Shaghik Issakhani
Alex Lin
Zhenxiang Peng
Manuel Miranda

Assignment 1 Report
(Web Crawling)

Approach:

We used the Python programming language for this assignment. We have 6 classes. The First one that we worked on was the general class which has a few housekeeping basic methods needed for most of the python programs. The Second class is the Spider class itself. When the spider first starts out, from the home page it starts gathering all of the URLs. Each link that it finds, it's going to add into a waiting list. Whenever it is done crawling that page, it is gonna add that URL into a file so we can keep track of what pages have already been crawled, and what links we have found that are still in the waiting list file. Also the program is multithreaded which means it crawls the pages simultaneously, and that way we can save some time. The 'append to file' function is the one responsible for adding the new URLs to the existing files. We also included a class variable in our spider class that can be shared among all of the instances. The program becomes multithreaded when the spider gathers a few links in the waiting list, so basically the first spider doesn't run on a thread. We have another file called domain, and the functions will be responsible for extracting the domain name. For language detection we used the "langdetect" library. The "requests" library was used to make HTML requests and save them, while the "BeautifulSoup" library was used to parse the HTML. We follow the allow and disallow restrictions on robots.txt files by using the robot parser provided by the "urllib" library. The word frequency calculator is a program that runs separately from the web crawler.. This allows us to collect all the HTML data we want parse through first.

Challenges:

One of the significant difficulties that we face has been time management. Since everyone is busy with school, work, and personal responsibilities, it was hard to meet up at times that benefit all parties. But we overcome it by checking in every other day or whenever we can. Github makes completing the assignment easier because anyone can push and pull the code to work on it in their free time. We would assign at least one task/feature to finish, and everyone would participate and try to knock down the task. It wasn't easy at first because some of us were unfamiliar with pulling from and pushing to Github, but it is becoming more straightforward as we become more accustomed to

using it. Since some of us had only basic knowledge of the Python language, it made it more challenging, but we got through since it was not hard to learn it. Since we all are Computer Science majors and know other languages, it was easier for us to learn Python. Another difficulty came in the form of understanding the assignment complexity. When discussing how thoroughly to web crawl through all the pages to get the frequent words and other details, we had a difficult time figuring out how in-depth we were supposed to crawl through a page and how to make our crawler follow the robot.txt page. We also tried to implement a crawl depth to our crawler, but couldn't get that to work in time for submission.

Group member's contribution:

During the initial group meeting, we spent most of the time fixing the errors that we were getting on our initial program. We did the meeting over a Discord voice channel and everyone showed up on time and already had basic knowledge of the concept and what needs to be done, and how they can contribute on their part.

Zhenxiang worked on switching Python packages midway through creating the crawler to facilitate some of the things we needed to achieve. He also worked on adding language detection, fixed robot parser, changed the word detection class to work on downloaded files, and fixed file saving issues that we were having. He was also mainly the one who was keeping the GitHub Repository active.

Alex worked mostly on the get_words file and did the implementation for it. He also worked on writing the challenges section of the report. He also worked on trying to fix the bugs as well. He created google document to start working on the project report.

Manuel worked on downloading the HTML into separate files. He also worked on the robots.txt parser. He also worked on fixing the bugs that we were getting in our program. Manuel also created the GitHub repository. He also worked on the approach and challenges section of the report. He also configured and ran the crawler and word frequency programs on the three target domains to acquire the output data.

Shaghik was responsible for writing the initial code for some of the files. She also worked on fixing the part of the code responsible for downloading the HTML into files. Shaghik also worked on the approach and group member contribution sections of the report.

https://www.tudn.com/	
word	# of times
de	21823 times
el	11384 times
en	11287 times
la	10614 times
del	4430 times
que	4230 times
con	3693 times
para	3595 times
por	3315 times
se	3197 times
al	3072 times
los	3023 times
champions	2618 times
no	2492 times
mx	2485 times
un	2439 times
liga	2404 times
las	2279 times
su	2078 times
univision	1822 times
noticias	1736 times
américa	1695 times
concacaf	1676 times
mls	1391 times
más	1376 times
ante	1351 times
es	1325 times
uefa	1265 times
deportes	1238 times

https://www.gmarket.co.kr/	
word	# of times
상품	14122 times
닫기	7389 times
무료배송	6343 times
관심상품	4602 times
팝업	3957 times
추가	2766 times
있습니다	2761 times
스마일클럽	2643 times
감소	2639 times
증가	2625 times
장바구니	2465 times
따라	2174 times
상세보기	2095 times
관한	1992 times
선택하기	1969 times
열기	1846 times
바로가기	1772 times
등록	1707 times
결제	1677 times
담기	1664 times
내일(금)	1620 times
도착예정	1615 times
판매자	1535 times
적립	1479 times
안내	1464 times
g마켓	1402 times
구매	1387 times
브랜드	1368 times
plus	1324 times

