

DATA TIYDING

Soheil Shirvani

810195416



مقدمه

در این پروژه می خواهیم با استفاده از سایت ورزش 3 که یک سایت معتبر برای اخبار ورزشی در ایران است دادههاای را به واسطه صفحه های وب آن استخراج کنیم. برای این کار 5 فایل اکسل حاوی داده های صفحه وب به ما داده شده است تا داده ها را استخراج کنیم. این دادههای استخراج شده سپس به تیم یادگیری ماشین داده می شود تا با پیدا کردن مدل های مناسب خروجی های مورد نظر را پیدا کنند.

داده های استخراج شده

√ جام حذفي:

داده های این جدول مانند زیر است:

	Year	Stage	team1	team2	score1	score2	time	date	Penalty1	Penalty2
0	جام حذفي ايران 98-99	جام حذفي الران 1/16 نهالِي	يرسيوليس	ماشينسازىتبريز	1	0	15:00	1398/07/08	0	0
1	جام حذفي ايران 98-99	جام حذفي الران 1/16 نهالِي	صنعت نفت أبادان	قشقایی شیراز	2	1	15:00	1398/07/07	0	0
2	جام حذفي ايران 98-99	جام حذفي الران 1/16 نهالِي	مس نوین کرمان	استقلال ماهشهر	1	2	16:00	1398/07/08	0	0
3	جام حذفي ايران 98-99	جام حذفي الران 1/16 نهالِي	شهرداری ماهشهر	سردار بوكان	2	1	15:00	1398/07/09	0	0
4	جام حذفي ايران 98-99	جام حذفي الران 1/16 نهالِي	سياهان	خيبر خرمآباد	3	0	16:30	1398/07/08	0	0
396	جام حذفي ابران 86-87	جام حذفی ایران یک چهارم نهایی	سياهان	داماش گیلان	2	4	14:00	1387/03/07	0	0
397	جام حذفي ابران 86-87	جام حذفی ایران نیمه نهایی	استقلال	فولاد	1	1	14:00	1387/03/12	2	4
398	جام حذفي ايران 86-87	جام حذفی ایران نیمه نهایی	برق شیراز	داماش گیلان	1	2	14:00	1387/03/09	0	0
399	جام حذفي ايران 86-87	جام حذفی ایران فیدل	استقلال	داماش گیلان	0	1	14:00	1387/03/20	0	0
400	جام حذفي ابران 86-87	جام حذفی ایران فیدل	داماش گیلان	استقلال	0	3	14:00	1387/03/27	0	0

401 rows × 10 columns

این داده ها شامل:

Year: سال مسابقه، Stage: مرحله مسابقه در جام حذفي، team1: تيم ميزبان،

team2: تیم مهمان، score1: امتیاز تیم میزبان، score2: امتیاز تیم مهمان، score1: زمان شروع مسابقه، date: تاریخ شروع مسابقه، penalty1: امتیاز گرفته شده در پنالتی برای تیم میزبان، penalty2: امتیاز گرفته شده در پنالتی نرود عدد penalty2: امتیاز گرفته شده در پنالتی برای تیم مهمان در صورتی که بازی به پنالتی نرود عدد آن 0 است.

این داده ها شامل 401 ردیف که نشان دهنده تعدادی از مسابقات جام حذفی در سالهای گذشته است.

√ تيم ها:

این داده ها شامل:

	Competition	تيم.1	بازيها	امتياز
0	آرشیو جدول های لیگ برتر ایران	گل گهرسپرجان	2	6
1	آرشیو جدول های لیگ برتر ایران	نساجى مازندران	2	4
2	آرشیو جدول های لیگ برتر ایران	فولاد	2	4
3	آرشیو جدول های لیگ برتر ایران	يرسيوليس	2	4
4	آرشیو جدول های لیگ برتر ایران	شهر خودرو	1	3
236	أرشيو جدول هاى ارديويسه هلند	هيرنفين	8	16
237	أرشيو جدول هاى ارديويسه هلند	آلكمار	7	11
238	أرشيو جدول هاى ارديويسه هلند	اوترخت	8	10
239	أرشيو جدول هاى ارديويسه هلند	دايمخن	34	34
240	جام جهائی 2018 روسیِه	دانمارک	6	10

