Installation Guide

Automatic Audio Labeler for Speaker Verification (A2LSV)

Project Owner

Ali AGDENIZ

Contents

1. Introducing A2LSV Installation Guide

This guide explains how to plan the deployment of A2LSV system, how to set it up and make it operational.

1.1. System requirements for A2LSV

This section lists the requirements that must be fulfilled by the target system in order to successfully install and run A2LSV.

Platforms	Ubuntu 18.04 LTS and other Ubuntu releases All Debian Distros
Processors (CPUs)	Multi-core x64 compatible processors
Memory	8 GB minimum (depending on data volumes, more may be required)
Disk space	500 MB total required to install. A2LSV stores dataset files in local disk. The disk space requirements for A2LSV are directly related to the amount of data being labeled and stored. Recommended disk space is 200 GB.
Security	Django Auth and CSRF
Web Framework	Django
Python	Version 3.6
Messaging system	Apache Kafka
Database	MongoDB

2. Installing the Prerequisites

A2LSV requires several tools to be installed on the computer. These tools include the Kafka server, mongoDb server, ffmpeg and several python libraries. These packages can be installed from the software repositories of your Linux distribution.

2.1. Installing for Kafka

2.1.1. Prerequisites for Kafka

- One Ubuntu 18.04 server or any debian distro and a non-root user with sudo privileges.
- At least 4GB of RAM on the server. Installations without this amount of RAM may cause the Kafka service to fail, with the Java virtual machine (JVM) throwing an "Out Of Memory" exception during startup.
- OpenJDK 8 must be installed on your server. To install this version, follow these instructions on installing specific versions of OpenJDK.

2.1.2. Creating a User for Kafka

Logged in as your non-root sudo user, create a user called kafka with the useradd command:

```
sudo useradd kafka –m
```

Set the password using passwd:

```
sudo passwd Kafka
```

Add the kafka user to the sudo group with the adduser command, so that it has the privileges required to install Kafka's dependencies:

```
sudo adduser kafka sudo
```

Log into this account using su:

```
su -l kafka
```

2.1.3. Downloading and Extracting the Kafka Binaries

Create a directory in /home/kafka called Downloads to store your downloads:

```
mkdir ~/Downloads
```

Use curl to download the Kafka binaries:

```
curl "https://www.apache.org/dist/kafka/2.1.1/kafka_2.11-2.1.1.tgz" -o ~/Downloads/kafka.tgz
```

Create a directory called kafka and change to this directory. This will be the base directory of the Kafka installation:

```
mkdir ~/kafka && cd ~/Kafka
```

Extract the archive you downloaded using the tar command:

```
tar -xvzf ~/Downloads/kafka.tgz --strip 1
```

2.1.4.Configuring the Kafka Server

Kafka's default behavior will not allow us to delete a topic, the category, group, or feed name to which messages can be published. To modify this, let's edit the configuration file.

Kafka's configuration options are specified in server.properties. Open this file with nano or your favorite editor:

```
nano ~/kafka/config/server.properties
```

Then add a setting that will allow us to delete Kafka topics. Add the following to the bottom of the file:

```
delete.topic.enable = true
```

Save the file, and exit nano.

2.1.5.Creating Systemd Unit Files and Starting the Kafka Server

You must create systemd unit files for the Kafka service. This will help you to perform common service actions such as starting, stopping, and restarting Kafka in a manner consistent with other Linux services.

Create the unit file for zookeeper:

```
sudo nano /etc/systemd/system/zookeeper.service
```

Enter the following unit definition into the file:

```
[Unit]
Requires=network.target remote-fs.target
After=network.target remote-fs.target

[Service]
Type=simple
User=kafka
ExecStart=/home/kafka/kafka/bin/zookeeper-server-start.sh
/home/kafka/kafka/config/zookeeper.properties
ExecStop=/home/kafka/kafka/bin/zookeeper-server-stop.sh
Restart=on-abnormal

[Install]
WantedBy=multi-user.target
```

Next, create the systemd service file for kafka:

```
sudo nano /etc/systemd/system/kafka.service
```

