AGE AND SENTIMENT BASED E-COMMERCE CLOTHING RECOMMENDATION SYSTEM
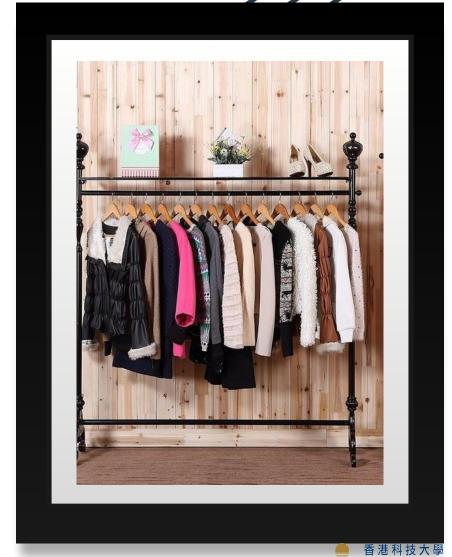
MSBD5003: BIG DATA COMPUTING

BANERJEE, Rohini

JUANG, Stefan Warren

MEHROTRA, Priya

PARK, Seoyoung
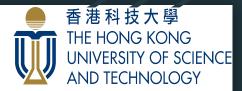
Background and Objective

- E-commerce has built their success story on top of their large user base and its feedbacks.

- A strong user data support base is now the very backbone of every e-commerce business .

- However, active interruption to the customers tend to cause more harm than good .

- To gain an edge in the competition, we propose a less intrusive machine learning model.

- Furthermore, we aim to improve user identification as well as bring a more intimate shopping experience to the users.

- We hope to build a personalized recommendation system which yields both customer satisfaction and business revenue.
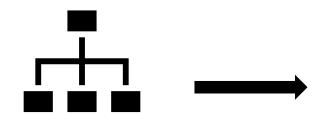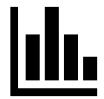
Dataset

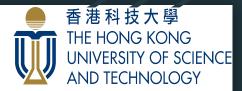23486 records, 10 features

→

Features:
{Clothing ID, Age, Title, Review, Rating, Recommendation IND, Positive Feedback Count, Division Name, Dept. Name, Class Name}

