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).

P-value

- To quantify the idea of *statistical significance*.
- The p-value is defined as the probability, under the null hypothesis H, about the unknonw 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.