طراحی، ساخت و فراهم کردن محتوی برای بانک اطلاعاتی رابطه‌ای استخراج شده از یاگو

مقدمه

هدف این تمرین کارکردن با داده‌های استخراج شده از وب معنایی است. همانطور که می‌دانید صفحات وب (از جمله دانش‌نامه‌هایی مانند ویکی‌پدیا) مملو از اطلاعات و داده‌های بدون ساختار هستند. ساختارمند کردن این داده‌ها از آرزوهای بشری است؛ زیرا در آن صورت به صورت سیستماتیک قابل استفاده خواهند بود. از اولین گام‌هایی که در این راه برداشته شده، استخراج داده‌ها با استفاده از پردازش زبان طبیعی است که نتایج آن (هرچند ابتدایی) هیجان‌انگیز است! در این پروژه، قصد داریم با داده‌های یکی از این منابع دانشی که از ویکی‌پدیا استخراج شده کار کنیم. نام این بانک دانش یاگو 3 است که در موسسه ماکس پلانک آلمان تولید شده است (سایت یاگو). از ویژگی‌های مهم یاگو می‌توان به دو نکته اشاره کرد: اولا، زبان فارسی را نیز در استخراج داده‌ها مورد توجه قرار داده است. ثانیا داده‌ها با دقت بالایی صحت‌سنجی شده‌اند و داده‌های بعضا غلط در ویکی‌پدیا را فیلتر کرده‌اند (با دقت نزدیک 95 درصد). ما در این پروژه قصد داریم داده‌های استخراج شده موجود در این بانک دانش را به فرمت یک بانک اطلاعاتی رابطه‌ای در آوریم تا استفاده از داده‌ها آسان‌تر و در دسترس‌تر باشد.

گام‌ها

اولین گام، بدست‌آوردن داده‌های مورد نیاز است. کل داده‌ها در فرمت (TSV-tab separated value) در دسترس همگان قرار گرفته است. می‌توانید داده‌ها را از سایت یاگو دانلود کنید. البته نیاز نیست شما این کار را انجام بدین، من این کار را کردم و فایل‌هایی را که فکر میکردم به درد پروژه بخورد را جدا کرده و در یک فایل زیپ داخل دراپ باکس درس قرار دادم (شما محدود به این فایل ها نیستید و میتوانید از فایل های دیگر بنا به نیاز استفاده کنید). فایل‌ها به صورت تک تک هم در سایت یاگو در دسترس هستند پیوند). هر کدام از این فایل های tsv تنها یک جنبه از داده را نشان می‌دهند. فرمت آن‌ها به این صورت است که هر خط متشکل از یک چندتایی است (3 یا 4 تایی) که با تب از هم جدا شده‌اند. به طور مثال فایل **yagoLabels.tsv** شامل تمام برچسب‌های استخراج شده از ویکی‌پدیا است که سطری مانند زیر در آن وجود دارد. (مثلا معنای سطر زیر این است که بر چسب "حسن روحانی" در زبان فارسی (fas) متعلق است به موجودیتی با نام Hassan_Rouhani).

```
row: <id_j6oh03_1ia_10zjm53>    <Hassan_Rouhani>    rdfs:label    "حسن روحانی"@fas
```

گام بعدی این است فایل های بدست آمده را با یک زبان برنامه نویسی پردازش کنید و یا در یک بانک اطلاعاتی رابطه ای بارگذاری کنید و از زبان SQL و امکانات تطبیق رشته ها مانند like استفاده کنید تا به اهدافی که در زیر به آنها اشاره شده برسید: در این پروژه از شما انتظار میرود **اطلاعات موجود پیرامون اتفاقات سیاسی و سیاسیون** را پیدا کرده و در رابطه های یک بانک اطلاعاتی دسته بندی کنید. پس یکم، **کار از جنس گشتن هست**، دوم، **طراحی بانک اطلاعاتی رابطه ای** مطرح است، و سوم، **فراهم آوردن محتوا برای رابطه‌ها در این بانک اطلاعاتی** دارای اهمیت می باشد.

من پروسه گشتن را به نحوی که خودم انجام میدهم، برای شما با یک مثال شرح میدهم. این میتواند یک روش کار برای شما باشد. من اول از برچسب ها استفاده میکنم. مثلا میدانم که حسن روحانی یک سیاست مدار است و به احتمال زیاد در داخل بانک دانش یاگو اطلاعاتی برای وی موجود باشد. پس به سراغ فایل **yagoLabels.tsv** میروم به دنبال تمام رکوردهایی میگردم که با "حسن روحانی" منطبق می شوند:

```
<id_j6oh03_qkd_1r9pjxd>    <Hassan_Rouhani>    <redirectedFrom>    "شیخ حسن روحانی"@eng
```

```
<id_128kwea_1ia_1h6kyl5>	<Hassan_Rouhani_presidential_campaign,_2013>	rdfs:label	"حسن روحانی در انتخابات ریاست‌جمهوری ایران (۱۳۹۲)"@fas
<id_j6oh03_qkd_1t0ooja>	<Hassan_Rouhani>	<redirectedFrom>	"دکتر حسن روحانی"@fas
<id_j6oh03_qkd_1t0ooja>	<Hassan_Rouhani>	<redirectedFrom>	"دکتر حسن روحانی"@eng
<id_128kwea_qkd_pwsd95>	<Hassan_Rouhani_presidential_campaign,_2013>	<redirectedFrom>	"حسن روحانی در انتخابات ریاست‌جمهوری ایران (1392)"@fas
<id_j6oh03_qkd_1yuye2g>	<Hassan_Rouhani>	<redirectedFrom>	"دکترحسن روحانی"@fas
<id_128kwea_1ia_1q1kp9>	<Hassan_Rouhani_presidential_campaign,_2013>	rdfs:label	"حسن روحانی در انتخابات ریاست‌جمهوری ایران"@fas
<id_j6oh03_qkd_10zjm53>	<Hassan_Rouhani>	<redirectedFrom>	"حسن روحانی"@eng
<id_j6oh03_qkd_10zjm53>	<Hassan_Rouhani>	<redirectedFrom>	"حسن روحانی"@ara
<id_128kwea_qkd_gfn7np>	<Hassan_Rouhani_presidential_campaign,_2013>	<redirectedFrom>	"حسن روحانی در انتخابات ریاست جمهوری ایران (1392)"@fas
<id_j6oh03_1ia_10zjm53>	<Hassan_Rouhani>	rdfs:label	"حسن روحانی"@fas
<id_j6oh03_1ia_10zjm53>	<Hassan_Rouhani>	rdfs:label	"حسن روحانی"@mzn
<id_j6oh03_1ia_10zjm53>	<Hassan_Rouhani>	rdfs:label	"حسن روحانی"@urd
<id_128kwea_qkd_17pfszp>	<Hassan_Rouhani_presidential_campaign,_2013>	<redirectedFrom>	"حسن روحانی در انتخابات ریاست جمهوری ایران (۱۳۹۲)"@fas
```