241 rows × 4 columns

که در آن

Competition: نشان دهنده مسابقه انجام شده، تیم: نشان دهنده اسم تیمی که در مسابقه شرکت کرده، بازیها: نشاندهنده ی تعداد بازی های انجام شده توسط آن تیم در آن مسابقه امتیاز: نشان دهنده امتیاز گرفته شده توسط آن تیم در آن مسابقه است.

این داده ها شامل 241 ردیف که نشان دهنده 240 تیم است که در مسابقات مختلف شرکت کردهاند و امتیاز و تعداد بازیهای آن ها در همان مسابقه را نشان می دهد. تیم های خارجی و داخلی ایران هر دو در این جدول حضور دارند.

√ جدول ها:

این داده ها شامل:

	Year	Competition	تيم	بازيها	يرد	مساوى	يلقت	گل زده	گل خورده	تقاضل كُل	امتياز
0	لْدِگ برئر (99-00)	آرشیو جدول های لیگ برتر ایران	گل گهرسپرجان	2	2	0	0	5	1	4	6
1	لْدِگ برىز (99-00)	آرشیو جدول های لیگ برتر ایران	تساجى مازندران	2	1	1	0	4	1	3	4
2	لْدِگ برىز (99-00)	أرشيو جدول های ليگ برتر ايران	فولاد	2	1	1	0	4	3	1	4
3	لْدِگ برىز (99-00)	أرشيو جدول های ليگ برتر ايران	يرسيوليس	2	1	1	0	1	0	1	4
4	لْدِگ برى (99-00)	أرشيو جدول های ليگ برتر ايران	شهر خودرو	1	1	0	0	3	0	3	3
1956	C گروه	جام جهائی 2018 روسیِه	استراليا	3	0	1	2	2	5	-3	1
1957	B گروه	جام جهائی 2018 روسیِه	اسيانيا	3	1	2	0	6	5	1	5
1958	B گروه	جام جهائی 2018 روسیِه	برتغال	3	1	2	0	5	4	1	5
1959	B گروه	جام جهائی 2018 روسیِه	ايران	3	1	1	1	2	2	0	4
1960	B گروه	جام جهائی 2018 روسیِه	مراكش	3	0	1	2	2	4	-2	1

1961 rows × 11 columns

که نشان دهنده بازی های فوتبال و فوتسال است. این جدول شامل:

Year: سال انجام بازی، Competition: نام مسابقه ی بازی، تیم: اسم تیمی که در ان مسابقه شرکت کرده، بازیها: تعداد بازی ای که آن تیم در آن مسابقه انجام داده است، برد: تعداد برد های آن تیم در آن مسابقه، باخت: تعداد باخت های آن تیم در آن مسابقه، باخت: تعداد باخت های آن تیم در آن مسابقه، گل زده و گل خورده: تعداد گل های زده و خورده ی آن تیم در آن لیگ، امتیاز: امتیاز گرفته شده توسط آن تیم در آن لیگ.

این داده ها شامل 1960 تیم است که در لیگ های مختلف فوتبال و فوتسال شرکت کرده اند و امتیازات این تیم ها نوشته شده است.