→

Extracted Features:
{Clothing ID, Age, Title, Review, Rating}

→

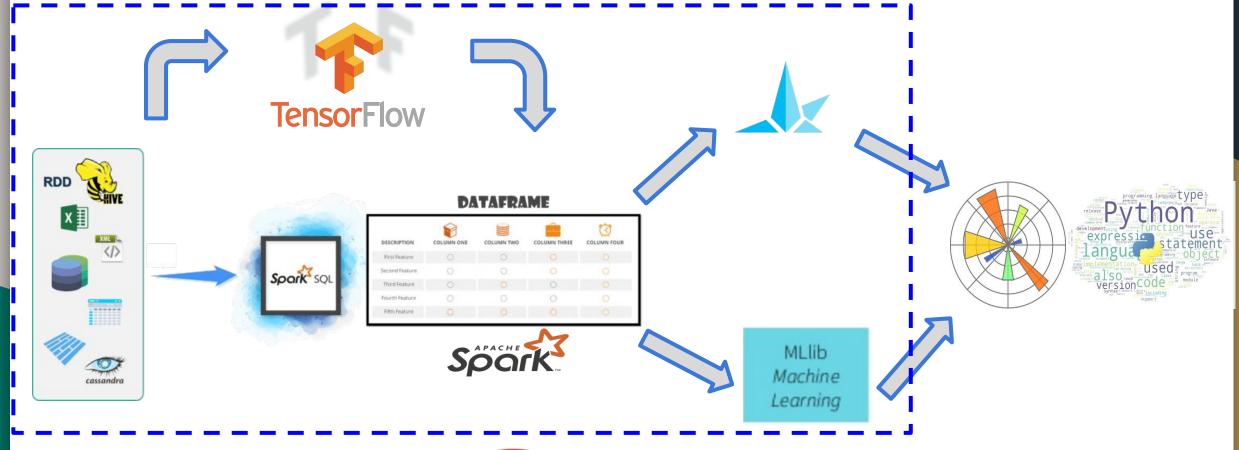NLP: Word embeddings of Reviews

Source: Kaggle

Actual Data View

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Reduce Rating Classes

```
[ ]  reviews_df.select('Rating').show()
```

```
[ ]  #reviews_df.filter(reviews_df.Rating == 2).select('Rating') = 1
     reviews_df = reviews_df.withColumn("Rating", F.when(reviews_df["Rating"] == 2, 1).otherwise(reviews_df["Rating"]))
     reviews_df = reviews_df.withColumn("Rating", F.when(reviews_df["Rating"] == 3, 2).otherwise(reviews_df["Rating"]))
     reviews_df = reviews_df.withColumn("Rating", F.when(reviews_df["Rating"] == 4, 3).otherwise(reviews_df["Rating"]))
     reviews_df = reviews_df.withColumn("Rating", F.when(reviews_df["Rating"] == 5, 3).otherwise(reviews_df["Rating"]))
     reviews_df.show()
```

```
+---+-----------+---+-----------------+-----------------+------+---------------+---------------------+---------------+---------------+-----------+
|_c0|Clothing ID|Age|            Title|      Review Text|Rating|Recommended IND|Positive Feedback Count|  Division Name|Department Name|Class Name|
+---+-----------+---+-----------------+-----------------+------+---------------+---------------------+---------------+---------------+-----------+
|  0|        767| 33|             null|Absolutely wonder...|     3|              1|                    0|       Initmates|       Intimate|  Intimates|
|  1|       1080| 34|             null|Love this dress! ...|     3|              1|                    4|        General|        Dresses|    Dresses|
|  2|       1077| 60|Some major design...|I had such high h...|     2|              0|                    0|        General|        Dresses|    Dresses|
|  3|       1049| 50|  My favorite buy!|I love, love, lov...|     3|              1|                    0|General Petite|        Bottoms|      Pants|
|  4|        847| 47|  Flattering shirt|This shirt is ver...|     3|              1|                    6|        General|           Tops|    Blouses|
|  5|       1080| 49|Not for the very ...|I love tracy rees...|     1|              0|                    4|        General|        Dresses|    Dresses|
|  6|        858| 39|Cagrcoal shimmer fun|I aded this in my...|     3|              1|                    1|General Petite|           Tops|      Knits|
|  7|        858| 39|Shimmer, surprisi...|I ordered this in...|     3|              1|                    4|General Petite|           Tops|      Knits|
|  8|       1077| 24|        Flattering|I love this dress...|     3|              1|                    0|        General|        Dresses|    Dresses|
|  9|       1077| 34|  Such a fun dress!|I'm 5"5' and 125 ...|     3|              1|                    0|        General|        Dresses|    Dresses|
| 10|       1077| 53|Dress looks like ...|Dress runs small ...|     2|              0|                   14|        General|        Dresses|    Dresses|
| 11|       1095| 39|             null|This dress is per...|     3|              1|                    2|General Petite|        Dresses|    Dresses|
| 12|       1095| 53|        Perfect!!!|More and more i f...|     3|              1|                    2|General Petite|        Dresses|    Dresses|
| 13|        767| 44|         Runs big|Bought the black ...|     3|              1|                    0|       Initmates|       Intimate|  Intimates|
| 14|       1077| 50|Pretty party dres...|This is a nice ch...|     2|              1|                    1|        General|        Dresses|    Dresses|
| 15|       1065| 47|Nice, but not for...|I took these out ...|     3|              1|                    3|        General|        Bottoms|      Pants|
| 16|       1065| 34|You need to be at...|Material and colo...|     2|              1|                    2|        General|        Bottoms|      Pants|
| 17|        853| 41|Looks great with ...|Took a chance on ...|     3|              1|                    0|        General|           Tops|    Blouses|
| 18|       1120| 32| Super cute and cozy|A flattering, sup...|     3|              1|                    0|        General|        Jackets|  Outerwear|
| 19|       1077| 47|Stylish and comfo...|I love the look a...|     3|              1|                    0|        General|        Dresses|    Dresses|
+---+-----------+---+-----------------+-----------------+------+---------------+---------------------+---------------+---------------+-----------+
only showing top 20 rows
```
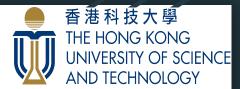
Technology System Architecture
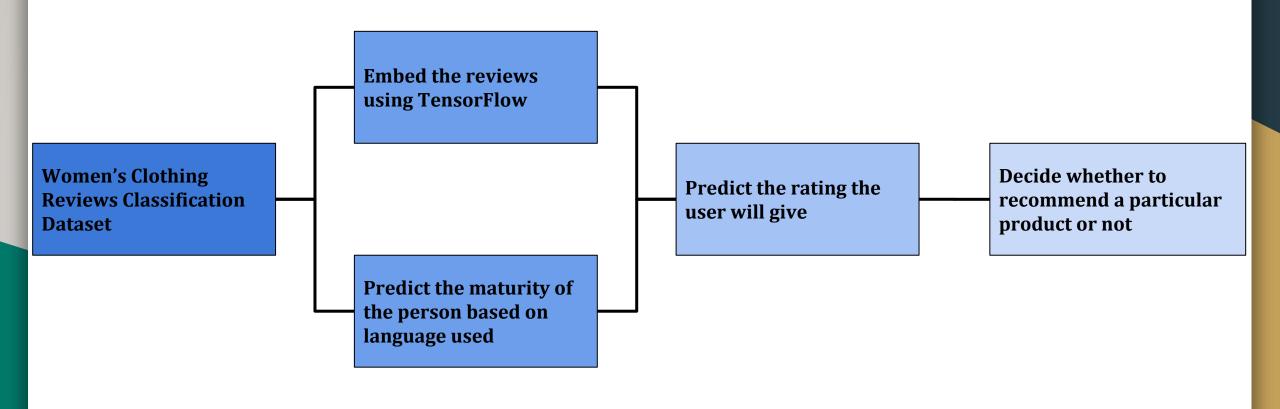
Process Flow Architecture

