Research Methods Worksheet 3: Experimentation

You have developed a new search tool that aims to improve the discovery of documents in a news service that combines news content in a Wiki format. The service is imaginatively called WikiNews. The search tool for WikiNews has to support both long and short news articles that are connected using hyperlinks (i.e. standard web links).

Established Methods

To compare your new search tool against an existing search algorithm, PageRank, you need to run an experiment. The standard method for testing the performance of a search algorithm includes:

1. A corpus (test set) of documents
2. A list of search topics that are associated with the test set, together with a list of documents in the collection that are relevant to each individual search topic.
3. Both algorithms are then run on the test set, providing a ranked list of matching documents for each algorithm on each search topic.
4. Two measurements are then calculated, for each algorithm:
    a. Precision – out of the top-ranked n (often 10) documents for each search topic, how many are found in the list of relevant documents? This percentage is averaged over all the search topics.
    b. Recall – out of all the known relevant documents from the search topic lists, how many are found in the result lists. This is a percentage of the known relevant documents.

    The ideal algorithm would score 100% for both measures.

The above method can be run automatically, without recruiting users. The datasets in automatic testing have used have been established for years. The original benchmark relevance judgments of the documents for each search topic were made by experts. The outcomes of automatic experiments are seen as highly reliable.

Some recent studies have started to vary this method. In the *interactive* method, judgments of relevance are made by users, typically with 30 or more users recruited to reduce the effects of individual preferences. For this, the number of relevant documents is recorded by the users in laboratory studies, and the precision is calculated by the number of documents that the average user found to be relevant in the top ten. The research community is divided on the reliability of the interactive method.

Test Collections

There are test collections available that are regularly used in running both automatic and interactive experiments. For you, the following are immediately available as test data sets:

1. TREC: a classic corpus for testing search tools. The material available includes collections of government documents and newspaper articles (there are a variety of separate collections of each type). The documents are not hyperlinked, but do contain news material. Most new articles are short. Most, but not all, content has an associated date.
2. NIST: a collection of government webpages. These documents are hyperlinked, but not in Wikipedia or news formats. All content has an associated date.
3. Reuters: a collection of Reuters newswire material – mostly, but not only, short news articles. This is more up-to-date than the TREC collections, and has some hyperlinks. All content has an associated date.
4. Wiki: a collection of Wikipedia pages. This is almost current, and has, mostly, a good volume of hyperlinks. There is some news content, but most material is not from a news source. All documents are given the date of the last change made to them.

For collections 1 and 2 there are a list of search topics, with associated lists of relevant documents. For collections 3 and 4, there is no existing set of search topics. If you use an automatic method, then these two collections would need you to arrange for the creation of relevant search topics, and corresponding lists of relevant documents.

A New Measurement

A final complication to consider is measurement – *precision* and *recall* are given above. However, in the context of news articles, recent articles are expected to be perceived as more relevant than older material. There is no standard measurement available for this.

Your Task

Design **two** studies: first, an initial study with limited time available (up to six weeks), and second, an ideal study given up to six months of time. Both time periods include the setting-up, running and reporting of the experiment.

1) Select either an *automatic* or *interactive* assessment method, explaining your choice.
2) Choose an appropriate dataset for the analysis. If choosing either dataset 3) or 4), explain how you would determine the relevance of documents. For each of the existing datasets, explain your reasons for choosing or rejecting it.
3) Identify what measurements you would use in the study. Particularly consider how you might address the issue of how recent the content is.
4) Report any limitations to the results that would be influenced by the method chosen.