AgeDetectionParser

Since TIKA-1988 Tika has the ability to extract a person's age from text. We use an Ensemble Maximum Entropy Classification and Linear Regression model built by the USC Information Retrieval & Data Science group, the paper that describes the approach is here.

```
@article{hong2016ensemble,
  title={Ensemble Maximum Entropy Classification and Linear Regression for Author Age Prediction},
  author={Hong, Joey and Mattmann, Chris and Ramirez, Paul},
  journal={arXiv preprint arXiv:1610.00852},
  year={2016},
  url={https://arxiv.org/abs/1610.00852}
}
```

Pre-requisites

None, all of the needed USC IRDS models (for age classification {{`}}classify-bigram.bin{{`}}, and {{`}}classify-bigram.bin{{`}}) are downloaded automatically and will be available in your `$TIKASRC/tika-nlp/model{{`}} directory.

Tests to Run Beforehand

It's worth trying to see if Age Prediction works for you before using it in Tika. To do so:

Download and Build AgePredictor

1. cd $HOME/src && git clone https://github.com/USCDataScience/AgePredictor.git 2. cd AgePredictor && mvn install

Test AgePredictor

1. {{`}}java -cp age-predictor-assembly/target/age-predictor-assembly-1.1-SNAPSHOT-jar-with-dependencies.jar edu.usc.irds.agepredictor.authorage.AgePredicterLocal I am actually very young now{{`}}