	Year	Competition	تيم	يازى	يرد	بلقت	گ ل زده	گل خورده	تقاضل كُل	امتياز
0	1398	أرشيو جدول های ليگ برتر بسکتبال ايران	شهرداری گرگان	26	24	2	1850	1413	437	50
1	1398	أرشيو جدول های ليگ برتر بسکتبال ايران	يتروشيمي ب.ا	26	21	5	1929	1514	415	47
2	1398	أرشيو جدول های ليگ برتر بسکتبال ايران	شيميين	26	20	6	2108	1831	277	46
3	1398	أرشيو جدول های ليگ برتر بسکتبال ايران	بِالابِش نفت أبادان	26	20	6	1948	1718	230	46
4	1398	أرشيو جدول های لميگ برتر بسکتبال ايران	مهرام	26	19	7	1943	1697	246	45
89	92-93	أرشيو جدول های ليگ برتر بسکتبال ايران	همپاری زنجان	22	9	13	1410	1503	-93	27
90	92-93	أرشيو جدول های ليگ برتر بسکتبال ايران	استقلال كثمم	22	8	14	1463	1556	-93	24
91	92-93	أرشيو جدول های ليگ برتر بسکتبال ايران	افرای تهران	22	6	16	1449	1624	-175	18
92	92-93	أرشيو جدول های ليگ برتر بسکتبال ايران	شهرداری گرگان	22	6	16	1346	1600	-254	18
93	92-93	أرشيو جدول های ليگ برتر بسکتبال ايران	ليروى زمينى	22	3	19	1313	1738	-425	9

94 rows × 10 columns

که شامل بازی های تیم ها در لیگ های بسکتبال است. این جدول شامل:

Year: سالی که مسابقه برگزار شده است، Competition: نام لیگ ای که مسابقه برگزار شده است، تیم: نام تیمی که در آن مسابقه شرکت کرده است، بازی، برد و باخت: شامل تعداد کل بازی های آن تیم، تعداد برد ها و تعداد باخت های آن تیم که در آن مسابقه بوده است، گل خورده و گل زده: نشان دهنده تعداد گل زده و خورده ی آن تیم در آن لیگ، تفاضل گل: نشان دهنده تفاضل گل خورده و زده ی آن تیم در آن لیگ، امتیاز: نشان دهنده ی امتیاز گرفته شده توسط آن تیم در آن مسابقه.

این جدول شامل 94 تیم است که در لیگ های بسکتبال شرکت کرده اند و امتیاز های این تیم ها در این جدول جمع آوری شده است.

Data Tidying Report

1	Year	Competition	تيم	امتياز	يازي	يرد	ياخت	30	31	32	 باخته	باخته	معدل	معدل	يرده	يرده	باخته	ياخته	معدل	معدل
0	(A سطح) 2017	ليگ جهاني واليبال	فرانسه	25	9	8	1	4	4	NaN	 7	673	3.714	1.177	26	792	7	673	3.714	1.177
1	(A سطح) 2017	ليگ جهاني واليبال	برزيل	19	9	6	3	1	5	NaN	 14	786	1.571	1.039	22	817	14	786	1.571	1.039
2	(A سطح) 2017	ليگ جهاني واليبال	صريستان	18	9	6	3	2	3	1	 14	767	1.500	1.039	21	797	14	767	1.500	1.039
3	(A سطح) 2017	ليگ جهاني واليبال	روسيه	14	9	5	4	3	NaN	2	 16	775	1.188	1.015	19	787	16	775	1.188	1.015
4	(A سطح) 2017	ليگ جهاني واليبال	كائادا	12	9	5	4	NaN	2	3	 20	858	0.900	0.984	18	844	20	858	0.900	0.984
182	(مرحله گروهي) 2015	جام جهاني واليبال	ايران	12	11	4	7	2	1	1	 24	893	0.667	1.013	16	905	24	893	0.667	1.013
183	(مرحله گروهي) 2015	جام جهاني والبيال	استزاليا	12	11	4	7	2	1	1	 24	925	0.625	0.898	15	831	24	925	0.625	0.898
184	(مرحله گروهي) 2015	جام جهاني واليبال	مصر	8	11	2	9	NaN	1	1	 30	976	0.433	0.911	13	889	30	976	0.433	0.911
185	(مرحله گروهي) 2015	جام جهاني واليبال	والزوئلا	3	11	1	10	NaN	NaN	1	 32	981	0.250	0.831	8	815	32	981	0.250	0.831
186	(مرحله گروهي) 2015	جام جهاني واليبال	ئونس	2	11	0	11	NaN	NaN	NaN	 33	924	0.152	0.808	5	747	33	924	0.152	0.808

