

BMS Project 1

Mid-term Review

Amrutha K
Rohit Vernekar

Data Exploration

Standard Names

	Analysis	Attribute	Standard names
0	CE-SDS (NON-REDUCED)	HHL	CE-SDS (NON-REDUCED) HHL
1	CE-SDS (NON-REDUCED)	PURITY	CE-SDS (NON-REDUCED) PURITY
2	CE-SDS (REDUCED)	PURITY	CE-SDS (REDUCED) PURITY
3	SDS-PAGE (NON-REDUCED)	PURITY	SDS-PAGE (NON-REDUCED) PURITY
4	SDS-PAGE (REDUCED)	PURITY	SDS-PAGE (REDUCED) PURITY
5	RP-HPLC	PURITY (MAIN PEAK)	RP-HPLC PURITY
6	IEF	ACIDIC PEAKS	IEF ACIDIC PEAKS
7	IEF	BASIC PEAKS	IEF BASIC PEAKS
8	IEF	MAIN PEAK	IEF MAIN PEAK
9	ICIEF	ACIDIC PEAKS	ICIEF ACIDIC PEAKS
10	ICIEF	BASIC PEAKS	ICIEF BASIC PEAKS
11	ICIEF	MAIN PEAK	ICIEF MAIN PEAK
12	CEX	ACIDIC PEAKS	CEX ACIDIC PEAKS
13	CEX	BASIC PEAKS	CEX BASIC PEAKS
14	CEX	MAIN PEAK	CEX MAIN PEAK
15	AEX	ACIDIC PEAKS	AEX ACIDIC PEAKS
16	AEX	BASIC PEAKS	AEX BASIC PEAKS
17	AEX	MAIN PEAK	AEX MAIN PEAK
18	SE-HPLC	HMW	SE-HPLC HMW
19	SE-HPLC	LMW	SE-HPLC LMW
20	SE-HPLC	MONOMER	SE-HPLC MONOMER
21	SE-UPLC	HMW	SE-UPLC HMW
22	SE-UPLC	LMW	SE-UPLC LMW
23	SE-UPLC	MONOMER	SE-UPLC MONOMER
24	HIAC OR UV	PARTICULATE-MATTER >= 10-UM	PARTICULATE-MATTER >= 10-UM
25	HIAC OR UV	PARTICULATE-MATTER >= 25-UM	PARTICULATE-MATTER >= 25-UM
26	CELL-BASED BIOASSAY	POTENCY	POTENCY BY CELL-BASED BIOASSAY
27	ELISA	BINDING ACTIVITY OR POTENCY	POTENCY BY BINDING ELISA
28	SPR/BIOCORE	BINDING ACTIVITY	SPR BINDING ACTIVITY
29	PH	PH	PH
30	A280	PROTEIN CONCENTRATION	PROTEIN CONCENTRATION (A280)
31	POLYSORBATE 80	POLYSORBATE 80	POLYSORBATE 80

Data(7730,2)

	analysis	Attribute_name
0	D_250475	IL2 INHIBITION ASSAY
1	D_95007196	PH
2	D_M00003744	ABATACEPT MAJOR BAND (REDUCED)
3	Y_SM_95011468_R	BIOASSAY
4	250684_CE_SDS_REDUC	SUM HEAVY AND LIGHT CHAIN
5	250580_TOTAL_PROT	PAAD
6	250580_TOTAL_PROT	SAMPLE 1 MASS
7	250684_CE_SDS_REDUC	SS RM 3 PURITY HC AND LC PEAKS
8	250683_CE_SDS_NON_RE	MAIN PEAK
9	Y_SM_95007441_R	B7 BINDING SPR

```
analysis
CE_SDS      1596
CE_SDS_2    1168
SDS_PAGE    1102
ICIEF       228
SE_HPLC     181
...
A_95011468  1
A_USP0236   1
A_95007196  1
A_249491    1
ELISA_5     1
Name: count, Length: 546, dtype: int64
```