و شاید جستجوی دقیق‌تر، حتی برچسب‌ها در زبان‌های دیگر را هم پیدا کند:

```
<id_j6oh03_1ia_ma5xuc>	<Hassan_Rouhani>	rdfs:label	"ﾛｳﾊｰﾆｰ・ｴﾝ"@jpn
<id_j6oh03_1ia_1trnstj>	<Hassan_Rouhani>	rdfs:label	"哈桑·鲁哈尼"@zho
<id_128kwea_1ia_1h6kyl5>	<Hassan_Rouhani_presidential_campaign,_2013>	rdfs:label	"حسن روحانی در انتخابات ریاست‌جمهوری ایران (۱۳۹۲)"@fas
<id_j6oh03_1ia_7fzx5t>	<Hassan_Rouhani>	rdfs:label	"Χαὶν ΡουDZνί"@ell
<id_j6oh03_1ia_jprvom>	<Hassan_Rouhani>	rdfs:label	"हसन रूहानी"@mar
<id_j6oh03_1ia_jprvom>	<Hassan_Rouhani>	rdfs:label	"हसन रूहानी"@hin
<id_j6oh03_1ia_hqogo1>	<Hassan_Rouhani>	rdfs:label	"Հասան Ռոհանի"@hye
<id_j6oh03_1ia_ywgvg7>	<Hassan_Rouhani>	rdfs:label	"Hasán Rouhaní"@spa
<id_j6oh03_1ia_1y7wz1i>	<Hassan_Rouhani>	rdfs:label	"Рухани, Хасан"@rus
<id_j6oh03_1ia_10zjm1h>	<Hassan_Rouhani>	rdfs:label	"حسن روحاني"@ara
<id_j6oh03_1ia_1onmteo>	<Hassan_Rouhani>	rdfs:label	"Həsən Ruhani"@aze
<id_j6oh03_1ia_1nlrafm>	<Hassan_Rouhani>	rdfs:label	"ჰასან რუჰანი"@kat
<id_j6oh03_1ia_1m5y92o>	<Hassan_Rouhani>	rdfs:label	"Hasan Ruhani"@lat
<id_j6oh03_1ia_1m5y92o>	<Hassan_Rouhani>	rdfs:label	"Hasan Ruhani"@tur
<id_j6oh03_1ia_1m5y92o>	<Hassan_Rouhani>	rdfs:label	"Hasan Ruhani"@bos
<id_j6oh03_1ia_1m5y92o>	<Hassan_Rouhani>	rdfs:label	"Hasan Ruhani"@epo
<id_j6oh03_1ia_1o75j09>	<Hassan_Rouhani>	rdfs:label	"Hasan Rúháni"@ces
```

```
<id_j6oh03_1ia_1qwa5o5>    <Hassan_Rouhani>    rdfs:label    "Hassan Rouhani'"@bcl
<id_j6oh03_1ia_14n6gx5>    <Hassan_Rouhani>    rdfs:label    "হাসান রuyানি"@ben
<id_j6oh03_1ia_1mj7ai7>    <Hassan_Rouhani>    rdfs:label    "Хассан Рухані"@ukr
<id_j6oh03_1ia_1e5a1iq>    <Hassan_Rouhani>    rdfs:label    "Hassan Rouhani"
<id_j6oh03_1ia_64qp2c>    <Hassan_Rouhani>    rdfs:label    "Хасан Рухані"@bel
<id_j6oh03_1ia_opo8qb>    <Hassan_Rouhani>    rdfs:label    "Hasan Rouhani"@spa
<id_j6oh03_1ia_opo8qb>    <Hassan_Rouhani>    rdfs:label    "Hasan Rouhani"@pol
<id_128kwea_1ia_1q1kp9>    <Hassan_Rouhani_presidential_campaign,_2013>    rdfs:label    "حسن روحانی در انتخابات ریاستجمهوری ایران"@fas
<id_fji5ly_1ia_eurz25>    <Inauguration_of_Hassan_Rouhani>    rdfs:label    "مراسم تحلیف یازدهمین دوره ریاست جمهوری در ایران"@fas
<id_j6oh03_1ia_1amoeh6>    <Hassan_Rouhani>    rdfs:label    "חסן רוחאני"@heb
<id_128kwea_1ia_3l0ur2>    <Hassan_Rouhani_presidential_campaign,_2013>    rdfs:label    "Hassan Rouhani presidential campaign"@eng
<id_j6oh03_1ia_8hgo0b>    <Hassan_Rouhani>    rdfs:label    "Hassan Rohani"@ron
<id_j6oh03_1ia_8hgo0b>    <Hassan_Rouhani>    rdfs:label    "Hassan Rohani"@scn
<id_j6oh03_1ia_8hgo0b>    <Hassan_Rouhani>    rdfs:label    "Hassan Rohani"@por
<id_j6oh03_1ia_8hgo0b>    <Hassan_Rouhani>    rdfs:label    "Hassan Rohani"@nld
<id_j6oh03_1ia_8hgo0b>    <Hassan_Rouhani>    rdfs:label    "Hassan Rohani"@swe
<id_j6oh03_1ia_8hgo0b>    <Hassan_Rouhani>    rdfs:label    "Hassan Rohani"@afr
<id_j6oh03_1ia_8hgo0b>    <Hassan_Rouhani>    rdfs:label    "Hassan Rohani"@fry
<id_j6oh03_1ia_8hgo0b>    <Hassan_Rouhani>    rdfs:label    "Hassan Rohani"@fra
<id_j6oh03_1ia_8hgo0b>    <Hassan_Rouhani>    rdfs:label    "Hassan Rohani"@dan
<id_j6oh03_1ia_8hgo0b>    <Hassan_Rouhani>    rdfs:label    "Hassan Rohani"@deu
<id_j6oh03_1ia_11bcphx>    <Hassan_Rouhani>    rdfs:label    "حەسەن ڕۆحانی"@ckb
<id_j6oh03_1ia_64qp1i>    <Hassan_Rouhani>    rdfs:label    "Хасан Рухани"@srp
<id_j6oh03_1ia_64qp1i>    <Hassan_Rouhani>    rdfs:label    "Хасан Рухани"@bul
<id_128kwea_1ia_12dxvgi>    <Hassan_Rouhani_presidential_campaign,_2013>    rdfs:label    "Hassan Rouhani presidential campaign, 2013"@eng
<id_j6oh03_1ia_1e5a1iq>    <Hassan_Rouhani>    rdfs:label    "Hassan Rouhani"@eng
<id_j6oh03_1ia_1e5a1iq>    <Hassan_Rouhani>    rdfs:label    "Hassan Rouhani"@fin
<id_j6oh03_1ia_1e5a1iq>    <Hassan_Rouhani>    rdfs:label    "Hassan Rouhani"@cat
<id_j6oh03_1ia_1e5a1iq>    <Hassan_Rouhani>    rdfs:label    "Hassan Rouhani"@msa
<id_j6oh03_1ia_1e5a1iq>    <Hassan_Rouhani>    rdfs:label    "Hassan Rouhani"@nob
<id_j6oh03_1ia_1e5a1iq>    <Hassan_Rouhani>    rdfs:label    "Hassan Rouhani"@oci
<id_j6oh03_1ia_1e5a1iq>    <Hassan_Rouhani>    rdfs:label    "Hassan Rouhani"@pam
<id_j6oh03_1ia_1e5a1iq>    <Hassan_Rouhani>    rdfs:label    "Hassan Rouhani"@ind
<id_j6oh03_1ia_1e5a1iq>    <Hassan_Rouhani>    rdfs:label    "Hassan Rouhani"@ita
<id_j6oh03_1ia_1e5a1iq>    <Hassan_Rouhani>    rdfs:label    "Hassan Rouhani"@vie
<id_j6oh03_1ia_1e5a1iq>    <Hassan_Rouhani>    rdfs:label    "Hassan Rouhani"@yor
<id_j6oh03_1ia_1e5a1iq>    <Hassan_Rouhani>    rdfs:label    "Hassan Rouhani"@tgl
<id_j6oh03_1ia_10zjm53>    <Hassan_Rouhani>    rdfs:label    "حسن روحانی"@fas
<id_j6oh03_1ia_10zjm53>    <Hassan_Rouhani>    rdfs:label    "حسن روحانی"@mzn
```

```
<id_j6oh03_1ia_10zjm53>    <Hassan_Rouhani>    rdfs:label    "حسن روحانی"@urd
<id_j6oh03_1ia_1c85x3f>    <Hassan_Rouhani>    rdfs:label    "அசன் ர□□□னி"@tam
<id_j6oh03_1ia_10di4g3>    <Hassan_Rouhani>    rdfs:label    "Хассан Роухани"@chv
<id_fji5ly_1ia_bujaed>    <Inauguration_of_Hassan_Rouhani>    rdfs:label    "Inauguration of Hassan Rouhani"@eng
<id_j6oh03_1ia_xpsg12>    <Hassan_Rouhani>    rdfs:label    "ਹਸਨ ਰੂਹਾਨੀ"@pan
<id_j6oh03_1ia_1m2nhii>    <Hassan_Rouhani>    rdfs:label    "Hasan Rohani"@bre
<id_j6oh03_1ia_1m2nhii>    <Hassan_Rouhani>    rdfs:label    "Hasan Rohani"@hrv
<id_j6oh03_1ia_1ji8t8p>    <Hassan_Rouhani>    rdfs:label    "Рухани"@rus
<id_j6oh03_1ia_10zjm1g>    <Hassan_Rouhani>    rdfs:label    "حسن روحانی"@arz
<id_j6oh03_1ia_s93g98>    <Hassan_Rouhani>    rdfs:label    "하산 로우하니"@kor
```

