR(p_j; t) \mid < \epsilon$$

- Google's founders reported that the PageRank algorithm for a network consisting of 322 million links (in-edges and out-edges) converges to within a tolerable limit in 52 iterations.
- For large networks, the number of iterations is $O(\lg n)$.

Correlation

- Correlation is any statistical association. Most commonly refers to the degree to which a pair of variables are *linearly related*.
- See example: Spark/correlation/CrossCorrelation.java
- *A causal relationship implies correlation but the presence of a correlation is not sufficient to infer the presence of a causal relationship*.
- Pearson correlation coefficient is commonly used. It reflects a linear relationship (which may be present even when one variable is a nonlinear function of the other).
- The Pearson correlation coefficient between two random variables X and Y with expected values μ_X and μ_Y and standard deviations σ_X and σ_Y is defined as:

$$\rho(X, Y) = corr(X, Y) = \frac{cov(X, Y)}{\sigma_X \sigma_Y} \quad (1)$$

$$= \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y} \quad (2)$$

- The value is in the range $(-1, +1)$. $+1$ is perfect increasing/direct linear relationship, -1 is perfect decreasing/inverse linear relationship. As it approaches zero there is less of a relationship (closer to uncorrelated).

P-value

- To quantify the idea of *statistical significance*.
- The p-value is defined as the probability, under the null hypothesis H, about the unknown distribution F of the random variable X, for the variate to be observed as a value equal to or more extreme than the value observed.
- For example, this could be $Pr(X \geq x|H)$ (right-tail event) or be $Pr(X \leq x|H)$ (left-tail event) or $min\{Pr(X \leq x|H), Pr(X \geq x|H)\}$ (double-tail event)
- The null hypothesis H is rejected if any of these probabilities is less than or equal to a small, fixed but arbitrarily pre-defined threshold value α, which is referred to as the level of significance.
- The setting of α is arbitrary. By convention, α is commonly set to 0.05, 0.01, 0.005, or 0.001.