**Interviewer 2:** [00:00:00] Okay.

**Interviewer 1:** [00:00:02] Okay, I would like to start by asking you some background questions about you. Could you tell us what is your current position at your job? What kind of job you're doing?

**Interviewee:** [00:00:13] At the moment I actually don't have a permanent job. I've been for the past couple of months.. I don't have a fixed job. Before that I was working at [Company Name: Removed for Anonymity] okay here in [City: Removed for Anonymity] doing a NLP related work.

**Interviewer 1:** [00:00:36] Okay, thank you. So could you tell us what is your overall work experience and your experience specifically in machine / deep learning systems?

**Interviewer 2:** [00:00:46] In terms of years.

**Interviewee:** [00:00:49] In terms of years... I've been doing machine learning for about three years.

**Interviewer 1:** [00:00:53] Okay, and your overall work experience, just for statistics.

**Interviewee:** [00:00:58] I've been a software engineer for over 20 years.

**Interviewer 1:** [00:01:02] Okay, thank you. And I wanted to ask what type of deep learning /machine learning networks have you developed implemented and by that I mean like supervised and unsupervised reinforcement learning and Etc.

When you work with text very often you work with unsupervised problems.

Okay. And you know, I've done also supervised things more as toy projects, but never I've never done a reinforcement learning so I don't have the experience to reinforcement learning.

Okay, and which programming languages and Frameworks have you been using in your experience?

So at the moment there is quite a bit.. But you know tensorflow, Keras, Scipy. Mostly the Python related tools for modeling itself. [Company Name: Removed for Anonymity] still uses a lot of C++ to do data processing. A lot of the data processing is still done with C++.

Okay, thank you. So for this interview we have one general question, which is what types of problems bark challenges you faced when developing deep learning machine learning [inaudible] and for your specifically we would also like to ask what kind of problems people ask often on Stackoverflow. If we could start from there, if you have anything on your mind, that would be very useful.

**Interviewee:** [00:02:46] Yeah, so I think the kind of questions on Stackoverflow are not [terribly, entirely?] representative of the problems you encounter on a machine learning project. So the kind of questions I tend to see on StackOverflow are more like.. Well first there's sort of beginner questions of people that don't really understand the documentation [inaudible] so you could say bad documentation is always a problem, but then you tend to see a lot of questions on how you develop custom metrics, for instanse, things that are not.. I don't know. I just don't think they are very representative of the day to day work in terms of developing a model. I mean, in terms of my experience, I don't know, the way I would put the challenges is first of all 80% of it is processing the data usually, you know, very often things are not like a Kaggle. In a Kegel competition you have a dataset that has been processed for you and in the real world you don't. You have, you know, either documents or a database that you have to scan and get the data from and usually there are errors in that data. So that's.. A lot of the ground work in a machine learning project just obtaining and cleaning up the data and it's usually tricky to make sure that you don't have bugs in your code as you're injecting the data and you're not dropping a lot of stuff that is useful or you are not doing stupid things when the value is missing. So that's usually time consuming and it's error-prone and I guess.. I don't know, the way I tend to do it.. A lot of times steps are error-prone and yeah, I guess model development itself is not deterministic. You don't know. You don't know which architectures are going to work or not. And so there's a lot of trial and error involved and sometimes some ideas work, sometimes they don't. You just have to spend a lot of time and [inaudible] with it.

I'm not sure if that's a good characterization.

**Interviewer 1:** [00:05:43] It definitely is, so maybe I could ask you a question about training dataset. So you said that there's a lot of ground work in collecting it. So did you work with existing datasets mostly or did you ever had to collect your own data? And if yes, what kind of problems you have faced in during that process?

**Interviewee:** [00:06:04] Mostly the later, at least four for commercial systems I had to collect data. A lot of the work I've been doing is based on web documents.