از همین نتیجه متوجه می‌شوم که یکم، هویتی با نام Hassan_Rouhani در بانک دانش وجود دارد (می‌توان بعدا مطمئن شد که این از نوع wikicat_Iranian_politicians هست) و دوم، اطلاعاتی مربوط به انتخابات ریاست جمهوری در اختیار است. از اینجا به بعد روی موجودیت Hassan_Rouhani تمرکز می‌کنم (اما تمام برچسب‌های مرتبط با آن در زبان فارسی را برای این موجودیت در یک جدول نگه میدارم. نگه‌داشتن این برچسب‌ها، برای اینکه بعدا این بانک اطلاعاتی برای پردازش های زبان فارسی قابل استفاده باشد، ضروری است). بعد از اینکه از نوع موجودیت Hassan_Rouhani مطمئن شدم (با استفاده از فایل yagoTypes)، به سراغ کشف حقایق پیرامون وی میروم. تا همینجا می‌دانم که یک جدول خواهم داشت با عنوان politicians و در رابطه هست با جدول election و یک سری برچسب فارسی متناظر هم دارد.

```
<id_j6oh03_88c_k9sqxo>    <Hassan_Rouhani>    rdf:type    <wikicat_Alumni_of_Glasgow_Caledonian_University>
<id_j6oh03_88c_1hw68kf>    <Hassan_Rouhani>    rdf:type    <wikicat_Combatant_Clergy_Association_politicians>
<id_j6oh03_88c_k157oh>    <Hassan_Rouhani>    rdf:type    <wikicat_Iranian_Majlis_Deputy_Speakers>
<id_j6oh03_88c_1rb0t67>    <Hassan_Rouhani>    rdf:type    <wikicat_Iranian_Majlis_Representatives>
<id_j6oh03_88c_juljkk>    <Hassan_Rouhani>    rdf:type    <wikicat_Iranian_academics>
<id_j6oh03_88c_1cewq2z>    <Hassan_Rouhani>    rdf:type    <wikicat_Iranian_clerics>
<id_j6oh03_88c_173d4kp>    <Hassan_Rouhani>    rdf:type    <wikicat_Iranian_diplomats>
<id_j6oh03_88c_xis0s7>    <Hassan_Rouhani>    rdf:type    <wikicat_Iranian_lawyers>
<id_j6oh03_88c_ef6kon>    <Hassan_Rouhani>    rdf:type    <wikicat_Iranian_politicians>
<id_j6oh03_88c_up6ot4>    <Hassan_Rouhani>    rdf:type    <wikicat_Iranian_reformists>
<id_j6oh03_88c_1fv0w2t>    <Hassan_Rouhani>    rdf:type    <wikicat_Iranian_revolutionaries>
<id_j6oh03_88c_1jibh39>    <Hassan_Rouhani>    rdf:type    <wikicat_Iranian_scholars>
<id_j6oh03_88c_1oqwaqw>    <Hassan_Rouhani>    rdf:type    <wikicat_Iranian_writers>
<id_j6oh03_88c_12dr1lc>    <Hassan_Rouhani>    rdf:type    <wikicat_Islamic_Republican_Party_politicians>
<id_j6oh03_88c_1nlz59o>    <Hassan_Rouhani>    rdf:type    <wikicat_Living_people>
<id_j6oh03_88c_1cf6iq>    <Hassan_Rouhani>    rdf:type    <wikicat_Muslim_theologians>
<id_j6oh03_88c_1x8jksx>    <Hassan_Rouhani>    rdf:type    <wikicat_People_from_Semnan_Province>
<id_j6oh03_88c_12hu0w7>    <Hassan_Rouhani>    rdf:type    <wikicat_People_of_the_Iranian_Revolution>
<id_j6oh03_88c_ug6gob>    <Hassan_Rouhani>    rdf:type    <wikicat_Presidents_of_Iran>
<id_j6oh03_88c_161i36g>    <Hassan_Rouhani>    rdf:type    <wikicat_Scholars_of_constitutional_law>
```

```
<id_j6oh03_88c_10dtai>    <Hassan_Rouhani>    rdf:type    <wikicat_Shia_clerics>
<id_j6oh03_88c_vafydj>    <Hassan_Rouhani>    rdf:type    <wordnet_officeholder_110371450>
```

