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
```

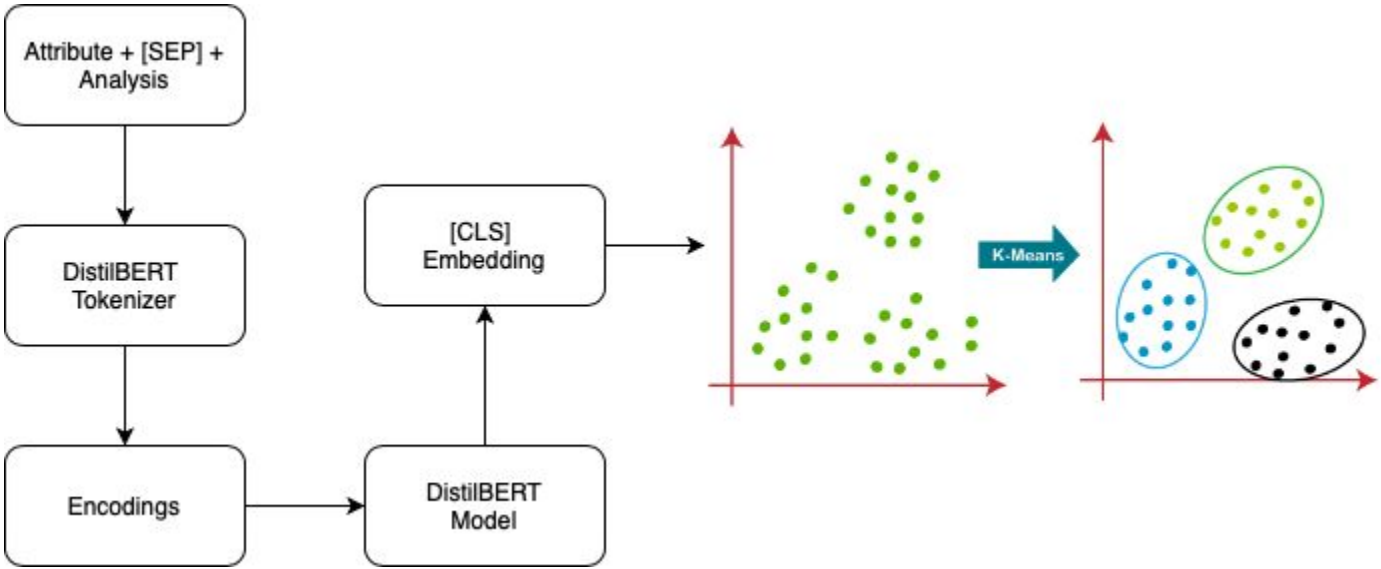
Imbalanced data

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

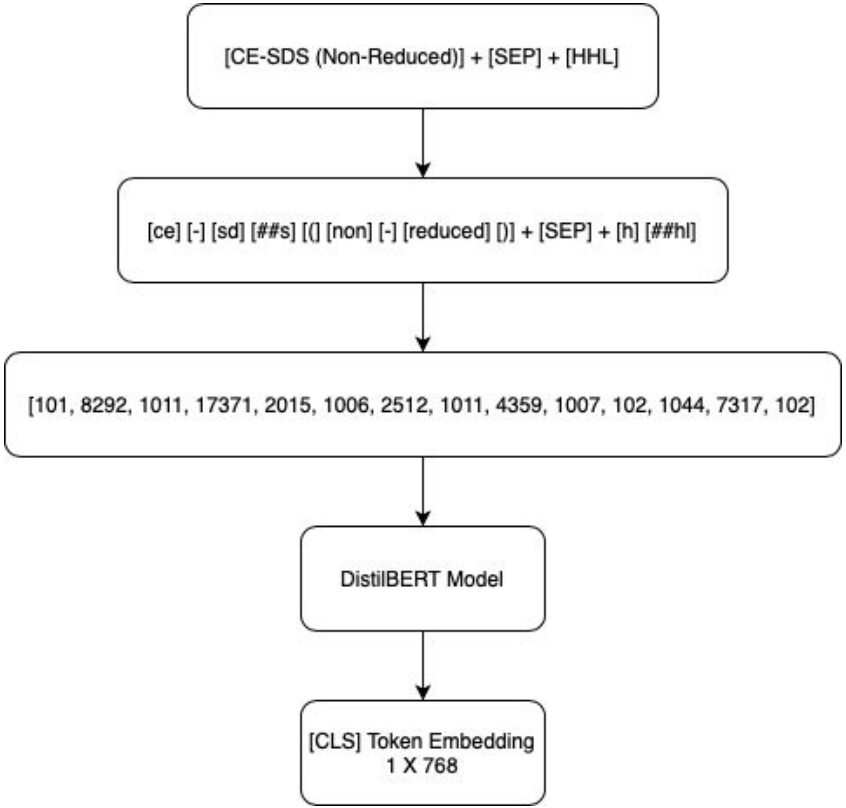
	att_match	ana_match	std_match	count
0	False	False	False	3923
1	False	True	False	2131
2	True	False	False	769
3	True	True	False	384
4	True	True	True	407

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

Approach 1 Kmeans clustering



Example


[CE-SDS (Non-Reduced)] + [SEP] + [HHL]
[ce] [-] [sd] [##s] [(] [non] [-] [reduced] [)] + [SEP] + [h] [##hl]
[101, 8292, 1011, 17371, 2015, 1006, 2512, 1011, 4359, 1007, 102, 1044, 7317, 102]
DistilBERT Model
[CLS] Token Embedding
1 X 768

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 Bioasay Nivolumab	18
BIOLOGICAL_ASSAY	Potency Relative to Rreference Standard	18
BIOLOGICAL_ASSAY	Potency (Cell Based)	18
BIOASSAY	Potency (Cell Based Biossay)	18
BIOASSAY	Human Cell IL-2 Inhibition Assay (Result)	18