187 rows × 25 columns

√ نقل و انتقالات

جدول شامل زیر است:

	نقل و انتقال	سال	تيم	ورودی ، قطعی	ورودی ، قرضی (بازگشت)	ورودی ، قرضی	ورودی ، احتمالی	خروجي ، قطعي	خروجی ، قرضی (بازگشت)	خروجی ، قرضی	غروجی ، احتمالی
0	نقل و انتقالات لیگ برکز	99-00	استقلال	بابک مرادی, سید احمد] موسوی, مثین کریم زاده, م	[سجاد آقابِی]	0	0	شاہین طاہرخانی, سیدسین] حسینی, علی نشئی, محس	0	[حسين يورحميدي, محمد بلبلي]	سید حسین] [حسینی
1	نقل و انتقالات لیگ براتر	99-00	آلومينيوم اراک	اسماعیِل شریفات، پوریا آریاکیِا،] جابر الصاری, ش	0	[حسين يورحميدي]	0	[محمود قائد رحمتي]	امیرحسین] [جدی	0	0
2	نقل و انتقالات لیگ براثر	99-00	پارس جنوبي جم	0	0	0	0	یوریا آریاکیا, ملین کریم زاده,] محمد نوری, فری	0	0	0
3	نقل و انتقالات لیگ براثر	99-00	يرسيوليس	أرمان رمضائي, احسان يهلوان,] حامد لک, سعيد أقا	0	0	فرشاد] [فرجی	علیرضا بیرانوند, محمد نادری,] ,شجاع خلیل زاده	0	[امیر روستایی]	0
4	نقل و انتقالات لیگ بریر	99-00	پیکان	اسماعیِل بابایی, حسین حیدری,] رضنا جبیره, شاهیِن	امیر حسین] [جدی	[امیر روستایی]	0	جابر انصباري, سعيد واسعي,] جلال الدين على محمدي	0	0	0
315	نقل و انتقالات اروپا	2015- 2016	نايولى	[په په ريدا, ولاد چريچس]	0	0	0	0	0	0	0
316	نقل و انتقالات اروپا	2015- 2016	والنسيا	[ايمن عبدنور]	0	0	[]	[نیکولاس او تامندی]	0	0	0
317	نقل و انتقالات اروپا	2015- 2016	وولفسبورگ	دانیلو دیئریج, روبرت اسچیک,] ژولین دراکسلر, سب	0	0	[]	کوین دی بروین, ایوان پریشیچ,] کارل کلاوس, دان	0	0	0
318	نقل و انتقالات اروپا	2015- 2016	وباردال	[دانیله بونر۱]	0	0		0	0	0	0
319	نقل و انتقالات اروپا	2015- 2016	يوولتوس	نتو, هرنانس, ألبرتو جرى,] ألكس ساندرو, اريك	[لوكا مارون]	گویدو وادالا, گویدو وادالا,] …ماریو لمینا, خوان		آرئورو ويدال, فردريک] سورنسن, ادواردو سريا, يا	0	کیدگزلی کومان, مارسل بوچل,] والریو لورنزو روست	0

320 rows × 11 columns

که در آن:

نقل و انتقال: نام لیگ ای که نقل و انتقال انجام شده است، سال: سالی که این نقل و انتقال در آن لیگ انجام شده است، تیم: نام تیمی که این نقل و انتقال را انجام داده است، ورودی و قطعی: بازیکنان ورودی به تیم که قطعی شده اند، ورودی قرضی بازگشت: بازیکنان ای که به تیم های دیگر قرض داده شده بودند و بازگشتند، ورودی قرضی: بازیکنان ای که از تیم های دیگر قرض گرفته شده اند و به تیم ورود کردند، ورودی احتمالی: بازیکنان ای که احتمالا قرار است در این فصل به این تیم ورود پیدا کنند، خروجی ها: همان حالت های ورودی را دارد با این تفاوت که این بازیکنان از تیک خارج شده اند.

