Spark Examples

Spark Examples

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

Correlation

- Correlation is any statistical association. Most commonly refers to the degree to which a pair of variables are *linearly related*.
- *A causal relationship implies correlation but the presence a correlation is not sufficient to infer the presence of a causal relationship*.
- Pearson correlation coefficientis 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).