Enter the following unit definition into the file:

```
[Unit]
Requires=zookeeper.service
After=zookeeper.service

[Service]
Type=simple
User=kafka
ExecStart=/bin/sh -c '/home/kafka/kafka/bin/kafka-server-start.sh
/home/kafka/kafka/config/server.properties > /home/kafka/kafka/kafka.log 2>&1'
ExecStop=/home/kafka/kafka/bin/kafka-server-stop.sh
Restart=on-abnormal

[Install]
WantedBy=multi-user.target
```

Now that the units have been defined, start Kafka with the following command:

```
sudo systemctl start Kafka
```

To ensure that the server has started successfully, check the journal logs for the kafka unit:

```
sudo journalctl -u Kafka
```

You should see output similar to the following:

```
Jun 24 18:38:59 kafka-ubuntu systemd[1]: Started kafka.service.
```

You now have a Kafka server listening on port 9092. To enable kafka on server boot, run:

```
sudo systemctl enable kafka
```

2.1.6.Restricting the Kafka User

Remove the kafka user from the sudo group:

```
sudo deluser kafka sudo
```

To further improve your Kafka server's security, lock the kafka user's password using the passwd command. This makes sure that nobody can directly log into the server using this account:

```
sudo passwd kafka -l
```

At this point, only root or a sudo user can log in as kafka by typing in the following command:

```
sudo su - kafka
```

2.2. Installing mongoDb

To install mondoDb on debian distros, please follow below official installation guide url link.

https://docs.mongodb.com/manual/tutorial/install-mongodb-on-debian/

After installing and configuring mongoDb, you must be sure that mongoDb server is running.

2.3. Installing ffmpeg

Start by updating the packages list:

```
sudo apt update
```

Next, install FFmpeg by typing the following command:

```
sudo apt install ffmpeg
```

To validate that the package is installed properly use the ffmpeg -version command which prints the FFmpeg version:

```
ffmpeg –version
```

The output should look something like this:

```
ffmpeg version 3.4.4-0ubuntu0.18.04.1 Copyright (c) 2000-2018 the FFmpeg developers
built with gcc 7 (Ubuntu 7.3.0-16ubuntu3)
```

2.4. Installing python3.6

To install python 3.6, please follow the below url page instructions.

https://wiki.python.org/moin/BeginnersGuide/Download

2.5. Installing pip

To install pip, please follow the below url page instructions.

https://pip.pypa.io/en/stable/installing/

2.6. Installing pipenv

To install pipenv, please run the the below command.

```
pip install pipenv
```

2.7. Create virtual environment

After extracting project files, open terminal, go to project path and run below command:

```
pipenv --python 3.6
```

Output of command:

```
(base) ali@ubuntu:~/Desktop/a2lsv$ pipenv --python 3.6
Creating a virtualenv for this project…
Pipfile: /home/ali/Desktop/a2lsv/Pipfile
Using /home/ali/anaconda3/bin/python3 (3.6.9) to create virtualenv…
⠸ Creating virtual environment...created virtual environment CPython3.6.9.final.0-64 in 466ms
  creator CPython3Posix(dest=/home/ali/.local/share/virtualenvs/a2lsv-QXf0Rz4O, clear=False, global=False)
  seeder FromAppData(download=False, pip=latest, setuptools=latest, wheel=latest, via=copy, app_data_dir=/home/ali/.local/share/virtualenv/seed-app-data/v1.0.1)
  activators BashActivator,CShellActivator,FishActivator,PowerShellActivator,PythonActivator,XonshActivator

✓ Successfully created virtual environment!
Virtualenv location: /home/ali/.local/share/virtualenvs/a2lsv-QXf0Rz4O
```

2.8. Activate virtual environment

In the project path, run below command:

```
pipenv shell
```

Output of command:

```
(base) ali@ubuntu:~/Desktop/a2lsv$ pipenv shell
Launching subshell in virtual environment…
 . /home/ali/.local/share/virtualenvs/a2lsv-QXf0Rz4O/bin/activate
```

2.9. Installing all required python packages

After extracting project files, open terminal, go to project path and run below command. This command will install django and all other required python packages.

```
pip install -r requirements.txt
```

Output of command:

```
(a2lsv) (base) ali@ubuntu:~/Desktop/a2lsv$ pip install -r requirements.txt
Processing /home/ali/.cache/pip/wheels/cd/fa/73/31558700376872867f3fa39ce740d42bc3babc913e31b6566e/djongo-1.3.1-py3-none-any.whl
Collecting numpy>=1.15.1
  Using cached numpy-1.19.0-cp36-cp36m-manylinux2010_x86_64.whl (14.6 MB)
Collecting scipy>=1.1.0
  Using cached scipy-1.5.0-cp36-cp36m-manylinux1_x86_64.whl (25.9 MB)
Collecting torch>=0.4.0
  Using cached torch-1.5.1-cp36-cp36m-manylinux1_x86_64.whl (753.2 MB)
Processing /home/ali/.cache/pip/wheels/a0/92/d5/6acb411afb95b4196e3782d32bf2a813524a4cb09623bb6826/umap_learn-0.4.4-py3-none-any.whl
Collecting matplotlib>=2.0.2
  Using cached matplotlib-3.2.2-cp36-cp36m-manylinux1_x86_64.whl (12.4 MB)
Processing /home/ali/.cache/pip/wheels/ba/22/1c/d4e9707bbb27d469c384efc4263d8c7125219c1f088937289c/webrtcvad-2.0.10-cp36-cp36m-linux_x86_64.whl
Processing /home/ali/.cache/pip/wheels/cb/1d/15/a479fa740849128d481333d2f354f97691be3e2c82480a3e00/librosa-0.7.2-py3-none-any.whl
Collecting tqdm
  Using cached tqdm-4.46.1-py2.py3-none-any.whl (63 kB)
Collecting sounddevice
  Using cached sounddevice-0.3.15-py2.py3-none-any.whl (30 kB)
Collecting google-api-python-client
  Using cached google_api_python_client-1.9.3-py3-none-any.whl (59 kB)
Collecting django
  Using cached Django-3.0.7-py3-none-any.whl (7.5 MB)
Collecting django-crispy-forms
  Using cached django_crispy_forms-1.9.1-py2.py3-none-any.whl (108 kB)
Collecting django-extensions
  Using cached django_extensions-2.2.9-py2.py3-none-any.whl (217 kB)
Collecting kafka-python
  Using cached kafka_python-2.0.1-py2.py3-none-any.whl (232 kB)
Collecting pydub
  Using cached pydub-0.24.1-py2.py3-none-any.whl (30 kB)
Collecting numba==0.48
  Using cached numba-0.48.0-cp36-cp36m-manylinux1_x86_64.whl (2.5 MB)
Collecting dataclasses>=0.1
  Using cached dataclasses-0.7-py3-none-any.whl (18 kB)
Processing /home/ali/.cache/pip/wheels/6e/2d/16/a19e18f0e795af35a27ce8415bc29ac9a495cc87ebf57cce5b/bson-0.5.8-py3-none-any.whl
```

2.10. Configuring “configs.json”

All configurations related to A2lSV system is in “configs.json” file. You must have a valid Youtube Data API developer key. Default values for kafka port and mongoDb address are below. Change them if you need.

```
{
        "kafkaPort": 9092,
        "mongoDbAddress" : "127.0.0.1:27017",
        "googleAPIDeveloperKey" : "laskjdalskjdblashfhbalksjfnalksjf"
}
```