https://www.cpp.edu/	
word	# of times
the	9336 times
and	7981 times
to	6895 times
of	5526 times
for	3037 times
in	2955 times
pomona	2634 times
your	2022 times
students	1950 times
student	1938 times
you	1932 times
about	1928 times
poly	1848 times
cal	1840 times
university	1658 times
our	1655 times
is	1601 times
on	1370 times
information	1346 times
we	1338 times
are	1328 times
or	1325 times
campus	1215 times
pm	1194 times
at	1177 times
with	1129 times
that	1055 times
services	1042 times
be	1019 times

tv	1098 times		버튼	1287 times		academic	1007 times
famosos	1097 times		등록되지	1277 times		privacy	983 times
gratis	1096 times		않음	1277 times		all	970 times
información	1095 times		정보	1241 times		college	937 times
personal	1093 times		자세히보기	1223 times		will	921 times
política	1089 times		배송	1081 times		by	858 times
tus	1086 times		상품금액	1006 times		safety	827 times
uforia	1084 times		이내	986 times		center	818 times
newsletters	1084 times		쿠폰	971 times		website	807 times
una	1083 times		연관상품	888 times		as	804 times
le	1041 times		순위	801 times		program	796 times
lo	997 times		이상	779 times		view	786 times
fútbol	975 times		판매자에게	770 times		health	763 times
league	963 times		이전	761 times		resources	762 times
hollywood	840 times		비즈	755 times		california	753 times
cargando	834 times		하고	750 times		contact	741 times
tudn	832 times		전자상거래등에서	748 times		state	741 times
publicidad	775 times		소비자	748 times		faculty	738 times
español	771 times		법률	748 times		an	695 times
sobre	768 times		제1항	748 times		engineering	692 times
estados	766 times		동법	748 times		more	689 times
unidos	765 times		하위메뉴	740 times		if	682 times
futbol	760 times		상품평	726 times		office	682 times
como	757 times		스마일	716 times		plan	681 times
boxeo	737 times		배송비	693 times		not	677 times
final	737 times		상품의	685 times		science	673 times
sin	734 times		다음	684 times		learn	660 times
méxico	727 times		또는	682 times		programs	653 times
mundo	722 times		할인	679 times		content	649 times
vida	718 times		최대	670 times		events	649 times
está	717 times		찾기	653 times		use	646 times

tu	705 times		제공	644 times		have	643 times
disparo	705 times		대하여	644 times		aid	641 times
vs	686 times		스마일배송	643 times		us	631 times
mi	682 times		더보기	640 times		make	630 times
comparte	679 times		경우	632 times		how	630 times
tras	657 times		있으며	628 times		main	618 times
europa	656 times		등록된	623 times		it	609 times
copa	651 times		정하는	622 times		skip	607 times
pero	644 times		배송안내	611 times		experience	588 times
anota	635 times		검색	588 times		from	588 times
estadísticas	634 times		위해	556 times		financial	587 times
madrid	621 times		고객만족우수	549 times		accessibility	571 times
cruz	617 times		new	544 times		polytechnic	569 times
fue	608 times		스마일페이	532 times		may	568 times
radio	600 times		스마일클럽안내	528 times		document	566 times
archivo	593 times		방송잔여시간	528 times		feedback	564 times
acerca	591 times		스마일카드	523 times		admissions	556 times
media	587 times		서비스	520 times		this	549 times
vivo	585 times		지마켓글로벌	520 times		support	549 times
video	584 times		상품명	512 times		alumni	546 times
ciudad	576 times		취소	508 times		cpp	546 times
reacciona	570 times		특가	504 times		education	544 times
of	569 times		당사	504 times		can	540 times
estilo	565 times		상품은	502 times		rights	537 times
coronavirus	563 times		아이템카드	500 times		readers	531 times
prende	561 times		등록하기	500 times		policy	512 times
sigue	559 times		문의하기	499 times		staff	512 times
latina	556 times		추가구성	498 times		west	500 times
partido	552 times		총상품금액	496 times		ca	488 times
shows	550 times		상품금액에	496 times		temple	481 times
news	548 times		상세설명	496 times		reserved	479 times

and	547 times		제목	496 times		bronco	474 times
servicios	547 times		문의보기	496 times		news	472 times
gol	547 times		구매안전	496 times		better	461 times
primera	546 times		법정대리인이	496 times		email	459 times
not	546 times		가입하기	496 times		review	451 times
in	545 times		응답율	496 times		am	451 times
uso	545 times		재생	496 times		keep	449 times
elecciones	545 times		가능	493 times		safe	449 times
do	544 times		무이자할부	492 times		popular	442 times