Google Cloud Engine
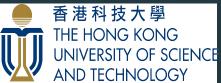
———

Benefits:

1. Per Second Pricing
2. Preemptible-VMs
3. Fast Startup Time

Control:

1. Web interface
2. Install gcloud
3. Cloud console

Dataproc

Google's cloud solution to HDFS and Spark

Create Preemptible Nodes

Fast and Affordable
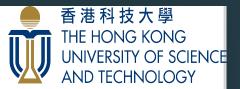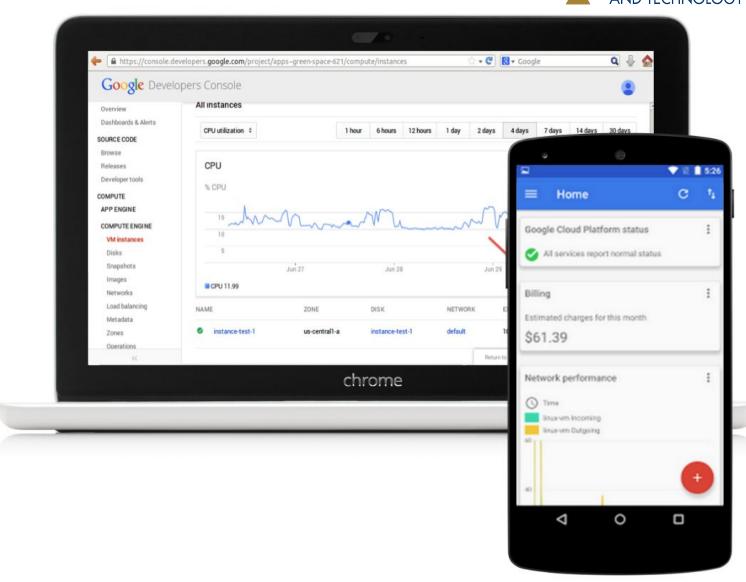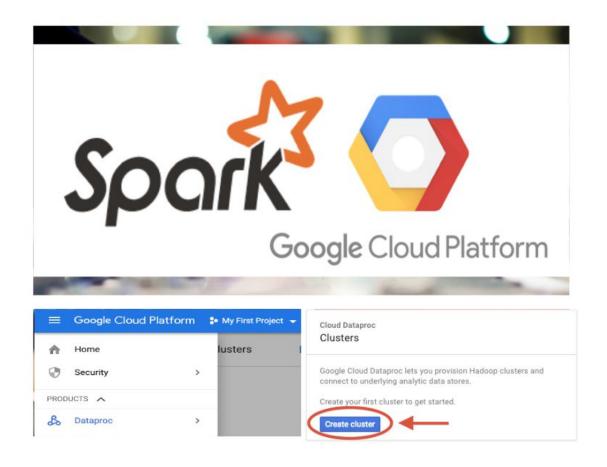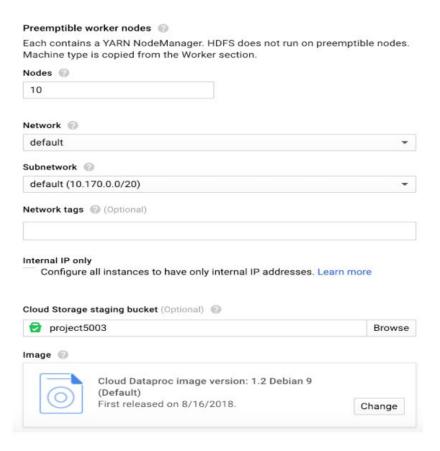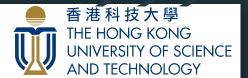
Cloud Bucket