اگر کمی عمیق‌تر شویم و به سراغ فایل **yagoFacts.tsv** برویم، می‌توانیم خصوصیت‌های بیشتری برای جدول politician استخراج کنیم یا روابط جدیدی را کشف کنیم. مثلا محل تولد حسن روحانی شهر سرخه هست که اگر با استفاده از سرخه دوباره جستجو کنید، متوجه می‌شوید که قسمتی از استان سمنان و کشور ایران است. شاید تصمیم بگیرم محل تولد را تنها به عنوان یک خاصیت نگه دارم؛ اما حزب سیاسی وی را بیشتر کنکاش کنم و همکاران حزبی وی را پیدا کنم. با همین حقایق بدست‌آمده تا اینجا، می‌دانم که وی جزء رهبران کشور ایران است، از دانشگاه گلاسگو فارغ التحصیل شده و عضو حزب روحانیون مبارز (Combatant_Clergy_Association) است (و حقایقی دیگر). خود حزب می‌تواند به یک جدول خوبی تبدیل شود که در یک رابطه چند به چند با سیاسیون قرار بگیرد.

```
<id_j6oh03_10x_ozmjfg>    <Hassan_Rouhani>    <graduatedFrom>    <Glasgow_Caledonian_University>
<id_j6oh03_10x_knw1r3>    <Hassan_Rouhani>    <graduatedFrom>    <University_of_Tehran>
<id_j6oh03_1c7_zjx1yu>    <Hassan_Rouhani>    <isLeaderOf>    <Iran>
<id_j6oh03_10x_qs91q9>    <Hassan_Rouhani>    <graduatedFrom>    <University_of_Glasgow>
<id_j6oh03_1s2_1oudnjq>    <Hassan_Rouhani>    <isAffiliatedTo>    <Combatant_Clergy_Association>
<id_j6oh03_14h_zjx1yu>    <Hassan_Rouhani>    <isPoliticianOf>    <Iran>
<id_j6oh03_1s2_1edkg9z>    <Hassan_Rouhani>    <isAffiliatedTo>    <Islamic_Republican_Party>
<id_j6oh03_10x_18jvp3p>    <Hassan_Rouhani>    <graduatedFrom>    <Qom_Hawza>
<id_j6oh03_1c7_bqvclm>    <Hassan_Rouhani>    <isLeaderOf>    <Non-Aligned_Movement>
<id_j6oh03_p3m_zkjp59>    <Hassan_Rouhani>    <hasGender>    <male>
<id_j6oh03_oyl_tk3qk6>    <Hassan_Rouhani>    <wasBornIn>    <Sorkheh>
```