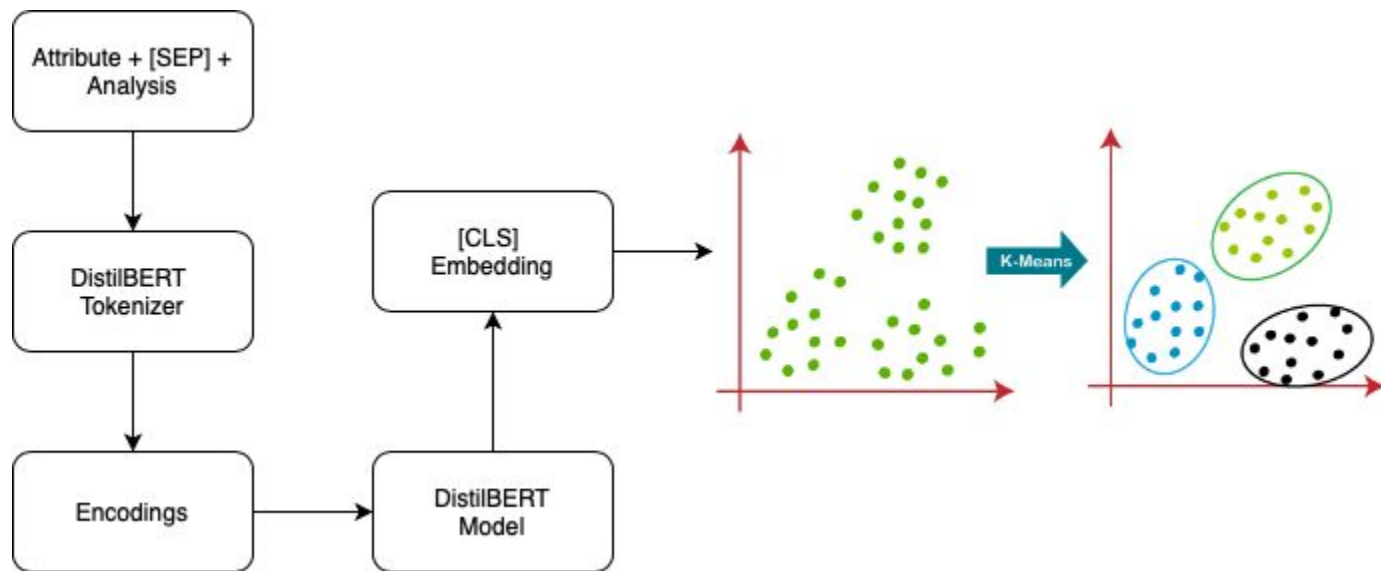
Imbalanced data

Tried to find for exact text match
(analysis,attribute,standard) to the
Attribute name in data

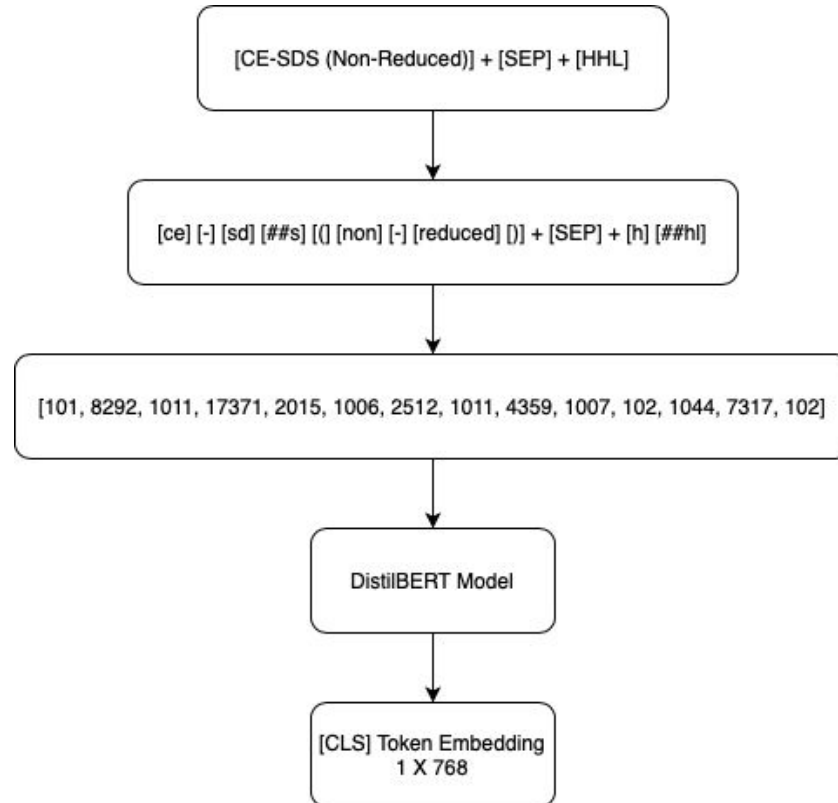
	att_match	ana_match	std_match	count
0	False	False	False	3923
1	False	True	False	2131
2	True	False	False	769
3	True	True	False	384
4	True	True	True	407

Was able to classify 400-700 rows
out of 7730. Not a good approach.

