Group W16: Speech and Speaker Recognition

Risto Rahulan

Frederik Raud

A Report

Business Understanding

Background: Our instructors, the lecturers and teaching assistants of the course Introduction to Data Science at the University of Tartu have tasked us with a data mining project to demonstrate that we can apply the knowledge gained at the course in practice. The completion of the course is required to pass the course. This project has two goals: grading and learning.

The first goal is one of grading. The project must follow a formal pattern that allows the instructors to grade the students with the end results of their students getting a grade on the course. This is necessary for the university, so that they can accredit the student with a degree. The requirements are strict standardized, so that looking at the reports, individuals can assess the credibility of the degree by the credibility of the university.

The second goal is one of learning. The Republic of Estonia is interested in specialist, so the intended aim of learning is not science- but business-oriented.

The students have chosen a speech categorization as the topic of the project. They will try to separate audio-clips by genders and accents. The project can be graded based on the poster and repository they create.

Business goals: The goal of this project is for the students to study the topic, put their knowledge into practice, and show their work on speech categorization in order that they may be graded.

The goal of the project to be graded is to produce an algorithm that can study speech and generalize based on that study the genders and accents of the speakers and to do this based on what has been taught on the course. The students must create a poster and a repository.

Business success criteria: The project is successful if it is submitted before the deadline and if the students are able to show their work in order to get a passing grade. Since self-improvement is subjective, there is no criteria for success.

The project is successful if it can categorize speakers better than a random number generator could. Preferably, the difference should be notable, as it is backed up by what has been taught on the course. By the end of this the students must have created a meaningful poster and a project repository.

Inventory of resources: In order to accomplish the project, the students will utilize the following resources: The Mozilla Voice Dataset (30GB), The Audio Feature Extractor of LibROSA, a processor running with a 2.80GHz frequency on 6 logical cores, at least 100GB of free space.

Requirements, assumptions, constraints: The project is time sensitive and must be finished by the 16th of December, 2019. The project must follow the set out requirements, so that it can be graded. The project must be done by two people, as including any more people on the project would be academic fraud.

The project assumes that the problem is solvable to a degree in the given time-frame and that the students already possess the knowledge to be the solvers. It also assumes that there even is a solution to the problem given the tools available. There assumption for the topic of study or the null-hypothesis is that the age, gender and accent of a speaker are related to the way they speak. We assume that if a human being is capable of perceiving distinct accents, so is a computer.

Risks and contingencies: Failing to finish the project by the 16th of December, 2019, results in failure, not following the set out requirements will result in a poorer grade or failure, committing academic fraud risks expulsion from the university or at least failure.

Terminology: audio, frequency, amplitude, time, spectrogram, zero crossing rate, spectral centroid, spectral roll-off, Mel-Frequency Cepstrum, speech, analysis, data mining, clustering, machine learning, categorization, generalization, poster, repository

Costs: By the end of the project, the students and instructors will have invested a significant amount of their time into completing a project. This takes the form of an opportunity costs: opportunity to complete other projects on other courses, opportunity to spend time on self-improvement, and the opportunity to live a life outside of being a student of Computer Science.

Benefits: By the end of the project, the students of Tartu University have proven that they can be useful in generating value by teaching machines to generalize. The students themselves will also learn to generalize on the knowledge they gained at the course.

Data-mining goals: The students will deliver a repository, containing the speech-categorization models and the results they have produced. They will also deliver a poster detailing their project.

Data-mining success criteria: The data-mining is successful if the instructors decided that it is successful. There is also a Kaggle competition[1] that can be used to decide the successfulness of the model in terms of accuracy.

Data Understanding

Data requirements outline: The data must take on the form of an audio-file that can be understood as a sum of frequencies with the amplitudes over a sequence of time and represent somebody speaking. The data must also be related to attribute regarding the speaker's gender, accent and possible age. For the sake of meta-data, it would be nice for the data to also be related to the data on which they are recorded.

Data availability verification: The data does exist as far as anything can be said to exist. *The Mozilla Voice Dataset*[2] and unless the 432 contributors on the git-hub repository of the project it is a part of[3] have been complete fabrications along with the 11,218 commits, this data can be said to exist. The dataset is open-source and therefore easily verifiable and free.

Data selection criteria: The actual audio-files are important only in-so-far as they can be translated to a usable result of features. In addition to the feature data we will produce, we will also worry the least about age and more about the accent and the gender of the speakers. We will also not have much use of the validation/in-validation rating, besides deciding whether we want to use that data or not.

Data description: The data consists of 1087 hours of speech produced by 39,577 voices of 29 different languages. In addition to audio, each voice is linked to their demographic data of age, sex and accent.

Data exploration: The data includes an id for each speaker. One speaker is responsible for multiple audio-files, to which it is related to by a file (path)name. Each audio-file is related to what the speaker said in that audio-file. The data includes how many people have validated or invalidated this correspondence of the audio-file to the sentence. In addition to that, each speaker comes with some demographic metadata, regarding sex (male or female), age (in bins of tens: twenties, thirties, forties etc.) and accent (for example with English, whether it is spoken by someone from England, Scotland, the US, New Zealand etc.). This indicated that the data

[1] https://www.kaggle.com/mozillaorg/common-voice
[2] https://voice.mozilla.org/en/datasets
[3] https://github.com/mozilla/voice-web

makes a hidden assumption that has been proven to be false about accents. There is no one-to-one relation between speakers and accents. Over the course of their life time, one speaker can produce various different accents.

Data quality: The data is not perfect. It's mass-validated and must necessarily only include the audio-file, the speaker, the sentence and the validation/invalidation counts of the correspondence. For over half the rows, age and gender are missing, and for about 2 thirds of the rows, the accent is missing. Since gender and accent are what we're categorizing, the amount of data is cut down in half.

Project Plan

1. Learning about modelling of audio-files and the features of audio-files

   Risto Rahulan (6 hours)

   Frederik Raud (6 hours)

   Dependent on: nothing, re-visited about evalutation
2. Data Selection

   Risto Rahulan (8 hours)

   Frederik Raud (8 hours)

   Dependent on: task 1 and evaluation
3. Data Modelling

   Risto Rahulan (8 hours)

   Frederik Raud (8 hours)

   Dependent on: task 1, task 2 and evaluation
4. Evaluation

   Risto Rahulan (5 hours)

   Frederik Raud (5 hours)

   Dependent on: task 3
5. Making the Poster

   Risto Rahulan (3 hours)

   Frederik Raud (3 hours)

   Dependent on: task 3 and task 4