The above should print something like:

```
$ java -cp age-predictor-assembly/target/age-predictor-assembly-1.1-SNAPSHOT-jar-with-dependencies.jar edu.usc.
irds.agepredictor.authorage.AgePredicterLocal I am actually very young now
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
17/07/06 23:23:24 INFO SparkContext: Running Spark version 2.0.0
17/07/06 23:23:24 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using
builtin-java classes where applicable
17/07/06 23:23:24 INFO SecurityManager: Changing view acls to: mattmann
17/07/06 23:23:24 INFO SecurityManager: Changing modify acls to: mattmann
17/07/06 23:23:24 INFO SecurityManager: Changing view acls groups to:
17/07/06 23:23:24 INFO SecurityManager: Changing modify acls groups to:
17/07/06 23:23:24 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users  with
view permissions: Set(mattmann); groups with view permissions: Set(); users  with modify permissions: Set
(mattmann); groups with modify permissions: Set()
17/07/06 23:23:25 INFO Utils: Successfully started service 'sparkDriver' on port 54970.
17/07/06 23:23:25 INFO SparkEnv: Registering MapOutputTracker
17/07/06 23:23:25 INFO SparkEnv: Registering BlockManagerMaster
17/07/06 23:23:25 INFO DiskBlockManager: Created local directory at /private/var/folders/n5
/1d_k3z4s2293q8ntx_n8sw54mm5n_8/T/blockmgr-aa033554-acff-4ea1-a5d1-250257f467dc
17/07/06 23:23:25 INFO MemoryStore: MemoryStore started with capacity 2004.6 MB
17/07/06 23:23:25 INFO SparkEnv: Registering OutputCommitCoordinator
17/07/06 23:23:25 INFO Utils: Successfully started service 'SparkUI' on port 4040.
17/07/06 23:23:25 INFO SparkUI: Bound SparkUI to 0.0.0.0, and started at http://172.20.10.2:4040
17/07/06 23:23:25 INFO Executor: Starting executor ID driver on host localhost
17/07/06 23:23:25 INFO Utils: Successfully started service 'org.apache.spark.network.netty.
NettyBlockTransferService' on port 54971.
17/07/06 23:23:25 INFO NettyBlockTransferService: Server created on 172.20.10.2:54971
17/07/06 23:23:25 INFO BlockManagerMaster: Registering BlockManager BlockManagerId(driver, 172.20.10.2, 54971)
17/07/06 23:23:25 INFO BlockManagerMasterEndpoint: Registering block manager 172.20.10.2:54971 with 2004.6 MB
RAM, BlockManagerId(driver, 172.20.10.2, 54971)
17/07/06 23:23:25 INFO BlockManagerMaster: Registered BlockManager BlockManagerId(driver, 172.20.10.2, 54971)
17/07/06 23:23:26 WARN SparkContext: Use an existing SparkContext, some configuration may not take effect.
```

```
17/07/06 23:23:26 INFO SharedState: Warehouse path is 'file:/Users/mattmann/git/AgePredictor/spark-warehouse'.
17/07/06 23:23:39 INFO MemoryStore: Block broadcast_0 stored as values in memory (estimated size 6.1 MB, free
1998.5 MB)
17/07/06 23:23:39 INFO MemoryStore: Block broadcast_0_piece0 stored as bytes in memory (estimated size 488.5
KB, free 1998.0 MB)
17/07/06 23:23:39 INFO BlockManagerInfo: Added broadcast_0_piece0 in memory on 172.20.10.2:54971 (size: 488.5
KB, free: 2004.1 MB)
17/07/06 23:23:39 INFO SparkContext: Created broadcast 0 from broadcast at CountVectorizer.scala:243
17/07/06 23:23:42 INFO CodeGenerator: Code generated in 173.162228 ms
17/07/06 23:23:42 INFO SparkContext: Starting job: first at AgePredicterLocal.java:114
17/07/06 23:23:42 INFO DAGScheduler: Got job 0 (first at AgePredicterLocal.java:114) with 1 output partitions
17/07/06 23:23:42 INFO DAGScheduler: Final stage: ResultStage 0 (first at AgePredicterLocal.java:114)
17/07/06 23:23:42 INFO DAGScheduler: Parents of final stage: List()
17/07/06 23:23:42 INFO DAGScheduler: Missing parents: List()
17/07/06 23:23:42 INFO DAGScheduler: Submitting ResultStage 0 (MapPartitionsRDD[3] at javaRDD at
AgePredicterLocal.java:112), which has no missing parents
17/07/06 23:23:42 INFO MemoryStore: Block broadcast_1 stored as values in memory (estimated size 10.9 KB, free
1998.0 MB)
17/07/06 23:23:42 INFO MemoryStore: Block broadcast_1_piece0 stored as bytes in memory (estimated size 5.4 KB,
free 1998.0 MB)
17/07/06 23:23:42 INFO BlockManagerInfo: Added broadcast_1_piece0 in memory on 172.20.10.2:54971 (size: 5.4 KB,
free: 2004.1 MB)
17/07/06 23:23:42 INFO SparkContext: Created broadcast 1 from broadcast at DAGScheduler.scala:1012
17/07/06 23:23:42 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 0 (MapPartitionsRDD[3] at
javaRDD at AgePredicterLocal.java:112)
17/07/06 23:23:42 INFO TaskSchedulerImpl: Adding task set 0.0 with 1 tasks
17/07/06 23:23:42 INFO TaskSetManager: Starting task 0.0 in stage 0.0 (TID 0, localhost, partition 0,
PROCESS_LOCAL, 6671 bytes)
17/07/06 23:23:42 INFO Executor: Running task 0.0 in stage 0.0 (TID 0)
17/07/06 23:23:42 INFO CodeGenerator: Code generated in 21.816953 ms
17/07/06 23:23:42 INFO Executor: Finished task 0.0 in stage 0.0 (TID 0). 3381 bytes result sent to driver
17/07/06 23:23:42 INFO TaskSetManager: Finished task 0.0 in stage 0.0 (TID 0) in 123 ms on localhost (1/1)
17/07/06 23:23:42 INFO TaskSchedulerImpl: Removed TaskSet 0.0, whose tasks have all completed, from pool
17/07/06 23:23:42 INFO DAGScheduler: ResultStage 0 (first at AgePredicterLocal.java:114) finished in 0.139 s
17/07/06 23:23:42 INFO DAGScheduler: Job 0 finished: first at AgePredicterLocal.java:114, took 0.221228 s

===================

Text received- 'I am actually very young now '
 Predicted Age - 34.983567


===================

17/07/06 23:23:43 INFO SparkContext: Invoking stop() from shutdown hook
17/07/06 23:23:43 INFO SparkUI: Stopped Spark web UI at http://172.20.10.2:4040
17/07/06 23:23:43 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
17/07/06 23:23:43 INFO MemoryStore: MemoryStore cleared
17/07/06 23:23:43 INFO BlockManager: BlockManager stopped
17/07/06 23:23:43 INFO BlockManagerMaster: BlockManagerMaster stopped
17/07/06 23:23:43 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
17/07/06 23:23:43 INFO SparkContext: Successfully stopped SparkContext
17/07/06 23:23:43 INFO ShutdownHookManager: Shutdown hook called
17/07/06 23:23:43 INFO ShutdownHookManager: Deleting directory /private/var/folders/n5
/1d_k3z4s2293q8ntx_n8sw54mm5n_8/T/spark-31e98436-00e3-4020-a4c5-784ec75b16de
$
```