این جدول شامل 320 ردیف شامل نقل و انتقالات تیم های مختلف در سال ها و لیگ های مختلف را نشان می دهد.

√ جام جهانی:

جدول آن به شکل زیر است:

	Year	Stage	team1	team2	score1	score2	time	date	Penalty1	Penalty2
0	جام جهانی 2014	جام جهانی یک هشتم نهایی	ثيثى	ىرزىڭ	1	1	20:30	1393/04/07	3	2
1	جام جهانی 2014	جام جهانی یک اشتم نهایی	اروگونه	كأمبيا	0	2	00:30	1393/04/08	0	0
2	جام جهانی 2014	جام جهانی یک هشتم نهایی	نبجريه	فرانسه	0	2	20:30	1393/04/09	0	0
3	جام جهانی 2014	جام جهانی یک هشتم نهایی	الجزاير	أثمان	1	2	00:30	1393/04/10	0	0
4	جام جهانی 2014	جام جهانی یک هشتم نهایی	مكزيك	هلتد	1	2	20:30	1393/04/08	0	0
5	جام جهانی 2014	جام جهانی یک هشتم نهایی	يونان	كاستاريكا	1	1	00:30	1393/04/09	5	3
6	جام جهانی 2014	جام جهانی یک هشتم نهایی	سوئيس	أرزانتين	0	1	20:30	1393/04/10	0	0
7	جام جهانی 2014	جام جهانی یک هشتم نهایی	أمريكا	بلڑیک	1	2	00:30	1393/04/11	0	0
8	جام جهانی 2014	جام جهانی یک چهارم نهایی	كأمييا	ىرزىك	1	2	00:30	1393/04/14	0	0
9	جام جهانی 2014	جام جهانی یک چهارم نهایی	آئمان	فرانسه	1	0	20:30	1393/04/13	0	0
10	جام جهانی 2014	جام جهانی یک چهارم نهایی	كاستاريكا	هلتد	0	0	00:30	1393/04/15	4	3
11	جام جهانی 2014	جام جهانی یک چهارم نهایی	بلڑیک	أرزانتين	0	1	20:30	1393/04/14	0	0
12	جام جهانی 2014	حام حياتى نبعه نيابي	آئمان	ىرزىك	7	1	00:30	1393/04/18	0	0
13	جام جهانی 2014	حام جهائى نبعه نهائي	أرژانتين	هلتد	0	0	00:30	1393/04/19	2	4
14	جام جهانی 2014	حام جهائى فينال	آرژانتین	آثمان	0	1	23:30	1393/04/22	0	0
15	جام جهاني 2018	جام جهانی یک هشتم نهایی	يرتغال	اروگونه	1	2	22:30	1397/04/09	0	0
16	جام جهاني 2018	جام جهانی یک هشتم نهایی	آرژانتین	فرانسه	3	4	18:30	1397/04/09	0	0
17	جام جهانی 2018	جام جهانی یک هشتم نهایی	مکزیک	برزيك	0	2	18:30	1397/04/11	0	0
18	جام جهاني 2018	جام جهانی یک هشتم نهایی	ڈان	بلڑیک	2	3	22:30	1397/04/11	0	0
19	جام جهاني 2018	جام جهانی یک هشتم نهایی	روسيه	اسهائها	1	1	18:30	1397/04/10	3	4
20	جام جهانی 2018	جام جهانی یک هشتم نهایی	دانمارک	كزواسى	1	1	22:30	1397/04/10	3	2
21	جام جهانی 2018	جام جهانی یک هشتم نهایی	سوئيس	سواد	0	1	18:30	1397/04/12	0	0
22	جام جهالی 2018	جام جهانی یک هشتم نهایی	انگلیس	كثمييا	1	1	22:30	1397/04/12	3	4
23	جام جهانی 2018	جام جهانی یک چهارم نهایی	فرانسه	اروگونه	2	0	18:30	1397/04/15	0	0
24	جام جهانی 2018	جام جهانی یک چهارم نهایی	بلڑیک	برزيل	2	1	22:30	1397/04/15	0	0
25	جام جهانی 2018	جام جهانی یک چهارم نهایی	كزواسى	روسيه	2	2	22:30	1397/04/16	3	4
26	جام جهالی 2018	جام جهانی یک چهارم نهایی	انگلیس	سوئد	2	0	18:30	1397/04/16	0	0
27	جام جهالی 2018	جام جهانی نبهه نهایی	بلڑیک	فرانسه	0	1	22:30	1397/04/19	0	0
28	جام جهالی 2018	جام جهانی نبهه نهایی	انگلیس	كزواسى	1	2	22:30	1397/04/20	0	0
29	جام جهانی 2018	جام جهانی ردهندی	انگلیس	بلڑیک	0	2	18:30	1397/04/23	0	0
30	جام جهانی 2018	جام جهائى فينال	كازواسى	فرانسه	2	4	19:30	1397/04/24	0	0