Approach



Example



Results

Some good clusters

Cell-Based Bioassay	Potency	Potency by Cell-Based Bioassay
BIOASSAY	Potency (Cell Based)	18
BIOLOGICAL_ASSAY	Potency Cell-based Assay	18
BIOASSAY	Cell-Based Bioassay Potency	18
BIOLOGICAL_ASSAY	Bioassay (Potency Relative to Reference Material)	18
BIOLOGICAL_ASSAY	Potency Relative to Reference Material	18
BIOASSAY	Cell Based Bioassay Ipilimumab (Result)	18
BIOASSAY	Potency (Cell-based Bioassay)	18
BIOASSAY	Bio-Assay	18
BIOASSAY	Cell Based Bioassay Nivolumab	18
BIOLOGICAL_ASSAY	Potency Relative to Reference Standard	18
BIOLOGICAL_ASSAY	Potency (Cell Based)	18
BIOASSAY	Potency (Cell Based Bioassay)	18
BIOASSAY	Human Cell IL-2 Inhibition Assay (Result)	18
BIOASSAY	Human Cell IL-2 Inhibition Assay (Result)	18
BIOLOGICAL_ASSAY	Potency (Cell Based Bioassay)	18

HIAC or UV	Particulate-Matter >= 10-um	Particulate-Matter >= 10-um
HIAC or UV	Particulate-Matter >= 25-um	Particulate-Matter >= 25-um
PARTICULATE_MATTER_2	Particles >= 2 Microns (Concentrate)	26
PARTICULATE_MATTER_2	Particles >= 3 micrometers	26
PARTICULATE_MATTER_3	Particulate Matter >= 25 micrometers	26
PARTICULATE_MATTER_MICRO_2	Particulate Matter (HIAC) >= 5-10 micrometers	26
PARTICULATE_MATTER_MICRO_2	Particulate Matter (HIAC) >= 25 micrometers	26
PARTICULATE_MATTER_2	Particulate Matter (HIAC) >= 3 micrometers (Concentrate)	26
PARTICULATE_MATTER_MICRO_2	Particulate Matter (HIAC) >= 10-25 micrometers	26
PARTICULATE_MATTER_MICRO_2	Particulate Matter (HIAC) >= 2-5 micrometers	26
PARTICULATE_MATTER_2	Particles >= 10 Microns (Concentrate)	26
PARTICULATE_MATTER_2	Particles >= 25 Microns (Concentrate)	26
PARTICULATE_MATTER	Subvisible Particulate Matter >= 5 and <10 micrometer	26
PARTICULATE_MATTER_2	Particle >= 25 micrometers	26
PARTICULATE_MATTER_2	Particle >= 10 micrometers	26

ELISA	Binding activity or Potency	Potency by Binding ELISA
ELISA_2	Potency ELISA Relative to Reference Standard	9
ELISA_BINDING_2	Activity Binding ELISA relative to reference standard	9
ELISA_BINDING	Potency (ELISA) binding capacity relative to reference stand	9
BINDING	Relative Binding ELISA	9
ELISA_2	Potency BCMA (ELISA)	9
ELISA_BINDING	Binding ELISA	9
ACTIVITY_ELISA	Binding Activity	9
ELISA_BINDING	Binding (ELISA) Nivolumab	9
ELISA_BINDING_2	Potency (ELISA)	9
ACTIVITY_ELISA	Binding Activity (ELISA)	9
ELISA_BINDING	Potency (ELISA) (Result)	9
ELISA_2	CD3 ELISA binding relative to reference standard	9

Results

CE_SDS_2	CE-SDS (Reduced) RRT of Unknown Minor Peak 3	1
CE_SDS_2	CE-SDS (Reduced) RRT of Unknown Minor Peak 2	1
CE_SDS	CE-SDS (Non-Reduced) Unknown Minor Peak 3 RRT	1
CE_SDS	CE-SDS (Non-Reduced) Unknown Minor Peak 1 RRT	1
CE_SDS_2	SDS-CGE (Non-Reduced) Main Peak	1
CE_SDS_2	CE-SDS (Reduced) Unknown Minor Peak 4	1
CE_SDS	CGE (Non-Reduced) Intact Main peak	1
CE_SDS_2	CE-SDS (Non-Reduced) Main Peak	1

CE_SDS_2	RRT~1.451/1	29
CE_SDS	RRT~1.404/1	29
CE_SDS	RRT~2.036/4	29
A_250494	PI RANGE 4.7 - 5.6%	29
CE_SDS_2	RRT~ 1.449	29
CE_SDS_2	RRT~1.782	29
95007425_ISOE_FOCUS	PI RANGE 4.3 TO 5.6	29
CE_SDS	RRT~ 2.045	29

CE_SDS	CE-SDS (Reduced) %: Undefined Peak_4>=LOQ	3
CE_SDS	CE-SDS (Reduced) %: Undefined Peak_5>=LOQ	3
CE_SDS	CE-SDS (Reduced) %: Undefined Peak_11>=LOQ	3
CE_SDS	CGE (Non-Reduced) Sum of Minor Peaks >= QL	3
CE_SDS	CGE (Non-Reduced) Sum of all minor peaks >= QL	3
CE_SDS	CE-SDS (Reduced) Sum of minor peaks >= LOQ	3
CE_SDS	CE-SDS (Non-Reduced) Cumulative area for all HMW peaks >= QL	3
CE_SDS	CE-SDS (Reduced) Sum of all minor peaks >= LOQ	3

Unknown Attributes

[illegible]

Junk data

Issues Faced

- Not enough labelled samples to train model
- Some Attribute-Analysis are very similar to each other

IEF	Acidic Peaks	IEF Acidic Peaks
IEF	Basic Peaks	IEF Basic Peaks
IEF	Main Peak	IEF Main Peak
iCIEF	Acidic Peaks	iCIEF Acidic Peaks
iCIEF	Basic Peaks	iCIEF Basic Peaks
iCIEF	Main Peak	iCIEF Main Peak

SE-HPLC	HMW	SE-HPLC HMW
SE-HPLC	LMW	SE-HPLC LMW
SE-HPLC	Monomer	SE-HPLC Monomer
SE-UPLC	HMW	SE-UPLC HMW
SE-UPLC	LMW	SE-UPLC LMW
SE-UPLC	Monomer	SE-UPLC Monomer

Solution - Clubbing classes

Analysis	Attribute	Standard names
CE-SDS (Non-Reduced)	HHL	CE-SDS (Non-Reduced) HHL
CE-SDS (Non-Reduced)	Purity	CE-SDS (Non-Reduced) Purity
CE-SDS (Reduced)	Purity	CE-SDS (Reduced) Purity
SDS-PAGE (Non-Reduced)	Purity	SDS-PAGE (Non-Reduced) Purity
SDS-PAGE (Reduced)	Purity	SDS-PAGE (Reduced) Purity
RP-HPLC	Purity (Main Peak)	RP-HPLC Purity
IEF	Acidic Peaks	IEF Acidic Peaks
IEF	Basic Peaks	IEF Basic Peaks
IEF	Main Peak	IEF Main Peak
ICIEF	Acidic Peaks	ICIEF Acidic Peaks
ICIEF	Basic Peaks	ICIEF Basic Peaks
ICIEF	Main Peak	ICIEF Main Peak
CEX	Acidic Peaks	CEX Acidic Peaks
CEX	Basic Peaks	CEX Basic Peaks
CEX	Main Peak	CEX Main Peak
AEX	Acidic Peaks	AEX Acidic Peaks
AEX	Basic Peaks	AEX Basic Peaks
AEX	Main Peak	AEX Main Peak

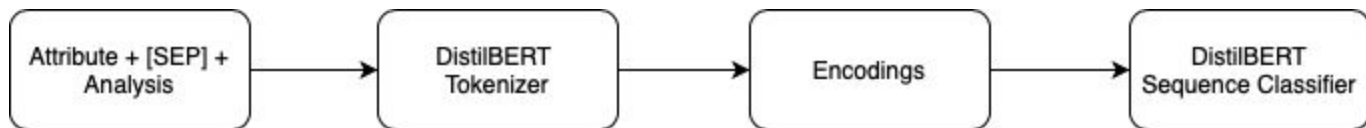
SE-HPLC	HMW	SE-HPLC HMW
SE-HPLC	LMW	SE-HPLC LMW
SE-HPLC	Monomer	SE-HPLC Monomer
SE-UPLC	HMW	SE-UPLC HMW
SE-UPLC	LMW	SE-UPLC LMW
SE-UPLC	Monomer	SE-UPLC Monomer
HIAC or UV	Particulate-Matter >= 10-um	Particulate-Matter >= 10-um
HIAC or UV	Particulate-Matter >= 25-um	Particulate-Matter >= 25-um
Cell-Based Bioassay	Potency	Potency by Cell-Based Bioassay
ELISA	Binding activity or Potency	Potency by Binding ELISA
SPR/Biocore	Binding activity	SPR Binding activity
pH	pH	pH
A280	Protein Concentration	Protein Concentration (A280)
Polysorbate 80	Polysorbate 80	Polysorbate 80

Solution - Training Data

- Manually labelled some of the data points
 - Were able to find 722 data points which could be classified directly
- Augmented data by replacing abbreviations with its full forms
 - With this we were able to increase the training size to 799,829 data points

BINDING_SPR	Binding Assay SPR Result (kd1)	SPR Binding activity
BINDING_Surface Plasmon Resonance	Binding Assay SPR Result (kd1)	SPR Binding activity
BINDING_SPR	Binding Assay Surface Plasmon Resonance Result (kd1)	SPR Binding activity

Approach



Results

CE_SDS_2	CGE(Reduced) Purity (LC and HC peaks)	0.999230623	0
CE_SDS_2	CE-SDS (Non-Reduced) Sum of Undefined Peaks Purity	0.999228954	0
CE_SDS_2	CE-SDS (Non-Reduced)Sum of Undefined Peaks Purity	0.999228954	0

ISOELECTRIC_FOCUSING	iCIEF Acidic Peaks Cumulative Area	0.999659538	2
CE_HPLC	Cation Exchange Chromatography Acidic Peak	0.999656558	2
ISOELECTRIC_FOCUSING	iCIEF Basic Peaks Cumulative Area	0.999653101	2
ISOELECTRIC_FOCUSING	IEF Acidic Peaks (Pre Peaks)	0.999651432	2

250432	PH 1 DECPT	0.731954038	8
PH	pH Determination at 25°C	0.7114712	8
D_JP_PH	PH	0.695997357	8
PH	pH (USP <791> EP: 2.2.3)	0.689385891	8

5683/7731 Classified with >0.99 P-Value

Results

J_TC-A-M51	√•-AE≈°√©,Ä°-è√Ø-°-èAssay	0.242341459	5
R_ORENCIAIV_VALORAC	Concentraci√É-≥n de Proteina - Abatacept	0.22535485	1
R_EMPLICITIINY_SEC	Mon√É-≥mero	0.223067954	3
Y_GM_95009656_R	OSMOLALITY	0.220089421	3
R_EMPLICITI_CONTDROG	Contenido de droga	0.216207653	2
A_250495_NONRED	SINGLE CHAIN	0.199646473	1

BIOLOGICAL_ASSAY_2	Potency(Cell Based) Nivolumab (Rel. to Reference Standard)	0.641247571	5
J_TB-OS-01	√§-[]-°√¶-≥-√•,Äπ,Äç√•-[]-Ø√©,Ä°-èmainMigration	0.641074777	2
T_M00000093	T_PUREZA_NOREDUC_PICO_PRINCIPAL	0.640622795	2
250683	LIGHT CHAIN 1DP	0.640277922	2

Next approach

- Use the outputs from BERT Sequence Classifier to further classify the samples into one of 32 classes.
 - Use Regex
 - Use KMeans clustering
- Use neural networks to take text input and outputs from BERT Classifier to classify the samples into one of 32 classes.

Cluster 0

Analysis	Attribute	Standard names
CE-SDS (Non-Reduced)	HHL	CE-SDS (Non-Reduced) HHL
CE-SDS (Non-Reduced)	Purity	CE-SDS (Non-Reduced) Purity
CE-SDS (Reduced)	Purity	CE-SDS (Reduced) Purity

N-gram	Frequency
cesds	~2800
sdspage	~1200
bicedsnr	~100
bicedsr	~50
ysmr	~20
cesdsreduc	~10
d	~5
cesdsreduc	~5
ysmnr	~5
a	~5
cesdsnonre	~5
c	~5
hplc	~5
yqds	~5
am	~5
m	~5
pm	~5
dm	~5
rabadspagesinred	~5
yusposds	~5
ymtcdesds	~5
gbl	~5
capillaryelectrophoresis	~5
purityimp	~5
ymrcases	~5
rabadspatiered	~5
anored	~5
asyn	~5

N-gram	Frequency
minor	2100
cerds	2050
peak	1850
loq	1350
unknown	1150
reduced	1100
of	900
rt	850
cpe	450
ql	400
for	350
time	300
peaks	300
migration	250
sum	200
peaklog	150
di	100
relative	100
ter	100
purity	100
nr	50
r	50
lmw	50
all	50
undefined	50
imp	50
retention	50
limit	50
lc	50
hmv	50
quantitation	50
main	50
nonreducedrt	50
number	50
red	50
reducedunknown	50
reducedrt	50
nonreducedunknown	50
hc	50
area	50
reducedrelative	50
peakseq	50
rel	50
nqhc	50
cehc	50
heavy	50

[illegible]

Can't classify them correctly to above category values as use of purity and hhl is found to be less. Need more information on these analysis and relation with terms like RRT(Relative Retention Time)
RMT(Relative Migration Time)
Minor peaks, LOQ