BIOASSAY	Human Cell IL-2 Inhibition Assay(Result)	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	Particluate 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 micrometersons (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 Rreference 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 (some unknown clusters)

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

J_TC-A-U13-3_MV1	√£Δí¬¶√£Δí,Äπ√£Δí∆í√£Δí∆ÜÜ√Ø¬º,Äò√Ø¬º¬ê√Ø¬º¬èUnit10	13
J_TC-A-U13-2_MV1	√£Δí¬¶√£Δí,Äπ√£Δí∆í√£Δí∆ÜÜ√Ø¬º,Äù√Ø¬º¬èUnit4	13
J_TC-A-U13-3_MV1	√£Δí¬¶√£Δí,Äπ√£Δí∆í√£Δí∆ÜÜ√Ø¬º,Ñ¢√Ø¬º¬èUnit9	13
J_TC-A-U13-3_MV1	√£Δí¬¶√£Δí,Äπ√£Δí∆í√£Δí∆ÜÜ√Ø¬º,Äù√Ø¬º¬èUnit4	13
J_TC-A-U13-2_MV1	√£Δí¬¶√£Δí,Äπ√£Δí∆í√£Δí∆ÜÜ√Ø¬º,Ñ¢√Ø¬º¬èUnit9	13
J_TC-A-U13-2_MV1	√£Δí¬¶√£Δí,Äπ√£Δí∆í√£Δí∆ÜÜ√Ø¬º,Äò√Ø¬º¬èUnit1	13
J_TC-A-U13-3_MV1	√£Δí¬¶√£Δí,Äπ√£Δí∆í√£Δí∆ÜÜ√Ø¬º,Äô√Ø¬º¬èUnit2	13
J_TC-A-R38	√£Δí,Äù√£Δí¬º√£,Äö¬Ø√Ø¬º,Äò/peak1	13
J_TC-A-R36	√•¬º¬©√•≈∏¬¬∫√¶,Ç¨¬ß√•ÀÜ,Ät√•¬≠¬ê√ß¬®¬Æ/BasicSpecies	13
J_TB-OS-03	√•¬çÀú√®,Ä°¬è√§¬Ω,Äú√Ø¬º¬èlgG	13
J_TC-A-R36	√§¬∏¬ª√£Δí,Äù√£Δí¬º√£,Äö¬Ø/main√£,Ç¨,Ç¨peak	13
J_TC-A-U13-2_MV1	√£Δí¬¶√£Δí,Äπ√£Δí∆í√£Δí∆ÜÜ√Ø¬º,Äì√Ø¬º¬èUnit6	13

