**DSCI 558: Building Knowledge Graphs**

**Quiz 2 – Crawling**

Question 1 (10 points)

List five (5) challenges of crawling. Briefly describe each challenge in one sentence.

* scale: the number of pages you want to crawl is usually huge
* deduplication: prevent crawlers to hit the same page multiple times
* cost: crawling is expensive because of:
  + dns
  + fetching
  + parsing/extracting
  + memory/disk
* speed: politeness can reduce speed in crawling
* errors, redirects: collected links (pages) can contains errors or redirects
* freshness: keep the data up to date
* deep web, forms: need to know the format of input to extract data
* counter-crawling/access (login, captchas, traps, fake errors, banning): websites design a set of techniques to prevent people from crawling their data
* localization: different versions of websites based on locale
* dynamic pages: need to interact with webpages
* infinite scrolling: need to interact with webpages
* archiving: