1. ANACONDA

folder: c:\anaconda

1. Download and install from:
    - https://www.anaconda.com/download

2. SPARK NLP

2.1 Java SE Development Kit (JDK) 8 (*required for Spark NLP models)

folder: c:\java\jdk, c:\java\jre

1. Download and install JDK 8 (jdk-8u333-windows-x64.exe) from:
    - https://www.oracle.com/java/technologies/downloads/#java8-windows
    - *Not to be confused with Java SE Runtime Environment (JRE) 8

2. Install JDK 8 in the Java home directory: *C:\Java\jdk*

3. Under Environment Variables, create JAVA_HOME path to *C:\Java\jdk*.

4. Add *%JAVA_HOME%\bin* to PATH

Resource: https://www.codejava.net/java-core/how-to-set-java-home-environment-variable-on-windows-10

2.2 Python

folder: c:\python\python37

1. Download Python 3.7.2:
    - https://www.python.org/downloads/release/python-372/
    - Note. This version is needed to run current Spark NLP models developed on v3.4.2

2. Select custom installation to install at *C:\Python\Python37*, and add to PATH

3. Under Environment Variables, create PYTHON_HOME path to *C:\Python\Python37*

4. Create python3.exe
    a. Go to C:\Python\Python37
    b. Copy python.exe and rename the copy as python3.exe (https://superuser.com/questions/1576758/how-do-i-alias-python3-on-windows)

2.3 Apache Spark (*required for Spark NLP models)

folder: c:\spark\ *spark-3.1.2-bin-hadoop2.7*

1. Download Spark 3.1.2 and Apache Hadoop 2.7:

- https://spark.apache.org/downloads.html

2. Create the Spark directory: *C:\Spark*. Move spark-3.1.2-bin-hadoop2.7.tgz file to the Spark directory and unzip to get a spark-3.1.2-bin-hadoop2.7 folder.

3. Under Environment Variables, create SPARK_HOME path to *C:\Spark\spark-3.1.2-bin-hadoop2.7*

4. Download winutils.exe:
   - https://github.com/cdarlint/winutils/blob/master/hadoop-2.7.7/bin/winutils.exe

5. Create the Hadoop directory: *C:\Hadoop\bin*. Move winutils.exe into the bin folder.

6. Under Environment Variables, create HADOOP_HOME path to *C:\Hadoop*.

7. Add *%SPARK_HOME%\bin* and %HADOOP_HOME%\bin to PATH

8. To verify successful installation of Spark and PySpark, open CMD and type:
   a. spark-shell
      *Output (Ignore warning messages):

   b. CRTL + D (exit scala)
   c. pyspark
      *Output (Ignore warning messages):

9. You may view your Spark environment by going to:
   - http://localhost:4040/environment/

10. Only if you encounter "Exception: Java gateway process exited before sending the driver its port number" error when running Spark NLP in Jupyter Notebook:
    a. Under Environment Variables, create PYSPARK_DRIVER_PYTHON path to *jupyter* and PYSPARK_DRIVER_PYTHON_OPTS path to *notebook*.

2.4 Virtual Environment

2.4.1 create conda env: sparknlp

1. Open Anaconda Prompt and create a new Python 3.7.2 environment for Spark NLP:
    a. conda create -n sparknlp python==3.7.2
    b. conda activate sparknlp

2. Install Jupyter:
    - pip install jupyter

3. Install PySpark:
    - pip install pyspark==3.1.2

4. Install Spark NLP:
    - pip install spark-nlp==3.4.2 (*this is the JSL_VERSION)

5. Install Spark NLP JSL:
    - pip install --upgrade spark-nlp-jsl==3.4.2 --user --extra-index-url https://pypi.johnsnowlabs.com/3.4.2-SECRET
    - Note. if you already have sparknlp installed and want to upgrade to new version, just run step 5 (no need step 4). It will upgrade your spark-nlp to the new version as well. Get the JSL_VERSION and SECRET in the license json file.