In this case you are good shape. If not, please report a bug here

Running AgeRecogniser (Tika's Parser)

To run AgeRecogniser, download and install Tika 1.16 or later, and then run the following (make sure you have a $TIKASRC/tika-nlp/model directory populated with models before running this per above)

1. {{`}}cd $HOME/src/ && git clone https://github.com/apache/tika.git{{`}} 2. {{`}}cd tika-nlp && echo "I am a test file" > test.txt{{`}} 2. {{`}}java -cp ../tika-app/target/tika-app-1.16-SNAPSHOT.jar:target/tika-nlp-1.16-SNAPSHOT-jar-with-dependencies.jar:./model org.apache.tika.cli.TikaCLI --config=src/test/resources/org/apache/tika/parser/recognition/tika-config-age.xml -m test.txt{{`}}

You should then see:

```
$java -cp ../tika-app/target/tika-app-1.16-SNAPSHOT.jar:target/tika-nlp-1.16-SNAPSHOT-jar-with-dependencies.
jar:./model org.apache.tika.cli.TikaCLI --config=src/test/resources/org/apache/tika/parser/recognition/tika-
config-age.xml -m test.txt
Jul 07, 2017 3:38:53 PM org.apache.tika.config.InitializableProblemHandler$3 handleInitializableProblem
WARNING: JBIG2ImageReader not loaded. jbig2 files will be ignored
See https://pdfbox.apache.org/2.0/dependencies.html#jai-image-io
for optional dependencies.
TIFFImageWriter not loaded. tiff files will not be processed
See https://pdfbox.apache.org/2.0/dependencies.html#jai-image-io
for optional dependencies.
J2KImageReader not loaded. JPEG2000 files will not be processed.
See https://pdfbox.apache.org/2.0/dependencies.html#jai-image-io
for optional dependencies.

Jul 07, 2017 3:38:53 PM org.apache.tika.config.InitializableProblemHandler$3 handleInitializableProblem
WARNING: Tesseract OCR is installed and will be automatically applied to image files unless
you've excluded the TesseractOCRParser from the default parser.
Tesseract may dramatically slow down content extraction (TIKA-2359).
As of Tika 1.15 (and prior versions), Tesseract is automatically called.
In future versions of Tika, users may need to turn the TesseractOCRParser on via TikaConfig.
Jul 07, 2017 3:38:53 PM org.apache.tika.config.InitializableProblemHandler$3 handleInitializableProblem
WARNING: org.xerial's sqlite-jdbc is not loaded.
Please provide the jar on your classpath to parse sqlite files.
See tika-parsers/pom.xml for the correct version.
INFO  Running Spark version 2.0.0
WARN  Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
INFO  Changing view acls to: mattmann
INFO  Changing modify acls to: mattmann
INFO  Changing view acls groups to:
INFO  Changing modify acls groups to:
INFO  SecurityManager: authentication disabled; ui acls disabled; users  with view permissions: Set(mattmann);
groups with view permissions: Set(); users  with modify permissions: Set(mattmann); groups with modify
permissions: Set()
INFO  Successfully started service 'sparkDriver' on port 60109.
INFO  Registering MapOutputTracker
INFO  Registering BlockManagerMaster
INFO  Created local directory at /private/var/folders/n5/1d_k3z4s2293q8ntx_n8sw54mm5n_8/T/blockmgr-253e0d76-
fef5-42bb-b2a9-1500e807797c
INFO  MemoryStore started with capacity 2004.6 MB
INFO  Registering OutputCommitCoordinator
INFO  Logging initialized @1726ms
INFO  jetty-9.2.z-SNAPSHOT
INFO  Started o.s.j.s.ServletContextHandler@12bcd0c0{/jobs,null,AVAILABLE}
INFO  Started o.s.j.s.ServletContextHandler@4879f0f2{/jobs/json,null,AVAILABLE}
INFO  Started o.s.j.s.ServletContextHandler@47db5fa5{/jobs/job,null,AVAILABLE}
INFO  Started o.s.j.s.ServletContextHandler@354fc8f0{/jobs/job/json,null,AVAILABLE}
INFO  Started o.s.j.s.ServletContextHandler@41813449{/stages,null,AVAILABLE}
INFO  Started o.s.j.s.ServletContextHandler@4678a2eb{/stages/json,null,AVAILABLE}
