I already loaded a ~3 GB dataset describing pageviews and link network between all pages on wikipedia in 2014, and how many times people navigated to those pages via certain ways (internal wikipedia link? google? bing? yahoo? other search engine?). The questions I want to answer are of the form “Think of k categories. Define a sample set of wikipedia pages associated with each category. Measure characteristics of distribution of x over k“. “proportion of search engines to internal navigation as compared to mean” is an interesting instance of x because a high-proportion of internal navigation means people are reading multiple related pages and researching/perusing something.

category, attribute pairs:

{pages in english, pages in spanish, pages in chinese … }, relative popularity of google and bing and yahoo.

{academic subjects, sexual subjects, <random sample>}, proportion of search engines to internal navigation as compared to mean (interpretation of question: how common is it to use wikipedia to quickly look up academic topics versus research into topics? do people often click on sex-related stuff when originally reading other things compared to the average amount of clicking on stuff (i.e. is there wikipedia clickbait?))

{pages related to microsoft, pages unrelated to microsoft}, relative popularity of bing compared to mean bing usage.

{relatively broad topics, relatively narrow topics}, proportion of search engines to internal navigation as compared to mean.

Another idea is to use graph algorithms to identify highly-connected subgraphs of the graph, and automatically count the “proportion of search engines to internal navigation as compared to mean” where samples are defined as members of the subgraphs, and use my human knowledge to look through the data and see trends.