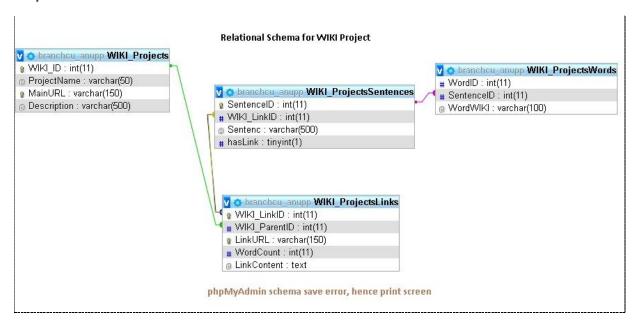
CODE CHALLENGE WIKI FOR WWI

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:



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) 1236	Avg(Char length(Sentenc)) 179.4676	<u>Country</u> 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_ProjectsWords where wordWIKI IN ('1914', '1915', '1916', '1917', '1918')) WWINNER join WIKI_ProjectsSentences 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 2	CNTRY Britain	<u>WWI_Years</u> 1914
2	Ireland	1914
1	Canada	1914
1	Serbia	1914
1	America	1914
4	Italy	1915

contcount 3	CNTRY Germany	
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 Itb_countries Itbc on Itbc.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.

Response:

```
!https://en.wikipedia.org/wiki/World War I: word count=135696
https://en.wikipedia.org/wiki/Europe: word count=120619
https://en.wikipedia.org/wiki/World War I casualties: word count=61816
https://en.wikipedia.org/wiki/Trench warfare: word count=47865
https://en.wikipedia.org/wiki/List of wars and anthropogenic disasters by death
_toll: word count=121020
https://en.wikipedia.org/wiki/World War II: word count=135301
https://en.wikipedia.org/wiki/Yugoslav nationalism: word count=20362
https://en.wikipedia.org/wiki/Sarajevo: word count=84799
https://en.wikipedia.org/wiki/July_Crisis: word count=55552
https://en.wikipedia.org/wiki/Kingdom of Serbia: word count=32271
https://en.wikipedia.org/wiki/Treaty of Brest-Litovsk: word count=29115
https://en.wikipedia.org/wiki/Interwar_period: word count=21552
https://en.wikipedia.org/wiki/Ernst Haeckel: word count=32238
https://en.wikipedia.org/wiki/Balance of power (international relations): word
count=25950
https://en.wikipedia.org/wiki/Balkans: word count=78466
https://en.wikipedia.org/wiki/Decline_of_the_Ottoman_Empire: word count=32302
https://en.wikipedia.org/wiki/Triple Alliance (1882): word count=12650
https://en.wikipedia.org/wiki/Bosniaks: word count=62897
https://en.wikipedia.org/wiki/Bosnian Serbs: word count=25560
https://en.wikipedia.org/wiki/Luxembourg: word count=70143
https://en.wikipedia.org/wiki/Neutrality_(international_relations): word
count=19054
https://en.wikipedia.org/wiki/Hague Conventions of 1899 and 1907: word
count=21581
https://en.wikipedia.org/wiki/British_Army: word count=74581
https://en.wikipedia.org/wiki/Cruiser: word count=37764
https://en.wikipedia.org/wiki/Blockade of Germany: word count=11544
https://en.wikipedia.org/wiki/U-boat: word count=21667
https://en.wikipedia.org/wiki/Destroyer: word count=32390
https://en.wikipedia.org/wiki/Aircraft carrier: word count=48482
https://en.wikipedia.org/wiki/Belgrade: word count=98464
https://en.wikipedia.org/wiki/Slovenia: word count=122526
https://en.wikipedia.org/wiki/Bosnia (region): word count=9298
https://en.wikipedia.org/wiki/Principality_of_Albania: word count=24041
```

```
https://en.wikipedia.org/wiki/Salonica: word count=124190
https://en.wikipedia.org/wiki/Constantine I of Greece: word count=33393
https://en.wikipedia.org/wiki/Eleftherios Venizelos: word count=60858
https://en.wikipedia.org/wiki/National Schism: word count=14700
https://en.wikipedia.org/wiki/Athens: word count=105893
https://en.wikipedia.org/wiki/Alexander_of_Greece: word count=36759
https://en.wikipedia.org/wiki/Bitola: word count=38764
https://en.wikipedia.org/wiki/Budapest: word count=120788
https://en.wikipedia.org/wiki/Vienna: word count=76501
https://en.wikipedia.org/wiki/Central_Powers: word count=31978
https://en.wikipedia.org/wiki/Caucasus: word count=44033
https://en.wikipedia.org/wiki/Armenians: word count=44253
https://en.wikipedia.org/wiki/Greeks: word count=91523
https://en.wikipedia.org/wiki/Assyrian_people: word count=63702
https://en.wikipedia.org/wiki/Armenian_Genocide: word count=115654
https://en.wikipedia.org/wiki/Greek_Genocide: word count=52628
https://en.wikipedia.org/wiki/Assyrian_Genocide: word count=44018
https://en.wikipedia.org/wiki/Gallipoli Campaign: word count=80670
https://en.wikipedia.org/wiki/Mesopotamian_campaign: word count=29284
https://en.wikipedia.org/wiki/Mesopotamia: word count=40175
https://en.wikipedia.org/wiki/Siege of Kut: word count=16869
https://en.wikipedia.org/wiki/Baghdad: word count=53693
https://en.wikipedia.org/wiki/ANZAC Mounted Division: word count=32938
https://en.wikipedia.org/wiki/Egyptian Expeditionary Force: word count=38470
https://en.wikipedia.org/wiki/Sinai Peninsula: word count=26763
https://en.wikipedia.org/wiki/Sinai: word count=26793
https://en.wikipedia.org/wiki/Enver_Pasha: word count=25436
https://en.wikipedia.org/wiki/Battle of Sarikamish: word count=29267
https://en.wikipedia.org/wiki/Sherif Hussein: word count=14972
https://en.wikipedia.org/wiki/Mecca: word count=62413