香港科技大學
THE HONG KONG
UNIVERSITY OF SCIENCE
AND TECHNOLOGY

5003project

Objects   Overview   Permissions   Bucket Lock

[Upload files] [Upload folder] [Create folder] [Manage holds] [Delete]

🔍 Filter by prefix...

Buckets / 5003project

	Name	Size	Type	Storage class	Last modified	Public access ❓	Encryption ❓	Retention expiration date ❓	Holds ❓
☐	📁 embedded dataset/	—	Folder	—	—	Per object	—	—	-

• • •

Upload 5 of 10 complete		Cancel	—
5003embed2part2.csv		0%	✕
5003embed3part2.csv		0%	✕
5003embed4part2.csv		69%	✕
5003embed3part1.csv		59%	✕
5003embed2part1.csv		0%	✕
.DS_Store	Finished		✕
5003embed1part1.csv	Finished		✕
5003embed1part2.csv	Finished		✕
5003embed5_check.csv	Finished		✕
5003embed4part1.csv	Finished		

Bonus Story

Nov 1 – 21, 2018	Compute Engine Storage PD Capacity: 5709.527 Gibibyte-days (Source:My First Project [awesome-wares-196616])
Nov 1 – 19, 2018	Compute Engine Network Internet Egress from APAC to EMEA: 92.047 Gibibytes (Source:My First Project [awesome-wares-196616])
Nov 1 – 19, 2018	Compute Engine Network Internet Egress from APAC to China: 215.786 Gibibytes (Source:My First Project [awesome-wares-196616])
Nov 1 – 19, 2018	Compute Engine Network Internet Egress from APAC to Americas: 304.929 Gibibytes (Source:My First Project [awesome-wares-196616])
Nov 1 – 19, 2018	Compute Engine N1 Standard Instance Ram running in APAC: 719.762 Gibibyte-hours (Source:My First Project [awesome-wares-196616])
Nov 1 – 19, 2018	Compute Engine N1 Standard Instance Core running in APAC: 191.932 Hours (Source:My First Project [awesome-wares-196616])
Nov 1 – 19, 2018	Compute Engine N1 Standard Instance Core running in Montreal: 0.199 Hours (Source:My First Project [awesome-wares-196616])

