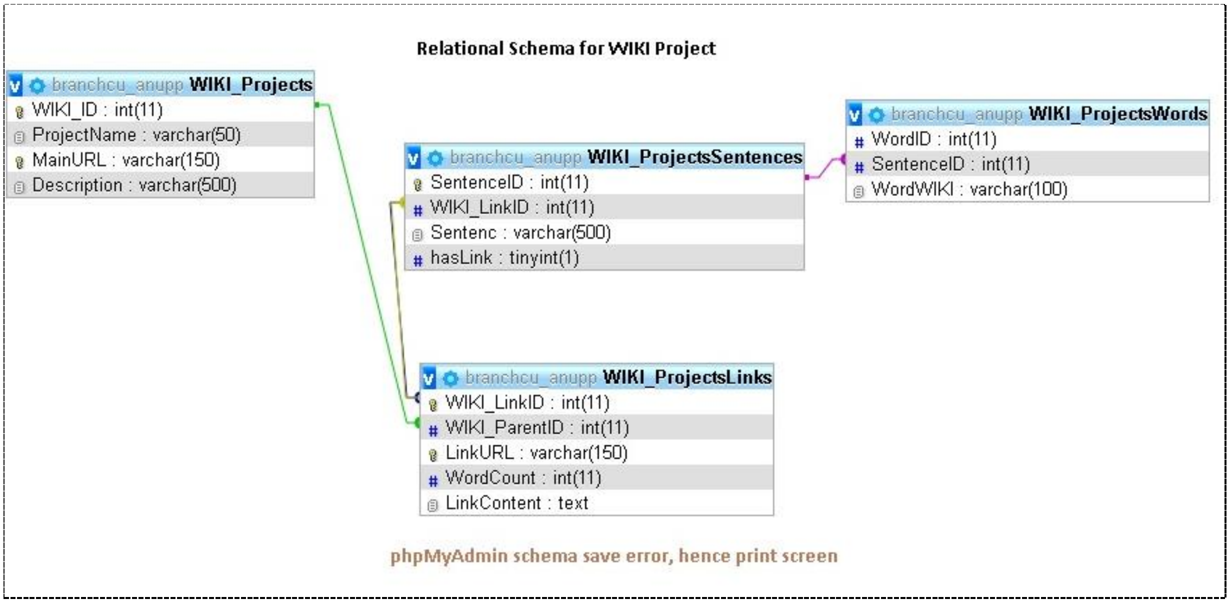
CODE CHALLENGE WIKI FOR WWI

Part 1

For the Wikipedia page on World War 1, https://en.wikipedia.org/wiki/World_War_I, download and store the main and all linked pages (just go one level in). From this dataset, design and create a data model for answering the questions below. Load the data into the data model to answer the questions below using the SQL flavor of your choice.

Response:



Relational Schema for WIKI Project

branchcu_anupp **WIKI_Projects**
WIKI_ID : int(11)
ProjectName : varchar(50)
MainURL : varchar(150)
Description : varchar(500)

branchcu_anupp **WIKI_ProjectsSentences**
SentenceID : int(11)
WIKI_LinkID : int(11)
Sentenc : varchar(500)
hasLink : tinyint(1)

branchcu_anupp **WIKI_ProjectsWords**
WordID : int(11)
SentenceID : int(11)
WordWIKI : varchar(100)

branchcu_anupp **WIKI_ProjectsLinks**
WIKI_LinkID : int(11)
WIKI_ParentID : int(11)
LinkURL : varchar(150)
WordCount : int(11)
LinkContent : text

phpMyAdmin schema save error, hence print screen

Part 2

Q1) What are the top 10 most used words on these pages and the total number of times they appear? Exclude articles and conjunctions.

Response

1) War: 4341 times

2) German: 4091

3) Army: 3028

4) British: 3017

5) French: 2180

6) World: 1879

7) City: 1877

8) Ottoman: 1853

9) Battle: 1665

10) Division: 1631

Q2) What are the top 10 countries referenced on these pages and what is the average length of the sentence they appear in?

Response

count(WordWIKI)	Avg(Char_length(Sentenc))	Country
1236	179.4676	Germany
901	177.7725	France
789	170.3308	Russia
538	173.8755	Italy
420	178.8381	Britain
348	175.9023	Serbia
337	180.3858	Bosnia
313	209.8147	Jordan
298	162.8389	Ireland
296	183.8986	Austria

SQL USED:

```
Select count(WordWIKI), Avg(Char_length(Sentenc)), Country from ltb_countries ltc inner join
WIKI_ProjectsWords WW on ltc.Country = WW.WordWiki inner join WIKI_ProjectsSentences Sen on
WW.SentenceID=Sen.SentenceID group by Country order by count(wordWIKI) desc limit 30
```

Q3) For the main WW1 page, estimate how many words are used to describe each year of the war (1914 – 1918) and the top 5 locations referenced for each year? *Bonus, test the code on the main WW2 page and see if the results make sense given the differences and years in the wars.

