Parse patent assignments in 2015 of the XML format to the CSV format

1. Drop patent assignments
   (a) without patent id
   (b) conveyance text is not “Assignment of Assignors Interest”
2. Each row records the patent assignment for each patent ID. It includes the information, including assignor, assignees, assignees addresses, last update date, reel number, and frame number.
3. Drop the duplicated patent assignments for each patent id based on assignor and assignee names
   Some patent assignments have duplicated entries in the record.
   We should only keep the most recent patent assignment to decrease the running time.
   EX:

address-1_0	assignor	city_0	country-name_0	frame-no	last_update_date	name_0	patent_ids	postcode_0	reel-no	total_number_assignees
7-7-1, NAKAGAITO, DAITO-SHI	LEXMARK INTERNATIONAL, INC.	OSAKA	JAPAN	1	20150203	FUNAI ELECTRIC CO., LTD	0423569	5740013	30416	1
7-7-1, NAKAGAITO, DAITO-SHI	LEXMARK INTERNATIONAL, INC.	OSAKA	JAPAN	1	20150217	FUNAI ELECTRIC CO., LTD	0423569	5740013	30416	1
7-7-1, NAKAGAITO, DAITO-SHI	LEXMARK INTERNATIONAL, INC.	OSAKA	JAPAN	1	20150303	FUNAI ELECTRIC CO., LTD	0423569	5740013	30416	1

Identify not meaningful patent transfer

1. Sort patent assignments by their patent ids and last updated date:
   The data does not record the address for the assignor. Doing so allows us to find the address of the assignor when it is the assignee.
   EX:

```
temp = total[total['patent_ids'] == patent_ids[-10]]
temp = temp.dropna(how = 'all', axis = 1)
temp = temp.drop_duplicates( ['assignor', 'name_0'], keep = 'first')
temp
```

]:

	address-1_0	assignor	city_0	frame-no	last_update_date	name_0	patent_ids	postcode_0	reel-no	state_0	total_number_assignees
0	5 HIGH RIDGE PARK	WALKER, JAY S.	STAMFORD	840	20150303	WALKER DIGITAL CORPORATION	RE45401	06905	34423	CONNECTICUT	1
0	1 HIGH RIDGE PARK	WALKER DIGITAL CORPORATION	STAMFORD	869	20150303	WALKER DIGITAL, LLC	RE45401	06905	34423	CONNECTICUT	1
0	2 HIGH RIDGE PARK	WALKER DIGITAL, LLC	STAMFORD	929	20150303	INVENTOR HOLDINGS, LLC	RE45401	06905	34423	CONNECTICUT	1

2. Standardize company names before comparing their similarities:
   (1) Remove symbols including ‘.’, ‘,’, ‘*’
   (2) Drop end words such as inc, company and corporation
3. Identify whether the patent transfer is meaningful
   (1) If we could not find the address of the assignor:
   a. Compare the similarity between the assignor name and the assignee name
   We could get the score measuring the similarity by FuzzyMatching
   If the score is > 80 (80% similairity), we think that the names are similar. Thus, the patent transfer is not meaningful.

   b. In many cases, the main company word in the name does not change.
   If the word in the assignor name and the assignee name overlaps, we consider the patent transfer

as not meaningful.

EX: Assignor name – Fuji Electric Device Technology

Assignee name: Fuji Electric Systems

	address-1_0	assignor	city_0	country-name_0	frame-no	last_update_date	name_0	patent_ids	postcode_0	reel-no	total_number_assignees
240	11-2 OSAKI 1-CHOME, SHINAGAWA-KU	FUJI ELECTRIC DEVICE TECHNOLOGY CO., LTD.	TOKYO	JAPAN	438	20150310	FUJI ELECTRIC SYSTEMS CO., LTD.	0587662	141-0032	24252	1

(2) If we could find the address of the assignor:

Compare the similarity of each component of the address between the assignor and the assignee

If the similarity of each component of the address is larger than 85%, the patent transfer is not meaningful.

EX:

```
temp = total[total['patent_ids'] == patent_ids[-10]]
temp = temp.dropna(how = 'all', axis = 1)
temp = temp.drop_duplicates( ['assignor', 'name_0'], keep = 'first')
temp
```

	address-1_0	assignor	city_0	frame-no	last_update_date	name_0	patent_ids	postcode_0	reel-no	state_0	total_number_assignees
0	5 HIGH RIDGE PARK	WALKER, JAY S.	STAMFORD	840	20150303	WALKER DIGITAL CORPORATION	RE45401	06905	34423	CONNECTICUT	1
0	1 HIGH RIDGE PARK	WALKER DIGITAL CORPORATION	STAMFORD	869	20150303	WALKER DIGITAL, LLC	RE45401	06905	34423	CONNECTICUT	1
0	2 HIGH RIDGE PARK	WALKER DIGITAL, LLC	STAMFORD	929	20150303	INVENTOR HOLDINGS, LLC	RE45401	06905	34423	CONNECTICUT	1

(3) If the first patent transfer is from a person to its company

(a) Sort patents by their last update dates

(b) Only keep patents with the earliest date (The first transfer)

(c) Split the assignor name to individual words

(d) Use individual words to identify the pattern “Last name First Name Middle Name Abbreviation”. If this is case, the assignor is the person.

(e) Use the First Names and Last Names dataset to identify whether each individual word in the assignor name is people name. If they are all people names, the assignor is the person.

Reference: https://github.com/philipperemy/name-dataset

Sample Output:

	assignor	name_0	notmeaningful	notmeanigful_firsttransfer
1	robert bosch	ip	True	True
23	fabbri jason	adobe systems	True	True
99	hays bill j	hays barbara b	True	True
109	hays bill j	hays barbara b	True	True
110	traeger joseph p	traeger pellet grills	True	True
432	traeger joseph p	traeger pellet grills	True	True
443	hays bill j	hays barbara b	True	True
939	hays bill j	hays barbara b	True	True
1003	michelson gary karlin	sdgi	True	True
1097	michelson gary karlin	sdgi	True	True
1154	michelson gary karlin	sdgi	True	True
1280	michelson gary karlin	sdgi	True	True