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

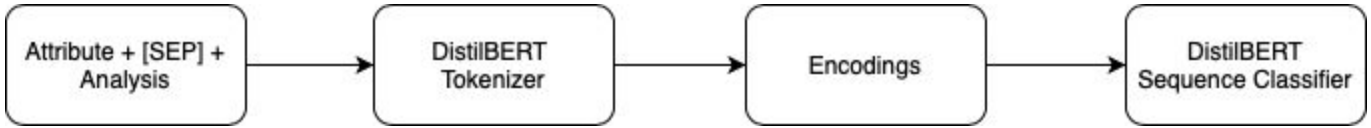
SE-HPLC	HMW	SE-HPLC HMW
SE-HPLC	LMW	SE-HPLC LMW
SE-HPLC	Monomer	SE-HPLC Monomer
SE-UPLC	HMW	SE-UPLC HMW
SE-UPLC	LMW	SE-UPLC LMW
SE-UPLC	Monomer	SE-UPLC Monomer
HIAC or UV	Particulate-Matter >= 10-um	Particulate-Matter >= 10-um
HIAC or UV	Particulate-Matter >= 25-um	Particulate-Matter >= 25-um
Cell-Based 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 around 2000 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


Attribute + [SEP] + Analysis
DistilBERT Tokenizer
Encodings
DistilBERT Sequence Classifier

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√Ç¬∞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	√•¬Æ≈°√©,Ä°¬è√Ø¬º¬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	√§¬∏¬ª√¶¬≥¬≥√•,Äπ,Ä¢√•¬∏¬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

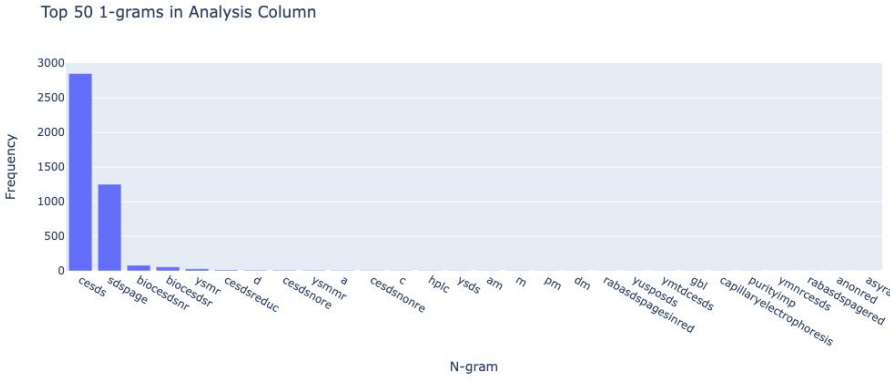
- Use the outputs from BERT Sequence Classifier to further classify the samples into one of 32 classes.