3. Django Deployment

To deploy web server, firstly you need to make migrations. Change working directory to a2lsv_web and run below commands:

```
python manage.py makemigrations web_interface
```

```
python manage.py migrate
```

```
(a2lsv) (base) ali@ubuntu:~/Desktop/a2lsv/a2lsv_web$ python manage.py makemigrations
 web_interface
Migrations for 'web_interface':
  web_interface/migrations/0001_initial.py
    - Create model Languages
    - Create model Datasets
    - Create model User
(a2lsv) (base) ali@ubuntu:~/Desktop/a2lsv/a2lsv_web$ python manage.py migrate
Operations to perform:
  Apply all migrations: admin, auth, contenttypes, sessions, web_interface
Running migrations:
  Applying contenttypes.0001_initial... OK
  Applying contenttypes.0002_remove_content_type_name... OK
  Applying auth.0001_initial... OK
  Applying auth.0002_alter_permission_name_max_length... OK
  Applying auth.0003_alter_user_email_max_length... OK
  Applying auth.0004_alter_user_username_opts... OK
  Applying auth.0005_alter_user_last_login_null... OK
  Applying auth.0006_require_contenttypes_0002... OK
  Applying auth.0007_alter_validators_add_error_messages... OK
  Applying auth.0008_alter_user_username_max_length... OK
  Applying auth.0009_alter_user_last_name_max_length... OK
  Applying auth.0010_alter_group_name_max_length... OK
  Applying auth.0011_update_proxy_permissions... OK
  Applying web_interface.0001_initial... OK
  Applying admin.0001_initial... OK
  Applying admin.0002_logentry_remove_auto_add... OK
  Applying admin.0003_logentry_add_action_flag_choices... OK
  Applying sessions.0001_initial... OK
```

If you want to insert some language record, you can run below command:

```
python manage.py loaddata fixtures.json
```

```
^C(a2lsv) (base) ali@ubuntu:~/Desktop/a2lsv/a2lsv_web$ python manage.py loaddata fix
res.json
Installed 2 object(s) from 1 fixture(s)
```

To start server, run below command:

```
python manage.py runserver
```

```
(a2lsv) (base) ali@ubuntu:~/Desktop/a2lsv/a2lsv_web$ python manage.py runserver
Watching for file changes with StatReloader
Performing system checks...

System check identified no issues (0 silenced).
June 25, 2020 - 13:00:57
Django version 3.0.7, using settings 'a2lsv_web.settings'
Starting development server at http://127.0.0.1:8000/
Quit the server with CONTROL-C.
[25/Jun/2020 13:08:27] "GET / HTTP/1.1" 200 2775
[25/Jun/2020 13:08:30] "GET /accounts/signup/manager/ HTTP/1.1" 200 4381
[25/Jun/2020 13:08:33] "GET / HTTP/1.1" 200 2775
```

4. Starting Kafka Consumers and Producers

4.1. Activating virtual environment

You must activate virtual environment to start all kafka consumers and producers. To activate, change dir to project folder and run below command:

```
pipenv shell
```

```
(base) ali@ubuntu:~/Desktop/a2lsv$ pipenv shell
Launching subshell in virtual environment…
 . /home/ali/.local/share/virtualenvs/a2lsv-QXf0Rz4O/bin/activate
(base) ali@ubuntu:~/Desktop/a2lsv$  . /home/ali/.local/share/virtualenvs/a2lsv-QXf0R
z4O/bin/activate
```

