Overview:

In this part of the project, the datasets provided by https://archive.org/details/stackexchange was used. The data consist of up to 83 gigs of xml tables, and they are divided into 327 categories. The python script for this project is able to Extract-Transform-Load all categories at once or a specify category of choice. Choosing a category is strategically crucial when it comes to question and answer text analysis. For the extent of this project, I chose the category "apple" since it's about the very commercially popular company. I believe that the questions asked about "apple" would have benefitted since its more likely the questions will be "what, when, why" instead of "how" as one would find in the "askubuntu.com" category. It would be much harder to quantify knowledge gained from "how" questions compared to a "what" as it tends to be more objectively true.

Process:

Step 1: The user can choose up to 327 different categories they wise to have their texts and features extracted by extracting the xml folders into a path of their choice. I also removed some redundancy in the folder names for a cleaner and neater look.

```
class ETL:

    path = ('\\stackexchange')
    folders = os.listdir(path)
    categories = []

    def rename_files(path):
        removal = '.stackexchange.com'
        for folder in folders:  # keep this one
            os.rename(os.path.join(path, folder), os.path.join(path, folder.replace(removal, ' ')))

    def create_categories(folders):
        for a in folders:
            categories.append(a)
```

Step 2: The features then are selected, divided and extracted into a pandas dataframe

```
for root, dirname, filenames in os.walk(path):
    for filename in fnmatch.filter(filenames, 'Posts.*'):
        realname = os.path.join(root, filename)
        posts_path.append(realname)
for i in posts_path:
    posts_category_parsed.append(et.parse(i).getroot().findall('row'))
for j in posts_category_parsed:
    for row in j:
        posts_id_count_list.append(row.get('Id'))
        post_id_list.append(row.get('Id'))
        post_type_list.append(row.get('PostTypeId'))
        body_list.append(row.get('Body').encode('utf8'))
        views_list.append(row.get('ViewCount'))
        score_list.append(row.get('Score'))
        fav_list.append(row.get('FavoriteCount'))
        parent_list.append(row.get('ParentId'))
        userID_list.append(row.get('OwnerUserId'))
        answercount_list.append(row.get('AnswerCount'))
        commentcount_list.append(row.get('CommentCount'))
```

```
col_id = pd.Series(post_id_list)
col_type = pd.Series(post_type_list)
col_body = pd.Series(body_list)
col_views = pd.Series(views_list)
col_score = pd.Series(score_list)
col_fav = pd.Series(fav_list)
col_parentid = pd.Series(parent_list)
col_ownerid = pd.Series(userID_list)
col_answer = pd.Series(answercount_list)
col_comment = pd.Series(commentcount_list)

df_posts = pd.concat([col_id, col_type, col_body, col_views, col_score, col_fav, col_parentid, col_owner
                      col_answer, col_comment], join='outer', axis=1)
df_posts.columns = ['post_id', 'post_type', 'body', 'view_count', 'score', 'fav_count',
                    'parent_question_id',
                    'owner_user_id', 'answer_count', 'comment_count']
```

```
for l, m in enumerate(posts_id_count_list):
    if m == '1':
        posts_stop_list.append(l)

index_list = df_posts.index.tolist()

posts_stop_list.append(index_list[-1] + 1)

posts_stop_list2.append([y - x for x, y in zip(posts_stop_list, posts_stop_list[1:])])

for m in posts_stop_list2:
    for o in m:
        posts_stop_list3.append(o)

for p, q in zip(categories, posts_stop_list3):
    posts_final += [p for _ in range(q)]

for r, s in zip(posts_final, posts_id_count_list):
    posts_combined += [r + '_' + str(s)]

category = pd.Series(posts_combined)
df_posts['category'] = posts_combined
```

Step 3:

The pandas dataframes are split into questions and answers then uploaded to a mysql local server, although the same script will also work on the Ubuntu linux server.

```
engine = create_engine('mysql+pymysql://Ming:asdfg@localhost:3306/ist343')
engine2 = create_engine('mysql+mysqlconnector://ssh mingting@104.40.74.232/VotingBehaviorStudy')

df_q.to_sql(name='se_questions', con=engine, flavor='mysql', index=False)
df_a.to_sql(name='se_answers', con=engine, flavor='mysql', index=False)
```

Step 4:

Since I am using a local server for this demonstration, I will use mysql workbench. The following is a script that I will make a many to one join and output a .csv file to be used for analysis.

```
DROP TABLE IF EXISTS tfidf_q;
CREATE TABLE ist343.tfidf_q
(
Q_ID int(11) NOT NULL,
Q_BODY LONGTEXT NOT NULL,
primary key (Q_ID)
);

DROP TABLE IF EXISTS tfidf_a;
CREATE TABLE ist343.tfidf_a
(
PARENT_ID int(11) NOT NULL,
A_Body LONGTEXT,
foreign key (PARENT_Id) references ist343.tfidf_q(Q_ID)
);

DROP TABLE IF EXISTS tfidf;
CREATE TABLE ist343.tfidf AS
(
SELECT * FROM IST343.tfidf_q
LEFT JOIN ist343.tfidf_a on ist343.tfidf_q.Q_ID = ist343.tfidf_a.PARENT_ID
);

INSERT INTO ist343.tfidf_q(q_id, q_body)
SELECT ist343.se_questions.post_id, ist343.se_questions.body
FROM ist343.se_questions;

INSERT INTO ist343.tfidf_a(parent_id, a_body)
SELECT ist343.se_answers.parent_question_id, ist343.se_answers.body
FROM ist343.se_answers;

SELECT ist343.tfidf_q.q_id, ist343.tfidf_q.q_body, ist343.tfidf_a.parent_id, ist343.tfidf_a.a_body
FROM ist343.tfidf_q, ist343.tfidf_a
WHERE ist343.tfidf_q.q_id = ist343.tfidf_a.PARENT_ID;
```

Table view of the answers table:

post_id	post_type	body	view_count	score	fav_count	parent_question_id	owner_user_id	answer_count	com
6	2	One option would be to clone your startup drive to an external disk using something like Sup...	NULL	5	NULL	4	38	NULL	1
7	2	I originally wanted to do this with my first Mac a couple years ago as well, since that's how my...	NULL	5	NULL	5	41	NULL	4
8	2	Microsoft has dragged their heels on support for RTL languages such as Hebrew and Arabic f...	NULL	2	NULL	3	43	NULL	0
9	2	Hardware-wise: PowerPC is a microprocessor developed mainly by the three developing com...	NULL	16	NULL	1	18	NULL	0
13	2	"Why" is a question for Microsoft, but I'm guessing it boils down to a simple lack of resources ...	NULL	4	NULL	3	21	NULL	2
14	2	PPC Macs refers to the generation of Macintosh computers created in the mid to late 1990s t...	NULL	10	NULL	1	17	NULL	0
17	2	You can use <a href="http://www5e.biglobe.ne.jp/~arcana/index.en.html">StartSound.Pref...	NULL	26	NULL	11	22	NULL	7
18	2	One thing I know is that PPCs are big endian by default, but <a href="http://en.wikipedia.org...	NULL	0	NULL	1	50	NULL	2
20	2	From the end user point of view, you don't need to worry about it much. Many applications ...	NULL	1	NULL	1	48	NULL	0
22	2	No. Even restarting, complete power down and power up, will not remove the background a...	NULL	2	NULL	12	18	NULL	1
23	2	This is so called "Hibernation" (my first met in windows). When battery dies, the OS dumps w...	NULL	3	NULL	19	14	NULL	1
24	2	When it comes to Apple hardware, the differences between the last generation of PowerPC a...	NULL	6	NULL	1	21	NULL	0
25	2	This feature is called Safe Sleep. Apple notebooks will keep the RAM contents alive (sleep in PC...	NULL	7	NULL	19	57	NULL	0
26	2	Architecture: PowerPC: (short for Performance Optimization With Enhanced RISC ...	NULL	1	NULL	1	31	NULL	0
27	2	They aren't really "in memory," more like cached to disk if and when necessary. Many apps d...	NULL	3	NULL	12	10	NULL	0
31	2	Not with the default Apple Spaces. There is an alternative, though. CocoaBots mak...	NULL	4	NULL	10	18	NULL	0
32	2	There are two ways to approach this question. From the end user's perspective, t...	NULL	5	NULL	12	21	NULL	0
36	2	I haven't tried these but Jucy and EiskaltDC++ are two DirectConnect clients that work on m...	NULL	1	NULL	21	26	NULL	2
37	2	<a href="http://forums.macrumors.com/showthread.php?t=527801" rel="nofollow">Accor...	NULL	11	NULL	33	67	NULL	1
38	2	Boot with a Mac OS X DVD, then there's an option to change passwords. <img src...	NULL	20	NULL	34	35	NULL	1
39	2	AFAIK - there's word-of-the-day screensaver that comes packed with every Mac OS X (since ...	NULL	5	NULL	16	14	NULL	1
41	2	You can reboot into single user mode ...	NULL	7	NULL	34	10	NULL	0
42	2	Yes, Apple's warranties are international. See <a href="http://www.apple.com/legal...	NULL	15	NULL	30	17	NULL	0
43	2	Use the tools menu --> disk utility to erase the partition you want to install Snow Leopard on....	NULL	0	NULL	40	26	NULL	0