  - Use Regex
  - Use KMeans clustering

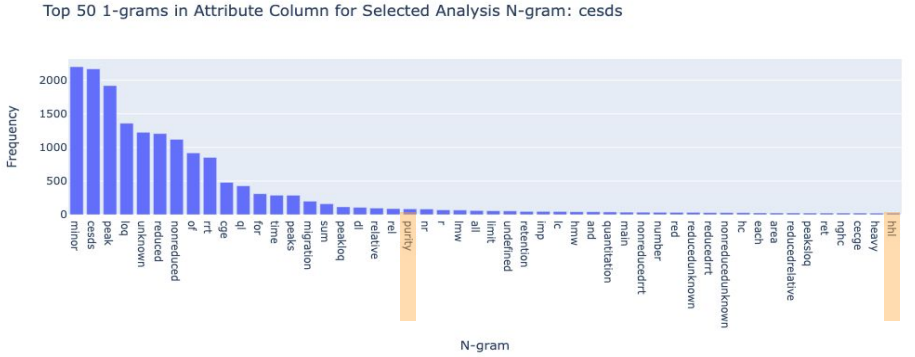
- Use neural networks to take text input and outputs from BERT Classifier to classify the samples into one of 32 classes.

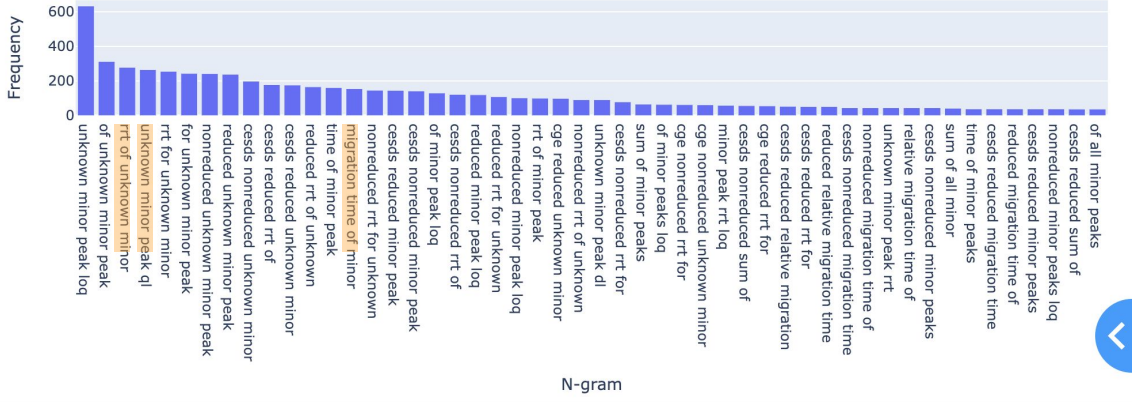
Issues

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







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

Solution

Consulted with the stakeholder regarding such data points and labelled them according to their domain knowledge, for improving training models.

Analysis	Attribute	label
CE_SDS_2	CGE(Reduced) Purity (LC and HC peaks)	CE-SDS (Reduced) purity
CE_SDS_2	CE-SDS (Non-reduced) Relative MigratioTime of Minor Peak 5	Irrelevant
CE_SDS	CE-SDS (Non-Reduced) Sum of Undefined Peaks	Irrelevant
CE_SDS_2	CE-SDS (Non-reduced) Sum of Minor Peaks >LOQ	Irrelevant
CE_SDS_2	CE-SDS (Reduced) Purity of heavy and light chain peaks	CE-SDS (Reduced) purity
CE_SDS	CE-SDS (Non-reduced) HHL Peak	CE-SDS (Non-reduced) HHL
CE_SDS	CE-SDS (Non-Reduced) Quantitative Limit	Irrelevant
CE_SDS_2	CE-SDS (Non-Reduced) RRT of Unknown Minor Peak <LOQ 1	Irrelevant
CE_SDS	CE-SDS (Reduced) Minor Peaks 6 (>LOQ)	Irrelevant
CE_SDS_2	CGE (Reduced) Purity (HC+LC)	CE-SDS (Reduced) purity
CE_SDS	CE-SDS Reduced Non-glycosylated Peak	Irrelevant
CE_SDS_2	CE-SDS (Reduced) Quantitation Limit	Irrelevant
CE_SDS	CE-SDS (Non-reduced) Limit of Quantitation	Irrelevant
CE_SDS	CE-SDS Reduced sum of heavy and light chain peaks	CE-SDS (Reduced) purity
CE_SDS	CE-SDS (NR) - Corrected Area% of IgG Purity	CE-SDS (Non-Reduced) purity
CE_SDS	CE-SDS (Reduced) Light Chain (LC)	Irrelevant
CE_SDS	CE-SDS (Non-Reduced) unknown imp RRT 2.46	Irrelevant
CE_SDS	CE-SDS (Reduced) Minor Peak 1 NLT LOQ	Irrelevant
CE_SDS	CE SDS (Non-Reduced) Monomer	CE-SDS (Non-Reduced) purity
CE_SDS	CE-SDS Reduced HMW Peak	Irrelevant
D_M00002527	PURITY REDUCED MAJOR BAND	CE-SDS (Reduced) purity