https://en.wikipedia.org/wiki/Fakhri Pasha: word count=8593
https://en.wikipedia.org/wiki/Medina: word count=34756
https://en.wikipedia.org/wiki/Siege_of_Medina: word count=8233
https://en.wikipedia.org/wiki/Senussi: word count=23394
https://en.wikipedia.org/wiki/Senussi Campaign: word count=34094
https://en.wikipedia.org/wiki/Treaty of London (1915): word count=9791
```

```
https://en.wikipedia.org/wiki/Field Marshal: word count=14926
https://en.wikipedia.org/wiki/Ljubljana: word count=101510
https://en.wikipedia.org/wiki/Alpini: word count=29251
https://en.wikipedia.org/wiki/Bolshevik: word count=20765
https://en.wikipedia.org/wiki/Treaty of Bucharest (1918): word count=17536
https://en.wikipedia.org/wiki/Warsaw: word count=109879
https://en.wikipedia.org/wiki/Brusilov Offensive: word count=19639
https://en.wikipedia.org/wiki/Bucharest: word count=69264
https://en.wikipedia.org/wiki/Nicholas II of Russia: word count=72185
https://en.wikipedia.org/wiki/Saint Petersburg: word count=103819
https://en.wikipedia.org/wiki/Allied intervention in the Russian Civil War: word
count=23862
https://en.wikipedia.org/wiki/Czechoslovakia: word count=28386
https://en.wikipedia.org/wiki/Russia: word count=185148
https://en.wikipedia.org/wiki/France: word count=156243
https://en.wikipedia.org/wiki/Italy: word count=133324
https://en.wikipedia.org/wiki/Austria-Hungary: word count=103244
https://en.wikipedia.org/wiki/Bolsheviks: word count=20734
https://en.wikipedia.org/wiki/David_Lloyd_George: word count=76238
https://en.wikipedia.org/wiki/German General Staff: word count=19933
https://en.wikipedia.org/wiki/Supreme War Council: word count=5464
https://en.wikipedia.org/wiki/Sinai and Palestine Campaign: word count=76514
https://en.wikipedia.org/wiki/XX Corps (United Kingdom): word count=5099
https://en.wikipedia.org/wiki/Desert Mounted Corps: word count=38982
https://en.wikipedia.org/wiki/Jerusalem: word count=144260
https://en.wikipedia.org/wiki/Friedrich Freiherr Kress von Kressenstein: word
count=7095
https://en.wikipedia.org/wiki/Cevat %C3%87obanl%C4%B1: word count=7196
https://en.wikipedia.org/wiki/Ottoman Army: word count=21908
https://en.wikipedia.org/wiki/Otto Liman von Sanders: word count=8642
https://en.wikipedia.org/wiki/Battle_of_Megiddo_(1918): word count=37058
https://en.wikipedia.org/wiki/Battle of Tulkarm (1918): word count=29305
https://en.wikipedia.org/wiki/Battle of Tabsor (1918): word count=34794
https://en.wikipedia.org/wiki/Battle of Arara: word count=22174
https://en.wikipedia.org/wiki/Battle_of_Nablus_(1918): word count=39684
https://en.wikipedia.org/wiki/Battle of Samakh (1918): word count=26850
```

```
https://en.wikipedia.org/wiki/Damascus: word count=57169
https://en.wikipedia.org/wiki/Amman: word count=69563
https://en.wikipedia.org/wiki/Armistice of Mudros: word count=15641
https://en.wikipedia.org/wiki/Aleppo: word count=80133
https://en.wikipedia.org/wiki/United States Navy: word count=80449
https://en.wikipedia.org/wiki/Submarine: word count=48382
https://en.wikipedia.org/wiki/US Marines: word count=87467
https://en.wikipedia.org/wiki/Krupp: word count=22920
https://en.wikipedia.org/wiki/South Caucasus: word count=29546
https://en.wikipedia.org/wiki/First Republic of Armenia: word count=27684
https://en.wikipedia.org/wiki/Azerbaijan Democratic Republic: word count=28184
https://en.wikipedia.org/wiki/Democratic_Republic_of_Georgia: word count=22210
https://en.wikipedia.org/wiki/South West Caucasian Republic: word count=9540
https://en.wikipedia.org/wiki/Transcaucasian_Democratic_Federative_Republic:
word count=15275
https://en.wikipedia.org/wiki/Treaty_of_Poti: word count=11089
https://en.wikipedia.org/wiki/Battle_of_Sardarabad: word count=16475
https://en.wikipedia.org/wiki/Dominion: word count=36204
https://en.wikipedia.org/wiki/Kiel: word count=32887
https://en.wikipedia.org/wiki/Social Democratic Party of Germany: word
count=29273
https://en.wikipedia.org/wiki/The Netherlands: word count=125678
https://en.wikipedia.org/wiki/Saloniki: word count=124192
https://en.wikipedia.org/wiki/Armistice with Germany: word count=21447
https://en.wikipedia.org/wiki/House of Romanov: word count=36197
https://en.wikipedia.org/wiki/House of Hohenzollern: word count=43085
https://en.wikipedia.org/wiki/Habsburg: word count=80413
https://en.wikipedia.org/wiki/Ottoman Dynasty: word count=63194
https://en.wikipedia.org/wiki/Treaty of Versailles: word count=50555
https://en.wikipedia.org/wiki/Termination_of_the_Present_War_(Definition)_Act_1
918: word count=5554
https://en.wikipedia.org/wiki/Treaty of Lausanne: word count=22767
https://en.wikipedia.org/wiki/War_memorial: word count=17144
https://en.wikipedia.org/wiki/Istanbul: word count=125996
https://en.wikipedia.org/wiki/Paris Peace Conference, 1919: word count=36953
```