این جدول شامل ستون هایی با همان تعریف های ستون های جام حذفی است با این تفاوت که جام آن یک جام جهانی در سال های 2014 و 2018 است. بقیه ستون ها همانند جام حذفی است.

1. برای پیشبینی قهرمان یک لیگ، از داده هایی که تیم قبلی به ما داده است چه اطلاعاتی بدست آوردیم؟ چه ویژگی هایی را استخراج کردیم و چه داده هایی را کنار گذاشتیم؟

داده های استخراج شده از صفحات مختلف در بالا توضیح داده شد. لازم به ذکر است که داده هایی مانند فیلم و عکس های بازیکنان، بازی های گذشته برخی تیم ها که در جدول جام حذفی گزارش شده بود، خبر های ورزشی که در رابطه با تیم و لیگ ها در صفحات مختلف گذاشته شده بود استخراج نشده اند و در صورت نیاز می توان آنها را نیز استخراج کرد ولی ممون است تاثیر چندانی در پیشبینی فینال ندارند.

2. هزينه استخراج ويژگي ها چقدر بوده است؟

در قسمت قبل در رابطه با داده های استخراج شده و داده هایی که تصمیم گرفته شد استخراج نشوند صحبت شد. هزینه این استخراج برای داده های بالا هزینه زمانی بوده است و زمان خوبی برای استخراج گرفته اند. برای استخراج داده های بیشتر از جمله داده های خبر ها و یا فیلم ها نیاز به حافظه بیشتر و در نتیجه دیتابیس بزگر تر نیز داریم که مثلا برای داده های فیلم و دانلود کردن آنها ممکن است هزینه بیشتری نیاز داشته باشیم. برای استخراج جزیات بیشتر هزینه زمانی زیادی نیاز داشتیم چرا که برای اینکار باید به صفحان مربوط به هر کدام از صفحات داده شده نیز مراجعه می کردیم و جداول مرتبط را استخراج می کردیم. در صورت بزرگ تر شدن جداول نیاز داریم تا آنها را وارد دیتابیس کنیم که کوری نوشتن آن نیز یک هزینه زمانی دیگر تحمیل می کند.

3. چه توجیهی برای این هزینه دارید؟ در آینده چه کمکی می تواند به ما بکند؟

این داده ها علارقم داشتن هزینه زمانی می توانند اطلاعات زیادی برای ما داشته باشند. مثلا در صورت داشتن اخبار مربوط به یک تیم می توانیم نقل و انتقالات تیم، معروف بودن بازیکن های مختلف، میزان رضایت مردم از تیم و ... را تشخیص داد. این اطلاعات می تواند هزینه ناشی از استخراج اطلاعات را جبران کند. هزینه داشتن دیتابیس های خوب نیز با داده API های مناسب به شرکت ها و موسسه ها می تواند هزینه را جبران کند. این داده ها چون هزینه بیشتری می طلب اند نیاز داریم تا با تیم یادگیری مدل نیز ارتباط داشته باشیم تا در صورت نیاز آنها را استخراج کنیم.