4.2. Starting Kafka Consumers and Producers

4.2.1. Start “youtubeSearch.py”

To start “youtubeSearch.py” consumer producer, activate environment and run below command:

```
python youtubeSearch.py
```

```
(a2lsv) (base) ali@ubuntu:~/Desktop/a2lsv$ python youtubeSearch.py
\Running Consumer..
request got: sample_dataset, tr, deneme
{'videoId': 'PkDVxX-qNx8', 'channelId': 'UCRBsX5EjfkEK-6TOJ_otlhA', 'titles': 'ONLİN
E DENEME SINAVI', 'channelTitles': 'osmanhocaylahazırlanıyorum', 'language': 'tr', '
downloaded': False, 'speakersDiarized': False, 'labelsConfirmed': False, 'speakersAd
ded': False, '_id': ObjectId('5ef30d605f627d48c7a87857')}
Message published successfully.
{'videoId': 'XAO804_DUfc', 'channelId': 'UCRBsX5EjfkEK-6TOJ_otlhA', 'titles': 'Onlin
e Deneme Sınavı-1', 'channelTitles': 'osmanhocaylahazırlanıyorum', 'language': 'tr',
 'downloaded': False, 'speakersDiarized': False, 'labelsConfirmed': False, 'speakers
Added': False, '_id': ObjectId('5ef30d605f627d48c7a87858')}
Message published successfully.
{'videoId': 'w6ZFx65qb8Q', 'channelId': 'UCrIj8ABnTsRoPIWUpavPIjg', 'titles': 'Makin
a nasıl denenir JCB SÜPER 3 CX ALIM İÇİN  DENEME', 'channelTitles': 'İŞ Makineleri K
EREM USTA', 'language': 'tr', 'downloaded': False, 'speakersDiarized': False, 'label
sConfirmed': False, 'speakersAdded': False, '_id': ObjectId('5ef30d605f627d48c7a8785
9')}
Message published successfully.
{'videoId': '5Ef4-hWhYlQ', 'channelId': 'UCPdN4Vx2DogwmjDJCkyVcXQ', 'titles': 'Yeni
LGS Kitapları ÇIKTI!', 'channelTitles': 'Tonguç 8.Sınıf', 'language': 'tr', 'downloa
ded': False, 'speakersDiarized': False, 'labelsConfirmed': False, 'speakersAdded': F
alse, '_id': ObjectId('5ef30d605f627d48c7a8785a')}
Message published successfully.
{'videoId': 'tkxoXTUSeh8', 'channelId': 'UCnu3buA0DHNXsTizUxx84Fw', 'titles': 'DENEM
E SINAVLARI NEDEN ÖNEMLİ?', 'channelTitles': 'Pervin Kaplan', 'language': 'tr', 'dow
nloaded': False, 'speakersDiarized': False, 'labelsConfirmed': False, 'speakersAdded
': False, '_id': ObjectId('5ef30d605f627d48c7a8785b')}
Message published successfully.
{'videoId': '2PTwlQNTTFM', 'channelId': 'UCRsazqgoMv4H4KSxY0KFvpg', 'titles': 'Telsi
zi eve kurdum dinleme ve gönderme deneme', 'channelTitles': 'huseyin 16 kartal', 'la
nguage': 'tr', 'downloaded': False, 'speakersDiarized': False, 'labelsConfirmed': Fa
lse, 'speakersAdded': False, '_id': ObjectId('5ef30d605f627d48c7a8785c')}
Message published successfully.
```

4.2.2.Start “youtubeAudioDownloader.py”

To start “youtubeAudioDownloader.py” consumer producer, activate environment and run below command:

```
python  youtubeAudioDownloader.py
```

```
(a2lsv) (base) ali@ubuntu:~/Desktop/a2lsv$ python youtubeAudioDownloader.py
Running Consumer..
got request: Vb9ox7zVAIg tr_dataset
video: Vb9ox7zVAIg attempt: 0
video: Vb9ox7zVAIg downloaded.
message sent: Vb9ox7zVAIg tr_dataset m4a
Message published successfully.
got request: cXq5kF7sYz4 tr_dataset
video: cXq5kF7sYz4 attempt: 0
video: cXq5kF7sYz4 downloaded.
message sent: cXq5kF7sYz4 tr_dataset m4a
Message published successfully.
got request: fPrqoKTnd4E tr_dataset
video: fPrqoKTnd4E attempt: 0
video: fPrqoKTnd4E downloaded.
message sent: fPrqoKTnd4E tr_dataset m4a
Message published successfully.
```

4.2.3.Start “speakerDiarization.py”

To start “youtubeAudioDownloader.py” consumer, activate environment and run below command:

```
python speakerDiarization.py
```

```
(a2lsv) (base) ali@ubuntu:~/Desktop/a2lsv$ python speakerDiarization.py
Running Consumer..
loading models ...
Loaded encoder "encoder/saved_models/pretrained.pt" trained to step 1564501
models loaded.
got request: ['Vb9ox7zVAIg', 'tr_dataset', 'm4a']
speakers are being diarized; videoId: Vb9ox7zVAIg
speaker diarization done; videoId:Vb9ox7zVAIg
got request: ['cXq5kF7sYz4', 'tr_dataset', 'm4a']
speakers are being diarized; videoId: cXq5kF7sYz4
speaker diarization done; videoId:cXq5kF7sYz4
got request: ['fPrqoKTnd4E', 'tr_dataset', 'm4a']
speakers are being diarized; videoId: fPrqoKTnd4E
```

5. Accessing final dataset files

You can find final dataset files in “a2lsv_web/static/datasets/(dataset_name)/final_dataset” directory. Here is a example screenshot.

Folder hierarchy is like speaker id => youtube video id => audio file.