SQL Used

Sum(Char_Length(Sentenc))	WORDWIKI
3020	1914
3736	1915
3410	1916
6056	1917
6397	1918

```
Select Sum(Char_Length(Sentenc)), WORDWIKI from (Select SentenceID, WordWiki from WIKI_ProjectsWo
rds where wordWIKI IN ('1914', '1915', '1916', '1917', '1918')) WWINNER join WIKI_ProjectsSentenc
es Sen on WW.SentenceID = Sen.SentenceID inner join WIKI_ProjectsLinks WL on WL.WIKI_LinkID= Sen.
WIKI_LinkID whereLinkURL='/wiki/World_War_I' group by WORDWIKI
```

Top 5 locations for each year:

contcount	CNTRY	WWI_Years
2	Britain	1914
2	Ireland	1914
1	Canada	1914
1	Serbia	1914
1	America	1914
4	Italy	1915

contcount	CNTRY	WWI_Years
3	Germany	1915
3	Britain	1915
2	France	1915
1	England	1915
1	Germany	1916
1	Macedonia	1916
1	Britain	1916
1	Ireland	1916
1	England	1916
3	Russia	1917
2	France	1917
1	Latvia	1917
1	Germany	1917
1	Estonia	1917
24	Germany	1918

```
(Select count(WAgain.WordWIKI) as contcount, WAgain.WordWIKI as CNTRY, WWI_Years from (Select
SentenceID, WordWiki as WWI_Years from WIKI_ProjectsWords where wordWIKI = ('1914')) WW INNER
join WIKI_ProjectsSentences Sen on WW.SentenceID = Sen.SentenceID inner join  WIKI_ProjectsLinks
WL on WL.WIKI_LinkID= Sen.WIKI_LinkID inner join WIKI_ProjectsWords WAgain on Sen.SentenceID =
WAgain.SentenceID inner join ltb_countries ltbc on ltbc.Country=WAgain.WordWIKI

 where LinkURL='/wiki/World_War_I' group by WAgain.WordWIKI order by count(WAgain.WordWIKI)
desc limit 5)

 Union All

 (Select  count(WAgain.WordWIKI) as contcount, WAgain.WordWIKI as CNTRY, WWI_Years from (Select
SentenceID, WordWiki as WWI_Years from WIKI_ProjectsWords where wordWIKI = ('1915')) WW INNER
join WIKI_ProjectsSentences Sen on WW.SentenceID = Sen.SentenceID inner join  WIKI_ProjectsLinks
WL on WL.WIKI_LinkID= Sen.WIKI_LinkID inner join WIKI_ProjectsWords WAgain on Sen.SentenceID =
WAgain.SentenceID inner join ltb_countries ltbc on ltbc.Country=WAgain.WordWIKI
```

```
 where LinkURL='/wiki/World_War_I' group by WAgain.WordWIKI order by count(WAgain.WordWIKI)
desc limit 5)

 Union all

 (Select count(WAgain.WordWIKI) as contcount, WAgain.WordWIKI as CNTRY, WWI_Years from (Select
SentenceID, WordWiki as WWI_Years from WIKI_ProjectsWords where wordWIKI = ('1916')) WW INNER
join WIKI_ProjectsSentences Sen on WW.SentenceID = Sen.SentenceID inner join  WIKI_ProjectsLinks
WL on WL.WIKI_LinkID= Sen.WIKI_LinkID inner join WIKI_ProjectsWords WAgain on Sen.SentenceID =
WAgain.SentenceID inner join ltb_countries ltbc on ltbc.Country=WAgain.WordWIKI

 where LinkURL='/wiki/World_War_I' group by WAgain.WordWIKI order by count(WAgain.WordWIKI)
desc limit 5)

 Union all

 (Select count(WAgain.WordWIKI) as contcount, WAgain.WordWIKI as CNTRY, WWI_Years from (Select
SentenceID, WordWiki as WWI_Years from WIKI_ProjectsWords where wordWIKI = ('1917')) WW INNER
join WIKI_ProjectsSentences Sen on WW.SentenceID = Sen.SentenceID inner join  WIKI_ProjectsLinks
WL on WL.WIKI_LinkID= Sen.WIKI_LinkID inner join WIKI_ProjectsWords WAgain on Sen.SentenceID =
WAgain.SentenceID inner join ltb_countries ltbc on ltbc.Country=WAgain.WordWIKI

 where LinkURL='/wiki/World_War_I' group by WAgain.WordWIKI order by count(WAgain.WordWIKI)
desc limit 5)

 Union all

 (Select count(WAgain.WordWIKI) as contcount, WAgain.WordWIKI as CNTRY, WWI_Years from (Select
SentenceID, WordWiki as WWI_Years from WIKI_ProjectsWords where wordWIKI = ('1918')) WW INNER
join WIKI_ProjectsSentences Sen on WW.SentenceID = Sen.SentenceID inner join  WIKI_ProjectsLinks
WL on WL.WIKI_LinkID= Sen.WIKI_LinkID inner join WIKI_ProjectsWords WAgain on Sen.SentenceID =
WAgain.SentenceID inner join ltb_countries ltbc on ltbc.Country=WAgain.WordWIKI

 where LinkURL='/wiki/World_War_I' group by WWI_Years order by count(WAgain.WordWIKI) desc limit
5)
```

Part 3

Q1) Using a scripting language of your choice and the raw html page

data, provide a list of all pages where the country “Turkey” is

referenced and estimate the number of words on the page.