Word2Vec Embedding
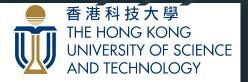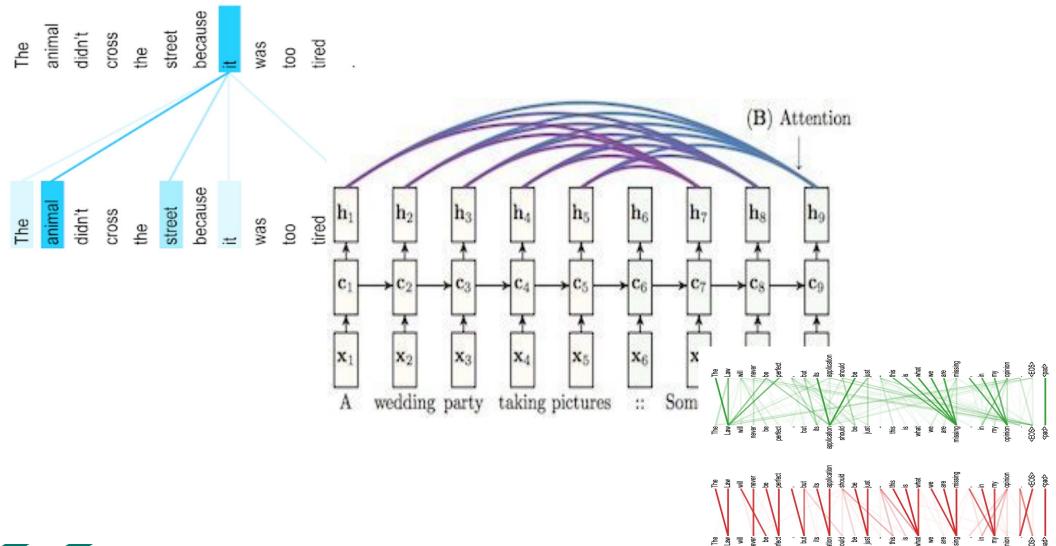

N = 1 : This is a sentence | unigrams: this, is, a, sentence

N = 2 : This is a sentence | bigrams: this is, is a, a sentence

N = 3 : This is a sentence | trigrams: this is a, is a sentence

Attention

(B) Attention

Spark SQL

Join Train Reviews and Embedded Train Reviews

```
reviews_train_df = reviews_train_df.withColumn("idx", F.monotonically_increasing_id())
windowSpec = W.orderBy("idx")
reviews_train_df = reviews_train_df.withColumn("idx", F.row_number().over(windowSpec))
```

```
embedded_train_df = embedded_train_df.withColumn("idx", F.monotonically_increasing_id())
windowSpec = W.orderBy("idx")
embedded_train_df = embedded_train_df.withColumn("idx", F.row_number().over(windowSpec))
```

```
joined_train_df = reviews_train_df.join(embedded_train_df, "idx", "inner")#.drop("id")
joined_train_df.sort('idx').show()
joined_train_df.count()
```

ion Name	Department Name	Class Name	buckets	0	1	2	3	4	5
nitmates	Intimate	Intimates	2.0	0.05118567	−0.041212857	−0.0057248306	−0.025765266	0.013969434	0.014644007
General	Dresses	Dresses	2.0	−0.049780436	−0.0056768777	−0.07220182	0.018154016	0.043466363	−0.06942644
General	Dresses	Dresses	8.0	−0.03431566	0.044290908	−0.08823467	0.049415097000000005	0.026357368	−0.08636762
l Petite	Bottoms	Pants	6.0	0.0051291888	−7.5167314E-5	−0.0039698277	0.016769648	0.03881027	0.026949976
General	Tops	Blouses	5.0	−0.06531402	0.01002855	−0.05822191400000...	0.025711114	0.050464835	−0.024829336
General	Dresses	Dresses	5.0	−0.028266564	0.027315315	−0.06372557	0.0023762516	0.019971153	−0.013511952
l Petite	Tops	Knits	3.0	−0.025104128	0.011437717	−0.05741755	−0.028304886	0.052708082	−0.02046696
l Petite	Tops	Knits	3.0	−0.045107137	−9.794285E-5	−0.02541966700000...	−0.053572617	0.00795994	−0.0112525895
General	Dresses	Dresses	0.0	−0.03791774	0.00579716	−0.06689097	0.009804043	0.0503048	−0.044445325
General	Dresses	Dresses	2.0	−0.08306482	0.005929275	−0.07538348	−0.0067714504	0.03828641	−0.03141347
General	Dresses	Dresses	6.0	−0.009730896	0.04085852	−0.08136655400000001	0.05665242700000001	−0.009846802	−0.059379723
l Petite	Dresses	Dresses	3.0	0.012596051	0.009572292	−0.030710638	0.046942815	0.03877946	−0.00679251099999...
l Petite	Dresses	Dresses	6.0	0.008398089000000001	0.09707141	−0.011302323	0.05413509	−0.015016393	−0.05515934
nitmates	Intimate	Intimates	4.0	−0.014400116	0.022485716	−0.054296497	−0.008358585	0.015852246	−0.06509562
General	Dresses	Dresses	6.0	−0.07168408	0.025857221	−0.057857223	0.039491888	−0.00998145	−0.08167605
General	Bottoms	Pants	5.0	0.0076641566	0.022749718	−0.07635751	0.026876854	0.04002443	−0.03100559700000...
General	Bottoms	Pants	2.0	−0.03321122799999...	0.0137859825	−0.06455126	0.022302862000000003	0.049921826	−0.038751185
General	Tops	Blouses	4.0	0.0025946386	0.045149054	−0.05982217	−0.0025410836	0.0013539282	0.0030134362
General	Jackets	Outerwear	2.0	−0.091148615	0.030286754	−0.049163673	−0.04572083	0.03219183	0.026147354
General	Dresses	Dresses	5.0	−0.04024081700000...	0.044220358	−0.08213224	0.011779263	0.023770964	−0.036075633

