General Evaluation Methodology

- Have humans to create a test collection where every document is tagged with the desired categories ("ground truth")
- Generate categorization results using a system on the test collection
- Compare the system categorization decisions with the human-made categorization decisions and quantify their similarity (or equivalently difference)
  - The higher the similarity is, the better the results are
  - Similarity can be measured from different perspectives to understand the quality of results in detail (e.g., which category performs better?)
  - In general, different categorization mistakes may have a different cost that inevitably depends on specific applications, but it is okay not to consider such a cost variation for **relative comparison of methods**