CE_SDS_2	CE-SDS (Non-Reduced) LC >=LOD	Irrelevant
CE_SDS_2	SDS-CGE (Non-Reduced) LMW Peaks	Irrelevant
CE_SDS	CGE (Non-Reduced) Quantitation Limit	Irrelevant
CE_SDS	CGE (Non-Reduced) F(ab')2	Irrelevant
CE_SDS	CGE (Non-Reduced) RRT for Unknown 1	Irrelevant
CE_SDS_2	CGE (reduced) LMW2	Irrelevant

A snapshot of that new labelled data
From stakeholder

	analysis	Attribute_name	class_probabilities	labels	max_prob	label
0	D_250475	IL2 INHIBITION ASSAY	[0.9641308188438416]	[5]	0.964131	5
1	D_95007196	PH	[0.997627317905426]	[8]	0.997627	8
2	D_M00003744	ABATACEPT MAJOR BAND (REDUCED)	[0.9975391626358032]	[0]	0.997539	0
3	Y_SM_95011468_R	BIOASSAY	[0.8629494905471802, 0.04849901795387268, 0.01...	[5, 9, 6, 10, 2]	0.862949	5
4	250684_CE_SDS_REDUC	SUM HEAVY AND LIGHT CHAIN	[0.9992501139640808]	[0]	0.99925	0
5	250580_TOTAL_PROT	PAAD	[0.9724606275558472]	[9]	0.972461	9
6	250580_TOTAL_PROT	SAMPLE 1 MASS	[0.9716528058052063]	[9]	0.971653	9
7	250684_CE_SDS_REDUC	SS RM 3 PURITY HC AND LC PEAKS	[0.9993009567260742]	[0]	0.999301	0
8	250683_CE_SDS_NON_RE	MAIN PEAK	[0.9989823698997498]	[0]	0.998982	0
9	Y_SM_95007441_R	B7 BINDING SPR	[0.979091227054596]	[7]	0.979091	7
10	Y_SM_95009421_R	FREE SULFHYDRAL GROUPS	[0.5766417980194092, 0.12531059980392456, 0.11...	[0, 8, 9, 7, 1, 4, 2]	0.576642	0
11	D_95007441	B7 BINDING	[0.977358341217041]	[7]	0.977358	7
12	D_95007423	SDS COOM REDUCED MAJOR BAND	[0.9993601441383362]	[0]	0.99936	0
13	250657_SIZE_HOMOG	MONOMER	[0.99868243932724]	[3]	0.998682	3
14	250684_CE_SDS_REDUC	SAMPLE HC RETENTION TIME	[0.9988405108451843]	[0]	0.998841	0
15	250683_CE_SDS_NON_RE	SAMPLE IGG RET TIME	[0.9987433552742004]	[0]	0.998743	0
16	250657_SIZE_HOMOG	HMW	[0.9988844990730286]	[3]	0.998884	3

Results on test data

Next steps

Classify them further to the appropriate standard by using various similarities score

1. cosine similarity with embeddings
2. tf_idf score with fuzzy matching