فرض کنید من قصد دارم اطلاعات مربوط به سیاسیون دیگر ایران را ببینم. فقط کافی هست که به دنبال چندتایی‌هایی در داخل فایل yagoFacts.tsv بگردم که **<isPoliticianOf> و <Iran>** دارند. مثلا به هاشمی رفسنجانی میرسم و با جستجو در مورد وی متوجه می‌شوم که کلی اطلاعات دیگر در مورد سیاسیون می‌توان بدست آورد، مثلا اینکه با چه کسی ازدواج کردند، فرزندان وی چی کسانی هستند و فرزندان سیاسی وی (به طور مثال رییس جمهور بعد از وی) چه کسانی هستند.

```
<id_hhhpjx_16x_3u9yei>    <Akbar_Hashemi_Rafsanjani>    <isMarriedTo>    <Effat_Marashi>
<id_3u9yei_16x_hhhpjx>    <Effat_Marashi>    <isMarriedTo>    <Akbar_Hashemi_Rafsanjani>
<id_hhhpjx_tje_zjx1yu>    <Akbar_Hashemi_Rafsanjani>    <isCitizenOf>    <Iran>
<id_hhhpjx_14h_zjx1yu>    <Akbar_Hashemi_Rafsanjani>    <isPoliticianOf>    <Iran>
<id_5mxh1z_11e_10709pj>    <Ali_Khamenei>    <hasChild>    <Hashemi_Rafsanjani>
<id_xwavld_10x_1vgp1mn>    <Mehdi_Hashemi_Rafsanjani>    <graduatedFrom>    <Sharif_University_of_Technology>
<id_cyt9jc_p3m_zkjp59>    <Mohammad_Hashemi_Rafsanjani>    <hasGender>    <male>
<id_xwavld_10x_knw1r3>    <Mehdi_Hashemi_Rafsanjani>    <graduatedFrom>    <University_of_Tehran>
<id_10709pj_11e_p2m5qk>    <Hashemi_Rafsanjani>    <hasChild>    <Mohammad_Khatami>
<id_hhhpjx_11e_o01lon>    <Akbar_Hashemi_Rafsanjani>    <hasChild>    <Faezeh_Hashemi>
<id_cyt9jc_10x_xvmift>    <Mohammad_Hashemi_Rafsanjani>    <graduatedFrom>    <University_of_California,_Berkeley>
```

```
<id_xwavld_oyl_5uo9mw>    <Mehdi_Hashemi_Rafsanjani>    <wasBornIn>    <Tehran>
<id_hhhpjx_11e_xwavld>    <Akbar_Hashemi_Rafsanjani>    <hasChild>    <Mehdi_Hashemi_Rafsanjani>
<id_hhhpjx_11e_aru4q9>    <Akbar_Hashemi_Rafsanjani>    <hasChild>    <Yasser_Hashemi_Rafsanjani>
<id_xwavld_12k_10it2g9>    <Mehdi_Hashemi_Rafsanjani>    <livesIn>    <England>
<id_1u17as0_oyl_1lfy8y>    <Mohsen_Hashemi_Rafsanjani>    <wasBornIn>    <Kerman>
<id_hhhpjx_1s2_z491o>    <Akbar_Hashemi_Rafsanjani>    <isAffiliatedTo>    <Executives_of_Construction_Party>
<id_10709pj_oyl_3429dq>    <Hashemi_Rafsanjani>    <wasBornIn>    <Bahreman>
<id_hhhpjx_1gi_1oyeh2y>    <Akbar_Hashemi_Rafsanjani>    <created>    <Islamic_Azad_University>
<id_aru4q9_p3m_zkjp59>    <Yasser_Hashemi_Rafsanjani>    <hasGender>    <male>
<id_10709pj_1s2_1oudnjq>    <Hashemi_Rafsanjani>    <isAffiliatedTo>    <Combatant_Clergy_Association>
<id_xwavld_p3m_zkjp59>    <Mehdi_Hashemi_Rafsanjani>    <hasGender>    <male>
<id_10709pj_14h_zjx1yu>    <Hashemi_Rafsanjani>    <isPoliticianOf>    <Iran>
<id_10709pj_10x_18jvp3p>    <Hashemi_Rafsanjani>    <graduatedFrom>    <Qom_Hawza>
<id_hhhpjx_11e_3nhrnu>    <Akbar_Hashemi_Rafsanjani>    <hasChild>    <Mohsen_Hashemi>
<id_xwavld_12k_3rux50>    <Mehdi_Hashemi_Rafsanjani>    <livesIn>    <Oxford>
<id_hhhpjx_ice_1pd7tss>    <Akbar_Hashemi_Rafsanjani>    <hasWebsite>    <http://www.rafsanjani.ir/>
<id_cyt9jc_oyl_1lfy8y>    <Mohammad_Hashemi_Rafsanjani>    <wasBornIn>    <Kerman>
<id_1v66fii_1sv_hhhpjx>    <Hussein-Ali_Montazeri>    <influences>    <Akbar_Hashemi_Rafsanjani>
<id_hhhpjx_p3m_zkjp59>    <Akbar_Hashemi_Rafsanjani>    <hasGender>    <male>
<id_hhhpjx_oyl_3429dq>    <Akbar_Hashemi_Rafsanjani>    <wasBornIn>    <Bahreman>
```

