Karthik Kalyanasundaram
Andre Wu

Project 2

1.We found a total of 144 unique pages that were in the domain

2. The largest page contained a total of 225 tokens.

2.The 50 most common words in this corpus are

['generation', 'text', 'coherent', 'discourse', 'propose', 'model', 'sentence', 'level', 'Generating', 'important', 'challenging', 'task', 'particularly', 'language', 'tasks', 'such', 'story', 'Despite', 'success', 'modeling', 'coherence', 'existing', 'models', 'still', 'struggle', 'maintain', 'event', 'sequence', 'throughout', 'generated', 'conjecture', 'because', 'difficulty', 'decoder', 'capture', 'semantics', 'structures', 'context', 'beyond', 'paper', 'long', 'represent', 'prefix', 'sentences', 'decoding', 'end', 'two', 'pretraining']

There are this many subdomains in the ics

apply.ics.uci.edu 0
campusgroups.ics.uci.edu 0
directory.ics.uci.edu 0
facebook.ics.uci.edu 0
forecast7.ics.uci.edu 0
instagram.ics.uci.edu 0
intranet.ics.uci.edu 6
mcs.ics.uci.edu 0
mds.ics.uci.edu 15
mhcid.ics.uci.edu 18
mswe.ics.uci.edu 15
secure.ics.uci.edu 0
twitter.ics.uci.edu 0
uci.ics.uci.edu 0
www.ics.uci.edu 37
youtube.ics.uci.edu 0