https://en.wikipedia.org/wiki/League of Nations: word count=78108

https://en.wikipedia.org/wiki/Article_231_of_the_Treaty_of_Versailles: word count=26399

https://en.wikipedia.org/wiki/World_War_I_reparations: word count=36984

https://en.wikipedia.org/wiki/Greater_Romania: word count=23268

https://en.wikipedia.org/wiki/Levant: word count=20739

https://en.wikipedia.org/wiki/Anatolia: word count=45356

https://en.wikipedia.org/wiki/Treaty_of_S%C3%A8vres: word count=31268

https://en.wikipedia.org/wiki/Turkish National Movement: word count=10781

https://en.wikipedia.org/wiki/Turkish_War_of_Independence: word count=44905

https://en.wikipedia.org/wiki/Lloyd_George: word count=76266

https://en.wikipedia.org/wiki/Chaim_Weizmann: word count=28145

https://en.wikipedia.org/wiki/Balfour_Declaration: word count=96590

https://en.wikipedia.org/wiki/Jewish_homeland: word count=18927

https://en.wikipedia.org/wiki/Hague_Convention_of_1907: word count=21640

https://en.wikipedia.org/wiki/Sir_Douglas_Haig: word count=71668

https://en.wikipedia.org/wiki/Conscription: word count=59254

https://en.wikipedia.org/wiki/Allied_Supreme_Council: word count=5506

https://en.wikipedia.org/wiki/lyasu_V: word count=16915

https://en.wikipedia.org/wiki/Ethiopia: word count=96639

https://en.wikipedia.org/wiki/Dervish_state: word count=16336

https://en.wikipedia.org/wiki/Addis_Ababa: word count=41162

https://en.wikipedia.org/wiki/Ethiopian_Empire: word count=28917

https://en.wikipedia.org/wiki/Syndicalism: word count=17091

https://en.wikipedia.org/wiki/Bertrand_Russell: word count=90198

https://en.wikipedia.org/wiki/Martial_law: word count=20453

https://en.wikipedia.org/wiki/Russian_Empire: word count=64017

https://en.wikipedia.org/wiki/Commonwealth_War_Graves_Commission: word count=40410

https://en.wikipedia.org/wiki/Tomb_of_the_Unknown_Soldier: word count=20124

https://en.wikipedia.org/wiki/Kansas_City,_Missouri: word count=70890

https://en.wikipedia.org/wiki/World_War_I_centenary: word count=20485

https://en.wikipedia.org/wiki/Posttraumatic_stress_disorder: word count=88188

https://en.wikipedia.org/wiki/Suffragette: word count=30073