سپس همکاران سیاسی وی را در داخل حزب بدست بیارم:

```
<id_1h8zcle_1s2_z491o>    <Eshaq_Jahangiri>    <isAffiliatedTo>    <Executives_of_Construction_Party>
<id_4tqtuo_1s2_z491o>    <Hassan_Habibi>    <isAffiliatedTo>    <Executives_of_Construction_Party>
<id_1ik5xgi_1s2_z491o>    <Ata'ollah_Mohajerani>    <isAffiliatedTo>    <Executives_of_Construction_Party>
<id_1on48g6_1s2_z491o>    <Gholamhossein_Karbaschi>    <isAffiliatedTo>    <Executives_of_Construction_Party>
<id_64kjit_1c7_z491o>    <Hossein_Marashi>    <isLeaderOf>    <Executives_of_Construction_Party>
<id_eoi4gz_1s2_z491o>    <fa/اسماعیل_جبارزاده>    <isAffiliatedTo>    <Executives_of_Construction_Party>
<id_z491o_z7a_zjx1yu>    <Executives_of_Construction_Party>    <isLocatedIn>    <Iran>
<id_z491o_z7a_5uo9mw>    <Executives_of_Construction_Party>    <isLocatedIn>    <Tehran>
<id_1mf37i9_1s2_z491o>    <Bijan_Namdar_Zangeneh>    <isAffiliatedTo>    <Executives_of_Construction_Party>
<id_hhhpjx_1s2_z491o>    <Akbar_Hashemi_Rafsanjani>    <isAffiliatedTo>    <Executives_of_Construction_Party>
<id_pbaum5_1s2_z491o>    <Ebrahim_Asgharzadeh>    <isAffiliatedTo>    <Executives_of_Construction_Party>
<id_expfk0_1s2_z491o>    <Mohammad-Hassan_Malekmadani>    <isAffiliatedTo>    <Executives_of_Construction_Party>
<id_bvr125_1s2_z491o>    <Mohammad-Ali_Najafi>    <isAffiliatedTo>    <Executives_of_Construction_Party>
```

```
<id_64kjit_1s2_z491o>    <Hossein_Marashi>    <isAffiliatedTo>    <Executives_of_Construction_Party>
<id_gk5jnw_1s2_z491o>    <Morteza_Alviri>    <isAffiliatedTo>    <Executives_of_Construction_Party>
```

و الی آخر. از جهات مختلف استخراج این روابط قابل گسترش است. مثلا در فایل yagoTypes اگر به دنبال نوع Iran بگردید متوجه می‌شوید که علاوه بر یک کشور بودن، عضو اپک هم هست و از اینجا می‌توانید تمام کشورهای عضو اپک را پیدا کنید و پیرامون آنها و اتفاقات مشترک آنها بانک اطلاعاتی را گسترش دهید:

```
<id_zjx1yu_88c_1ihryd7>    <Iran>    rdf:type    <wikicat_Countries>
<id_zjx1yu_88c_132mplg>    <Iran>    rdf:type    <wikicat_Developing_8_Countries_member_states>
<id_zjx1yu_88c_a0gvjm>    <Iran>    rdf:type    <wikicat_Islamic_states>
<id_zjx1yu_88c_2vqcl2>    <Iran>    rdf:type    <wikicat_Member_states_of_OPEC>
<id_zjx1yu_88c_9jq2e4>    <Iran>    rdf:type    <wikicat_Member_states_of_the_Organisation_of_Islamic_Cooperation>
<id_zjx1yu_88c_n39lbp>    <Iran>    rdf:type    <wikicat_Member_states_of_the_United_Nations>
<id_zjx1yu_88c_hk7knc>    <Iran>    rdf:type    <wikicat_Middle_Eastern_countries>
<id_zjx1yu_88c_6toa57>    <Iran>    rdf:type    <wikicat_Near_Eastern_countries>
<id_zjx1yu_88c_abola4>    <Iran>    rdf:type    <wikicat_Persian-speaking_countries_and_territories>
<id_zjx1yu_88c_pzpudq>    <Iran>    rdf:type    <wikicat_South_Asian_countries>
<id_zjx1yu_88c_1g43g7n>    <Iran>    rdf:type    <wikicat_Western_Asian_countries>
<id_zjx1yu_88c_czeqp5>    <Iran>    rdf:type    <wordnet_country_108544813>
```

```
<id_vky0j_88c_2vqcl2>    <fr/Subdivisions_des_Émirats_arabes_unis>    rdf:type    <wikicat_Member_states_of_OPEC>
<id_o6ta41_88c_2vqcl2>    <Nigeria>    rdf:type    <wikicat_Member_states_of_OPEC>
<id_1ublz3_88c_2vqcl2>    <Kuwait>    rdf:type    <wikicat_Member_states_of_OPEC>
<id_zjx1yu_88c_2vqcl2>    <Iran>    rdf:type    <wikicat_Member_states_of_OPEC>
<id_zjx1yx_88c_2vqcl2>    <Iraq>    rdf:type    <wikicat_Member_states_of_OPEC>
<id_10q9jcd_88c_2vqcl2>    <Libya>    rdf:type    <wikicat_Member_states_of_OPEC>
<id_1bn7eu5_88c_2vqcl2>    <Saudi_Arabia>    rdf:type    <wikicat_Member_states_of_OPEC>
<id_1fc9apl_88c_2vqcl2>    <Venezuela>    rdf:type    <wikicat_Member_states_of_OPEC>
<id_3xvwg3_88c_2vqcl2>    <United_Arab_Emirates>    rdf:type    <wikicat_Member_states_of_OPEC>
<id_10svr79_88c_2vqcl2>    <Qatar>    rdf:type    <wikicat_Member_states_of_OPEC>
<id_1w0uyi2_88c_2vqcl2>    <Angola>    rdf:type    <wikicat_Member_states_of_OPEC>
<id_1bw4mbl_88c_2vqcl2>    <Algeria>    rdf:type    <wikicat_Member_states_of_OPEC>
<id_wdrm11_88c_2vqcl2>    <Energy_in_the_OPEC_countries>    rdf:type    <wikicat_Member_states_of_OPEC>
<id_vitdtp_88c_2vqcl2>    <Ecuador>    rdf:type    <wikicat_Member_states_of_OPEC>
```