**Interviewer 1:** [00:06:17] Okay.

**Interviewee:** [00:06:18] And for instance parsing web documents is very time consuming and itself very.. To give you an to give you an example suppose that what you're trying to do is you want to classify news articles according to their topic, which is something else working on very often use articles have some ads on them and links to other news articles in the same property. So suppose you have a newspaper, there's an ad.. there are ads, they may mislead you if you just take the text of the ad and then you have the side links that point to a completely different topic. So you shouldn't really take the tokens from the side links. Because it totally confuses what is the talking about. So a problem like that. So scanning say CNN.com and figuring out which.. you know, extracting which tokens from the document are relevant and which things are noise can be your rather very time-consuming task and this just takes..

**Interviewer 1:** [00:07:38] Is it a part of cleaning data like removing these?

**Interviewee:** [00:07:42] Yeah, I mean you can consider it like that. It's a part of cleaning the data this case.. The web, sort of scraping data from the web is always very very noisy.

**Interviewer 1:** [00:07:57] Yeah, okay. And did you ever use existing data sets?

**Interviewee:** [00:08:06] Just mostly for toy projects. I mean it's sort of for training and try to play with ideas.. Yes, but for anything commercial - no.

**Interviewer 1:** [00:08:20] But I would think that on stackOverflow they do ask a lot about you know, when they work with existing datasets. Did you like notice any trend of the things that go wrong when people do that or maybe no?

**Interviewee:** [00:08:35] Right, so you're right. In StackOverflow it tend to have a lot of questions about people playing with the specific well-known dataset, because it's a good way to, you know, to try new algorithms. I don't know if there's a pattern there. I mean, sometimes people misunderstand the dataset and I guess it's always tricky.. Its error-prone typically to.. So even if you have a standard dataset, you often need to write some code that reads the dataset and and presents it to the library that you want to work with and there's often a lot of space for errors with things like.. With categorical features - how do you transform correctly, you know sort of token IDs into a categorical feature and yeah, it's sometimes easy to make mistakes on the code that reads the data set.

**Interviewer 1:** [00:09:49] Okay, so about model structure. Could you tell us about any problems related to the wrong model structure that you had in your experience like types of layers, numbers of layers and also the same about stackOverflow: do people confuse these things very often, do they pick wrong models - like the model itself the layer types itself..

**Interviewer 2:** [00:10:14] The number of nodes,also.

**Interviewee:** [00:10:17] So I mean, you have like the absolute beginners, but I don't think it's that representative. So if you look at it, there's sort of well-known techniques, so when you're dealing with images, you deal with one type of models, when you're dealing with audio and other type of models. So there's sort of, you know, sort of some 4-5 well-known starting points, depending on what kind of problem you are working with and let's assume that you'r practitioner that knows enough to start with the right starting point, right. Beginners may confuse it, so then once you once you start with something, that is a reasonable start, you mostly have an exploration space and you can do it manually, you can use library that explores.. So there are several libraries out there that you can say, well, run my model 10 times changing, you know, the number of layers or changing the number of layers and the learning rate. You can do that. Honestly, that's sort of like a final step for me because that takes like weeks typically because, you can imagine, for a non-trivial dataset each run can take days depending on what resources you have. So actually, you know, now I'm talking I'm talking to you, I was kind of reminded of the web article that I saw not so long ago with some tips on how to work on model development. I'll forward that to you after the call, now I'm not sure if I can if I can find it in a few seconds, but I'll forward that to you after the call and I think, that is not a hundred percent my experience, but I think it's a very well put-together web article. It at least characterizes some of the tasks you end up doing good with model development and so I'll send it to you afterwards.

**Interviewer 1:** [00:13:01] Yeah. Thanks a lot. Thank you. About hyper parameter tuning. So, -problems or bugs related to that in a way that: do people often confuse how to set it, is there any specific hyperparameter that you think is more challenging than the rest and etc.

**Interviewee:** [00:13:26] I don't have a good answer for you. I mean, typically, in my mind you tend to do a hyper parameter tuning when your model is already pretty decent and you're doing the final squeezing a few percent more and that is usually.. You know pretty well what you're tuning tuning. Its more of a question of working with the right library, so I don't think that's super challenging. Other than the time that it takes to..

**Interviewer 1:** [00:14:04] Okay, and about loss functions. Do you usually use a predefined loss function? Did you ever had a custom written ones, any problems related to the wrong use of loss function or it's wrong implementation?