Processing Features to Feed into MLLib

Rating Prediction

```
[ ]  dataset = joined_train_df
     assembler = VectorAssembler(
         inputCols=joined_train_df.columns[13:],
         outputCol="features")

     output = assembler.transform(dataset)

     output.select("features", "Rating").show(truncate=False)
```

```
------------------------------------------------------------------------------------------------------------------+------+
                                                                                                                   |Rating|
------------------------------------------------------------------------------------------------------------------+------+
9E-4,-0.05213137,0.07372937,-0.027677443,0.011817402,-0.0070387577]                                               |3     |
2,-0.04583441,-0.01133969,-0.030305952,-0.072569564,0.008480181,-0.017752068]                                     |3     |
0.019683326,0.011867737,0.037569955,-0.059360363,0.06531501599999999,-0.015076288,0.020703854,-0.0082388725,-0.05834136,-0.0011875583,-0.04037633]|2     |
4516897,-0.019131834,-0.064243935,-0.0040371413,0.008215551]                                                      |3     |
.01574329,-0.053998664]                                                                                            |3     |
                                                                                                                   |1     |
).04523472,0.031509113,-0.07728269,0.01016765,0.032739375,-0.08297086,-0.035459254,0.0058610677]                  |3     |
847,-0.06401726599999999,-0.007956901,0.007676328]                                                                |3     |
36,-0.0243513,0.013246751,0.025859231,-0.05608427,-0.03886247,0.011749377]                                        |3     |
699394,-0.011770535]                                                                                              |3     |
2290625E-4,-0.008017772,-0.0036323604,-0.050605442,-0.018519485,-0.04969981]                                      |2     |
0.003617297,-0.07195977,-0.015857775,0.019142233,-0.06537914,-0.004147601,-0.029163318]                           |3     |
24243562000000003,-0.07229552,-0.026330158,-0.079427995]                                                          |3     |
                                                                                                                   |3     |
44,-0.02484425,-0.04237938,-0.07592112,0.0010778197,-0.054639824]                                                 |2     |
,0.026265921,-0.06299139599999999,0.01821281,-0.06270449]                                                         |3     |
5440973,0.014821117,-0.015285174,0.059832316,-0.007449161700000002,-0.04662087,0.042234063,-0.067290634,0.021312706,-0.059656594]|2     |
1322058,-0.027671162000000003,-0.031308625]                                                                       |3     |
,0.06605594,-0.034671022999999995,-0.06663659,0.06551865,-0.08772673,0.025945032000000003,-0.07417156]            |3     |
20486,-0.08166112,-0.024973584,1.5873302E-4]                                                                      |3     |
------------------------------------------------------------------------------------------------------------------+------+
```
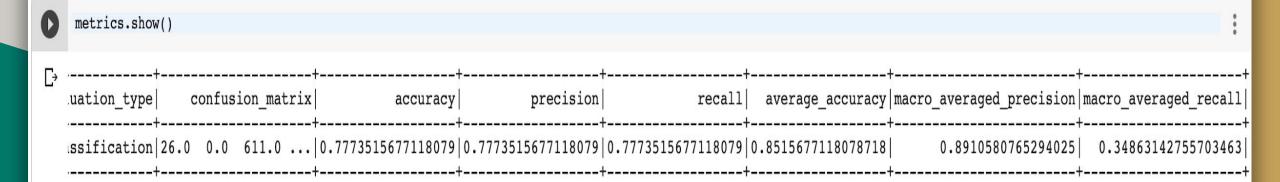
Microsoft Machine Learning for Apache Spark

```python
from mmlspark import TuneHyperparameters
from mmlspark.TrainClassifier import TrainClassifier
from pyspark.ml.classification import LogisticRegression, RandomForestClassifier, GBTClassifier
logReg = LogisticRegression()
randForest = RandomForestClassifier()
smlmodels = [logReg, randForest]
mmlmodels = [TrainClassifier(model=model, labelCol="Label") for model in smlmodels]
```