4. چگونه می سنجید که داده های استخراج شده قابل اعتماد است؟ فرایند کیفیت سنجی را توضیح دهید؟

می توانیم داده ها را با سایت های معتبر دیگر چک کنیم که آیا نتایج بدست آماده با واقعیت بقیه سایت ها هم خوانی دارد یا خیر. می توانیم از یک متخصص در حوزه ی داده ها بخواهیم تا داده ها را چک کند و بگوید آیا این داده ها می توانند درست باشند یا خیر. می توانیم داده های پرت مربوط به هر جدول را استخراج نمودار های پراکندگی و هیستوگرام مربوط به آن ها را کشیده و از روی این داده های آماری ببینیم که آیا داده های بدست آماده قابلیت استفاده در واقعیت و صنعت را دارند یا خیر. نیاز داریم تا این داده ها را تمیز کنیم و با واقعیت چک کنیم سپس بعد از دادن داده ها به تیم یادگیری مدل مطمئن شویم به دقت بالایی رسیده اند در غیر این صورت داده های بیشتر و تمیز تری را استخراج کنیم.

5. چگونه می توان این پروژه را در ابعاد بزرگ تر و در یک تیم انجام داد؟ چه وظایفی و چه فرایند کاری لازم است؟

برای این کار نیاز داریم تا مثلا به هر کس تعدادی از قسمت های داده ها را بدهیم تا با دقت و جزییات بیشتری این کار انجام دهد. باید یک رابط یا مدیر بین دو بخش استخراج داده و یادگیری مدل باشد تا به گروه بگوید که چه داده های برای این گروه واجب است را استخراج کنند. نیاز به یک مدیر برای تامین هزینه ها و افراد داریم تا بتوانیم افراد مناسب را با کمترین هزینه پیدا کنیم. تسک ها باید طوری بین افراد تقسیم شود که همپوشانی نداشته باشند و بتوانند با توجه به توانایی های خود داده های خوب را استخراج کنند. تیم استخراج داده باید با تیم یادگیری مدل تعامل داشته باشد تا بتواند داده های مناسب را استخراج کند. در این راه می توانیم یک گروه نیز برای تایید داده ها داشته باشیم تا بتوانیم با صراحت بگوییم داده قابل اعتماد هستند. برای این کار می توانیم متخصص مربوط به کار را در گروه داشته باشیم و از سایت ها و صفحه های دیگر این داده ها را چک کنیم تا مغایرت با واقعیت نداشته باشند.

6. نتایج حاصل از فعالیت تیم را چگونه به تیم بعد ارائه می دهید؟ توضیح دهید؟

ابتدا جداول و دیتابیس های استخراج شده را با یک گزارش از هر کدام با توضیح ویژگی های استخراج شده توضیح می دهیم. این توضیحات به زبان ساده و عام هست و نیاز به تخصص ندارند و قابل فهم هستند. سپس با توضیح این که داده ها قابل اعتماد هستند و چه ویژگی های دیگری نیز قابل استخراج هستند که به دلیل هزینه آن فقط در صورت نیاز استخراج می شوند ادامه می دهیم. سپس در رابطه با تمیز کردن داده ها و این که چه پروسه ای برای این کار انجام دادیم صحبت می کنیم تا در صورت نیاز بتوانند با روش های دیگر این کار را انجام دهند. و در آخر نمودار هایی مربوط به ویژگی های آماری داده نشان می دهیم و داده ها و پراکندگی آن ها که مثلا چه تعداد داده ی پرت موجود است گزارش می کنیم.

جدول های گفته شده به پیوست آمده اند و در صورت قبول هزینه بیشتر می توان داده های بیشتری مخصوصا داده های متنی بهتری نیز از این صفحات بدست اورد.