INFO  Started o.s.j.s.ServletContextHandler@5b43fbf6{/stages/stage,null,AVAILABLE}
INFO  Started o.s.j.s.ServletContextHandler@1080b026{/stages/stage/json,null,AVAILABLE}
INFO  Started o.s.j.s.ServletContextHandler@58ebfd03{/stages/pool,null,AVAILABLE}
INFO  Started o.s.j.s.ServletContextHandler@5b07730f{/stages/pool/json,null,AVAILABLE}
INFO  Started o.s.j.s.ServletContextHandler@1fdfafd2{/storage,null,AVAILABLE}
INFO  Started o.s.j.s.ServletContextHandler@a4b2d8f{/storage/json,null,AVAILABLE}
INFO  Started o.s.j.s.ServletContextHandler@dcfda20{/storage/rdd,null,AVAILABLE}
INFO  Started o.s.j.s.ServletContextHandler@6d304f9d{/storage/rdd/json,null,AVAILABLE}
INFO  Started o.s.j.s.ServletContextHandler@f73dcd6{/environment,null,AVAILABLE}
INFO  Started o.s.j.s.ServletContextHandler@5c87bfe2{/environment/json,null,AVAILABLE}
INFO  Started o.s.j.s.ServletContextHandler@2fea7088{/executors,null,AVAILABLE}
INFO  Started o.s.j.s.ServletContextHandler@40499e4f{/executors/json,null,AVAILABLE}
INFO  Started o.s.j.s.ServletContextHandler@51cd7ffc{/executors/threadDump,null,AVAILABLE}
INFO  Started o.s.j.s.ServletContextHandler@30d4b288{/executors/threadDump/json,null,AVAILABLE}
INFO  Started o.s.j.s.ServletContextHandler@4cc6fa2a{/static,null,AVAILABLE}
INFO  Started o.s.j.s.ServletContextHandler@40f1be1b{/,null,AVAILABLE}
INFO  Started o.s.j.s.ServletContextHandler@7a791b66{/api,null,AVAILABLE}
INFO  Started o.s.j.s.ServletContextHandler@6f2cb653{/stages/stage/kill,null,AVAILABLE}
INFO  Started ServerConnector@71b3bc45{HTTP/1.1}{0.0.0.0:4040}
INFO  Started @1831ms
INFO  Successfully started service 'SparkUI' on port 4040.
INFO  Bound SparkUI to 0.0.0.0, and started at http://192.168.1.65:4040
```

```
INFO  Starting executor ID driver on host localhost
INFO  Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 60110.
INFO  Server created on 192.168.1.65:60110
INFO  Registering BlockManager BlockManagerId(driver, 192.168.1.65, 60110)
INFO  Registering block manager 192.168.1.65:60110 with 2004.6 MB RAM, BlockManagerId(driver, 192.168.1.65,
60110)
INFO  Registered BlockManager BlockManagerId(driver, 192.168.1.65, 60110)
INFO  Started o.s.j.s.ServletContextHandler@7db0565c{/metrics/json,null,AVAILABLE}
WARN  Use an existing SparkContext, some configuration may not take effect.
INFO  Started o.s.j.s.ServletContextHandler@7692cd34{/SQL,null,AVAILABLE}
INFO  Started o.s.j.s.ServletContextHandler@32c0915e{/SQL/json,null,AVAILABLE}
INFO  Started o.s.j.s.ServletContextHandler@7dd712e8{/SQL/execution,null,AVAILABLE}
INFO  Started o.s.j.s.ServletContextHandler@22ee2d0{/SQL/execution/json,null,AVAILABLE}
INFO  Started o.s.j.s.ServletContextHandler@4b770e40{/static/sql,null,AVAILABLE}
INFO  Warehouse path is 'file:/Users/mattmann/tmp/tika1.15/tika-nlp/spark-warehouse'.
INFO  Block broadcast_0 stored as values in memory (estimated size 6.1 MB, free 1998.5 MB)
INFO  Block broadcast_0_piece0 stored as bytes in memory (estimated size 488.5 KB, free 1998.0 MB)
INFO  Added broadcast_0_piece0 in memory on 192.168.1.65:60110 (size: 488.5 KB, free: 2004.1 MB)
INFO  Created broadcast 0 from broadcast at CountVectorizer.scala:243
INFO  Code generated in 1537.312409 ms
INFO  Starting job: first at AgePredicterLocal.java:114
INFO  Got job 0 (first at AgePredicterLocal.java:114) with 1 output partitions
INFO  Final stage: ResultStage 0 (first at AgePredicterLocal.java:114)
INFO  Parents of final stage: List()
INFO  Missing parents: List()
INFO  Submitting ResultStage 0 (MapPartitionsRDD[3] at javaRDD at AgePredicterLocal.java:112), which has no
missing parents
INFO  Block broadcast_1 stored as values in memory (estimated size 10.5 KB, free 1998.0 MB)