Results: An Overview

```
classifier: LogisticRegression_408ab156626698a54d4b, featuresCol: TrainClassifier_4d8f85d47c080c62a615_features, labelCol: label, predictionCol: prediction
TrainClassifier_a83a96487761
```

```
metrics.show()
```

uation_type	confusion_matrix	accuracy	precision	recall	average_accuracy	macro_averaged_precision	macro_averaged_recall
ssification	26.0 0.0 611.0 ...	0.7773515677118079	0.7773515677118079	0.7773515677118079	0.8515677118078718	0.8910580765294025	0.34863142755703463

PCA: Information and Runtime Tradeoff

PCA
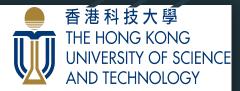
```
[30]  # Load training data
      data = output.select("features", "Rating")
      data = data.toDF('features', 'label')


      n_components = 100
      df = data.select("features", 'label')
      pca = PCA(k=n_components, inputCol="features", outputCol="pca_features")
      model = pca.fit(df)

      np.sum(model.explainedVariance)
      # DenseVector([0.7944, 0.2056])
```

```
0.9540048450840753
```

Spark MML: Rating Prediction Results

```
[33] layers = [n_components, 40, 10, 4]

     # create the trainer and set its parameters
     trainer = MultilayerPerceptronClassifier(maxIter=200, layers=layers, blockSize=128, seed=1234)

     # train the model
     model = trainer.fit(train)
```

```
[34] result = model.transform(test)
     p = result.select('prediction', 'label')
     e = MulticlassClassificationEvaluator(metricName='accuracy')
     print(str(e.evaluate(p)))
```

```
     0.8270513675783856
```

```
[35] p.select('prediction').distinct().show()
```

```
     +----------+
     |prediction|
     +----------+
     |       3.0|
     |       2.0|
     |       1.0|
     +----------+
```

Spark MML: Prediction of Maturity

21	Love these pants
23	Soft & lovely
23	
23	Beautifully made, but not versatile
23	Not a fan
23	Perfect fall shirt!
22	
21	Beautiful top!
23	Cuter is person
22	Not the best quality
23	Great dress, rough zipper
22	Love the lace up design!
23	Great too but not for women who have a large bust.
20	Super cute and flattering too
22	Fun jeggings
22	
23	Too short
23	The worst
22	Love it!
23	Amazing
23	Shrinks!!!

51	
52	Cute tee
51	Very vintage feel
54	Cozy casual - perfect for fall
52	Cute vest but...
52	Runs short
51	Great classic
53	Not impressed...
53	
52	Scratchy, uncomfortable top
52	Classy and cute
56	Just as pictured
55	Great purchase
51	Great for summer
52	
54	It's gorgeous, but...
52	Nice but too thin

Teenager

Old Timer