6. Install Spark NLP Display:
    - pip install spark-nlp-display

7. Install Matplotlib:
    - pip install matplotlib

8. Install numpy pandas sklearn
    - pip install numpy pandas sklearn

2.4.2 Test Spark NLP

*This is important to verify that required jar files are downloaded successfully

```
(sparknlp) D:\>python
Python 3.7.2 (default, Feb 21 2019, 17:35:59) [MSC v.1915 64 bit (AMD64)] :: Anaconda, Inc. on win32
Type "help", "copyright", "credits" or "license" for more information.
>>> import sparknlp
>>> spark=sparknlp.start()
```

Output:

```
:: loading settings :: url = jar:file:/C:/Spark/spark-3.1.2-bin-hadoop2.7/jars/ivy-2.4.0.jar!/org/apache/ivy/core/settings/ivysettings.xml
Ivy Default Cache set to: C:\Users\xxx\.ivy2\cache
```

```
The jars for the packages stored in: C:\Users\xxx\.ivy2\jars
com.johnsnowlabs.nlp#spark-nlp_2.12 added as a dependency
:: resolving dependencies :: org.apache.spark#spark-submit-parent-8d7ef308-70e5-4492-8e93-b8904b2af7bf;1.0
	confs: [default]
	found com.johnsnowlabs.nlp#spark-nlp_2.12;3.4.2 in central
	found com.typesafe#config;1.4.1 in central
	found org.rocksdb#rocksdbjni;6.5.3 in central
	found com.amazonaws#aws-java-sdk-bundle;1.11.603 in central
	found com.github.universal-automata#liblevenshtein;3.0.0 in central
	found com.google.code.findbugs#annotations;3.0.1 in central
	found net.jcip#jcip-annotations;1.0 in central
	found com.google.code.findbugs#jsr305;3.0.1 in central
	found com.google.protobuf#protobuf-java-util;3.0.0-beta-3 in central
	found com.google.protobuf#protobuf-java;3.0.0-beta-3 in central
	found com.google.code.gson#gson;2.3 in central
	found it.unimi.dsi#fastutil;7.0.12 in central
	found org.projectlombok#lombok;1.16.8 in central
	found org.slf4j#slf4j-api;1.7.21 in central
	found com.navigamez#greex;1.0 in central
	found dk.brics.automaton#automaton;1.11-8 in central
	found org.json4s#json4s-ext_2.12;3.5.3 in central
	found joda-time#joda-time;2.9.5 in central
	found org.joda#joda-convert;1.8.1 in central
	found com.johnsnowlabs.nlp#tensorflow-cpu_2.12;0.3.2 in central
	found net.sf.trove4j#trove4j;3.0.3 in central
downloading https://repo1.maven.org/maven2/com/johnsnowlabs/nlp/spark-nlp_2.12/3.4.2/spark-nlp_2.12-3.4.2.jar ...
	[SUCCESSFUL ] com.johnsnowlabs.nlp#spark-nlp_2.12;3.4.2!spark-nlp_2.12.jar (235594ms)
downloading https://repo1.maven.org/maven2/com/typesafe/config/1.4.1/config-1.4.1.jar ...
	[SUCCESSFUL ] com.typesafe#config;1.4.1!config.jar(bundle) (1435ms)
downloading https://repo1.maven.org/maven2/org/rocksdb/rocksdbjni/6.5.3/rocksdbjni-6.5.3.jar ...
	[SUCCESSFUL ] org.rocksdb#rocksdbjni;6.5.3!rocksdbjni.jar (304899ms)
downloading https://repo1.maven.org/maven2/com/amazonaws/aws-java-sdk-bundle/1.11.603/aws-java-sdk-bundle-1.11.603.jar ...
....
        ---------------------------------------------------------------------
        |                  |            modules            ||   artifacts   |
        |       conf       | number| search|dwnlded|evicted|| number|dwnlded|
        ---------------------------------------------------------------------
        |      default     |   21  |   0   |   0   |   0   ||   21  |   0   |
        ---------------------------------------------------------------------
```

*Ensure that the download for 21 jar files are successful. Otherwise, re-check your installation steps, remove the jars folder and redo this test.

2.4.3 Install TensorFlow 1.15 (*for training of NER model on v3.4.2)

1. Install Tensorflow (tf) 1.15 in sparknlp environment:
    - pip install tensorflow==1.15

2. To verify successful installation, type:
    import tensorflow as tf
    *If an error is obtained, you may require to downgrade to numpy==1.19.5