Table view of the newly created many answers to one question relationship table for cosine similarity:

q_id	q_body	parent_id	a_body
1	What is the hardware and software differences between Intel and PPC Macs?	1	Hardware-wise: PowerPC is a microprocessor developed mainly by the three developing compani...
1	What is the hardware and software differences between Intel and PPC Macs?	1	PPC Macs refers to the generation of Macintosh computers created in the mid to late 1990s thro...
1	What is the hardware and software differences between Intel and PPC Macs?	1	One thing I know is that PPCs are big endian by default, but <a href="http://en.wikipedia.org/wi...
1	What is the hardware and software differences between Intel and PPC Macs?	1	From the end user point of view, you don't need to worry about it much. Many applications wer...
1	What is the hardware and software differences between Intel and PPC Macs?	1	When it comes to Apple hardware, the differences between the last generation of PowerPC and t...
1	What is the hardware and software differences between Intel and PPC Macs?	1	Architecture: PowerPC: (short for Performance Optimization With Enhanced RISC – P...
1	What is the hardware and software differences between Intel and PPC Macs?	1	The Intel chips at the time of the transition were sourced to be far more thermal and power effici...
1	What is the hardware and software differences between Intel and PPC Macs?	1	Power PC has its unique set of instruction in which overall is labeled RISC architecture and the wa...
1	What is the hardware and software differences between Intel and PPC Macs?	1	I also wanted to know more on the Power architecture, I did find some good info on it. I'm glad ...
2	The VPN software I use for work (<a href="http://www.lobotomo.com/products/IPS...	2	There is no supported way to do this. Having said that, you can do it using <code>scutil</code...
3	I have Microsoft Office/2008 on my MacBook Pro. Office doesn't support RTL langua...	3	Microsoft has dragged their heels on support for RTL languages such as Hebrew and Arabic for y...
3	I have Microsoft Office/2008 on my MacBook Pro. Office doesn't support RTL langua...	3	"Why" is a question for Microsoft, but I'm guessing it boils down to a simple lack of resources on ...
3	I have Microsoft Office/2008 on my MacBook Pro. Office doesn't support RTL langua...	3	Microsoft probably doesn't have the manpower and it uses custom code for layout, plus the mar...
4	I had a power failure and upon rebooting noticed that the OS drive needed to be rep...	4	One option would be to clone your startup drive to an external disk using something like SuperD...
4	I had a power failure and upon rebooting noticed that the OS drive needed to be rep...	4	One option that doesn't require any external drives or disks: Disk Utility's repair disk is l...
5	I will often click on a button expecting it to be clicked but instead all that happens is th...	5	I originally wanted to do this with my first Mac a couple years ago as well, since that's how my Lin...
5	I will often click on a button expecting it to be clicked but instead all that happens is th...	5	This is freely possible for the Terminal and X11 : <blockquote> defaults write com.ap...
5	I will often click on a button expecting it to be clicked but instead all that happens is th...	5	Best little utility I stumbled upon is <a href="http://coderage-software.com/zooom/">Zooom/2...
5	I will often click on a button expecting it to be clicked but instead all that happens is th...	5	Amethyst (<a href="https://github.com/ianyh/Amethyst" rel="nofollow">https://github.com/ia...
10	In Spaces it's possible to specify which space a given application will open on -- for ex...	10	Not with the default Apple Spaces. There is an alternative, though. CocoaBots makes ...
10	In Spaces it's possible to specify which space a given application will open on -- for ex...	10	Stay App sounds like it might do w...
11	Everytime I turn on my Macbook Pro it makes a start up noise. This is annoying since...	11	You can use <a href="http://www5e.biglobe.ne.jp/~arcana/index.en.html">StartSound.PrefPan...
11	Everytime I turn on my Macbook Pro it makes a start up noise. This is annoying since...	11	I haven't noticed that sound on my MacBook Pro for ages, and today I figured out why. The MB...
11	Everytime I turn on my Macbook Pro it makes a start up noise. This is annoying since...	11	Open Terminal.app and type: <blockquote> sudo -s </blockquote> Give ...
11	Everytime I turn on my Macbook Pro it makes a start up noise. This is annoying since...	11	For Snow Leopard and earlier machines download and install "StartupSound.prefPane" which will i...