INFO  Block broadcast_1_piece0 stored as bytes in memory (estimated size 5.3 KB, free 1998.0 MB)
INFO  Added broadcast_1_piece0 in memory on 192.168.1.65:60110 (size: 5.3 KB, free: 2004.1 MB)
INFO  Created broadcast 1 from broadcast at DAGScheduler.scala:1012
INFO  Submitting 1 missing tasks from ResultStage 0 (MapPartitionsRDD[3] at javaRDD at AgePredicterLocal.java:
112)
INFO  Adding task set 0.0 with 1 tasks
INFO  Starting task 0.0 in stage 0.0 (TID 0, localhost, partition 0, PROCESS_LOCAL, 6477 bytes)
INFO  Running task 0.0 in stage 0.0 (TID 0)
INFO  Code generated in 14.170306 ms
INFO  Finished task 0.0 in stage 0.0 (TID 0). 3228 bytes result sent to driver
INFO  Finished task 0.0 in stage 0.0 (TID 0) in 80 ms on localhost (1/1)
INFO  Removed TaskSet 0.0, whose tasks have all completed, from pool
INFO  ResultStage 0 (first at AgePredicterLocal.java:114) finished in 0.094 s
INFO  Job 0 finished: first at AgePredicterLocal.java:114, took 0.154587 s
Content-Length: 17
Content-Type: text/plain
Estimated-Author-Age: 32.29913797083779
X-Parsed-By: org.apache.tika.parser.CompositeParser
X-Parsed-By: org.apache.tika.parser.recognition.AgeRecogniser
resourceName: test.txt
INFO  Invoking stop() from shutdown hook
INFO  Stopped ServerConnector@71b3bc45{HTTP/1.1}{0.0.0.0:4040}
INFO  Stopped o.s.j.s.ServletContextHandler@6f2cb653{/stages/stage/kill,null,UNAVAILABLE}
INFO  Stopped o.s.j.s.ServletContextHandler@7a791b66{/api,null,UNAVAILABLE}
INFO  Stopped o.s.j.s.ServletContextHandler@40f1be1b{/,null,UNAVAILABLE}
INFO  Stopped o.s.j.s.ServletContextHandler@4cc6fa2a{/static,null,UNAVAILABLE}
INFO  Stopped o.s.j.s.ServletContextHandler@30d4b288{/executors/threadDump/json,null,UNAVAILABLE}
INFO  Stopped o.s.j.s.ServletContextHandler@51cd7ffc{/executors/threadDump,null,UNAVAILABLE}
INFO  Stopped o.s.j.s.ServletContextHandler@40499e4f{/executors/json,null,UNAVAILABLE}
INFO  Stopped o.s.j.s.ServletContextHandler@2fea7088{/executors,null,UNAVAILABLE}
INFO  Stopped o.s.j.s.ServletContextHandler@5c87bfe2{/environment/json,null,UNAVAILABLE}
INFO  Stopped o.s.j.s.ServletContextHandler@f73dcd6{/environment,null,UNAVAILABLE}
INFO  Stopped o.s.j.s.ServletContextHandler@6d304f9d{/storage/rdd/json,null,UNAVAILABLE}
INFO  Stopped o.s.j.s.ServletContextHandler@dcfda20{/storage/rdd,null,UNAVAILABLE}
INFO  Stopped o.s.j.s.ServletContextHandler@a4b2d8f{/storage/json,null,UNAVAILABLE}
INFO  Stopped o.s.j.s.ServletContextHandler@1fdfafd2{/storage,null,UNAVAILABLE}
INFO  Stopped o.s.j.s.ServletContextHandler@5b07730f{/stages/pool/json,null,UNAVAILABLE}
INFO  Stopped o.s.j.s.ServletContextHandler@58ebfd03{/stages/pool,null,UNAVAILABLE}
INFO  Stopped o.s.j.s.ServletContextHandler@1080b026{/stages/stage/json,null,UNAVAILABLE}
INFO  Stopped o.s.j.s.ServletContextHandler@5b43fbf6{/stages/stage,null,UNAVAILABLE}
INFO  Stopped o.s.j.s.ServletContextHandler@4678a2eb{/stages/json,null,UNAVAILABLE}
INFO  Stopped o.s.j.s.ServletContextHandler@41813449{/stages,null,UNAVAILABLE}
INFO  Stopped o.s.j.s.ServletContextHandler@354fc8f0{/jobs/job/json,null,UNAVAILABLE}
```

```
INFO  Stopped o.s.j.s.ServletContextHandler@47db5fa5{/jobs/job,null,UNAVAILABLE}
INFO  Stopped o.s.j.s.ServletContextHandler@4879f0f2{/jobs/json,null,UNAVAILABLE}
INFO  Stopped o.s.j.s.ServletContextHandler@12bcd0c0{/jobs,null,UNAVAILABLE}
INFO  Stopped Spark web UI at http://192.168.1.65:4040
INFO  MapOutputTrackerMasterEndpoint stopped!
INFO  MemoryStore cleared
INFO  BlockManager stopped
INFO  BlockManagerMaster stopped
INFO  OutputCommitCoordinator stopped!
INFO  Successfully stopped SparkContext
INFO  Shutdown hook called
INFO  Deleting directory /private/var/folders/n5/1d_k3z4s2293q8ntx_n8sw54mm5n_8/T/spark-fd116873-eec8-437d-9ad0-
7b7a09889d92
$
```