امیدوارم با این مثال‌ها با نحوه گشتن در این چندتایی‌ها آشنا شده باشین. بعد از اینکه ساختار بانک اطلاعاتی استخراج شد، با نوشتن کدهای SQL یا هر زبان دیگری شروع به پرکردن جداول کنید. اگر ابهامی بود به تیم درس مراجعه کنید تا مسائل برای شما شفاف بشود.

حداقل‌ها

انتظاراتی که از شما در این پروژه می‌رود به شرح زیر است:

- استخراج حداقل ۱۰ تا رابطه که هر کدام حداقل ۱۰۰ سطر داشته باشند (شما می توانید روابط کوچک دیگری هم داشته باشید تا بانک اطلاعاتی کامل‌تری داشته باشیم) یا حداقل ۱۰۰۰ تا از رکوردهای یاگو ۳ پوشش داده شده باشند.
- تمام موجودیت‌های استخراج شده باید مرتبط با سیاست باشد و همه دارای حداقل یک برچسب فارسی باشند.
- بانک اطلاعاتی نهایی حداقل 3NF باشد.

خروجی

یک فایل زیپ سابمیت کنید که شامل موارد زیر باشد:

- مدل رابطه‌ای بانک اطلاعاتی که طراحی کردید به همراه توضیحات در مورد نحوه ساخت بانک اطلاعاتی (۳-۴ صفحه کفایت میکنه).
- فایل‌های SQL برای ساختن جداول بانک اطلاعاتی
- داده‌های موجود در جداول (که میشه به فرمت csv باشه)

قسمت امتیازی

کل پروژه ۱۰ درصد نمره کل هست اما میتواند تا ۲۰ درصد افزایش پیدا کند. برای این منظور یکی از دو کار زیر را انتخاب کرده و انجام دهید.

- تعداد رابطه‌ها را به ۲۰ رابطه افزایش دهید که هر کدام حداقل ۱۰۰ سطر داشته باشند یا حداقل ۲۰۰۰ تا از رکوردهای یاگو ۳ پوشش داده شده باشند.
- برنامه‌ای بنویسید که با استفاده از برچسب‌های زبان فارسی در بانک اطلاعاتی، هویت‌های موجود در متن اخبار را پیدا کرده، نوع آن را مشخص کرده و اطلاعات پیرامون آن هویت را به کاربر نشان دهد (نمونه اخبار در دراپ باکس درس قرار داده شده است، برای دیتاست کامل به من مراجعه کنید). برای منبع اخبار، همچنین به انتخاب خود، می توانید با استفاده از Google News API (پیوند)، اخبار فارسی، عربی و عبری درباره شخصیت‌های سیاسی ایران که در یاگو موجود است، بدست آورید و در پایگاه داده جانبی به نحوی مناسب ذخیره نمایید (خبر و هویت های موجود در خبر به همراه برچسب منطبق شده)