CP 423 Assignment1 Read Me

1. What this program does

Collects pages - starts at the Wikipedia page for Canada, then follows links only two clicks away.
Builds a word list - for every page it keeps the words (minus common filler words like “the”, “and”).
Stores where words appear - a quick look-up table: each word points to the pages that contain it.
Lets you search with AND / OR / NOT - a small prompt where you can type things like:
ontario AND population
(british AND columbia) OR alberta
canada AND NOT ontario

2. How it works

Step	What happens
Crawl	Visits up to **200** articles, never deeper than **2 links** from the seed.
Clean text	- lower-case - keep only a-z and 0-9 - remove 1-letter words - drop English stop-words.
Index	Saves each word in a Python dictionary so look-ups are instant.
Search	Parses the query, then does set math in memory to find matching pages.

3. Assumptions

- Required packages and imports are downloaded
- Internet connection is available while script runs

4. Results you can expect

Pages collected: This has been capped at 200 due to many pages available
Example Query outcomes:

- query> unicorn OR quebec -> 53 documents matched
- query> british NOT columbia -> 140 documents matched
- query> canada AND ontario -> 42 documents matched

5. Insights / notes

- A small, in-memory index is enough for quick Boolean search on a small crawl.
- Normalising URLs (lower-case, strip “#…”, remove last “/”) plus a 200-page cap prevents endless loops.
- The program can be extended easily:
  raise MAX_DEPTH and MAX_PAGES for more results