**Interviewee:** [00:14:21] So you tend to see a lot of questions about loss of function on stackOverflow.I mean, typically, I start with a traditional loss function. They've been some..a bunch of.. Often what you do have is.. When you see these in some Kaggle competitions is sometimes you want sort of a generic concept of a loss, but you want the loss to be averaged around classes so suppose that you you have five classes and they're not balanced and what your objective is, is to have reasonable error rate among these classes. So that's the most common reason to do a custom loss function is to take mean squared error, but then sort of average the error by the number of samples per class. So you get a more balanced loss function than just something that can predict one class very well and then screws up the other ones. So I guess you know like...one of the things that Kerass is probably missing is, you know, a pre-done, a predefined loss of function that can automatically balance by classes. That would be something that.. If there was a prepackaged one, it would save people some time. But other than that, I don't tend to use a custom loss. There's been some some work on better loss functions, but I don't know, I always get the sense that maybe you could use a better loss function to squeeze the last few percent out of a problem, but that is never the starting point unless you're doing this balancing of classes. Right, so, you know to start with a custom loss function makes sense when you're balancing classes, but not otherwise.

**Interviewer 1:** [00:16:58] Okay. Thank you. And I have a question about hardware. So did you have any problems related to hardware in your experience? Where do you usually train your models?

**Interviewee:** [00:17:10] Yeah, so it depends on what I'm doing so like, first, for [Company Name: Removed for Anonymity] it's easy to get a lot of hardware resources, although you have to work with the [Company Name: Removed for Anonymity] infrastructure to do so, which has its own challenges. To do sort of toy projects Kaggle is a really good resource because they give you VMs with [inaudible: connection problems, a sentence is lost] ... I mean, the resources out of the Google Colab thing. So for small things where you are just playing or learning Kaggle and Google Colab give you resources for free and it's very easy to use.For industrial stuff that depends on which company.. And yeah, building the infrastructure for ML is not simple but hopefully somebody else is doing that for you.

**Interviewer 1:** [00:18:24] Okay. Do you remember like any error messages that you got because of the use of Hardware?

**Interviewee:** [00:18:31] So yeah, it's very common that, for instance, when you're using GPUs there are errors of synchronizing between, you know.. Say, the model and the GPU and sometimes some libraries don't do a good job of propagating that. So like, there's a couple couple of days ag I lost about two hours because I was running this very simple notebook one on a computer at home and it was giving me just complete junk results. And that was because somehow the CUDA operations on the GPU were just returning garbage results. There was an error, but that was not in the notebook, it was hidden somewhere else, so as you can imagine those things.. If something fails silently and it can easily.. You could be looking at the results in a notebook and there's some log message in some other place that can be very confusing sometimes.

**Interviewer 1:** [00:19:46] Okay, I would ask you one like very very general question, I guess. So, in any project that you were doing, let's say you had some training data you put on some module, you trained it, you got some results and then, I guess, you did some steps to improve the accuracy of your model. Could you just give us examples of the steps that you would usually take to ensure that your model has a better accuracy?

**Interviewee:** [00:20:13] So a lot of the work that I've been doing sort of professionally was a [Comnapy Name: Removed for Anonymity] with unsupervised learning and there it is tricky - you don't have a definition of accuracy if you're doing [inaudible]..

**Interviewer 1:** [00:20:33] Yeah, yeah.

**Interviewee:** [00:20:35] Okay, so I mean if you're doing a simple problem where you have, you know, a test set, you can save a test that you don't use for validation and you can just evaluate it against that test set. When it's unsupervised, the only thing you can do is try to manually compare whether the results of model A and model B, whether they look better to you or not, and that is pretty tricky and it typically involves getting humans to write then the results of the model and it's.. Yeah on unsupervised problems evolving the model can be pretty challenging.

**Interviewer 1:** [00:21:28] So, you said that you need to humans who will act as an oracle, as I understand you, to say what's right and what's wrong and my question would be do humans agree often on what is what or do they rather disagree?

**Interviewee:** [00:21:41] No, you have to do.. Usually you have to get multiple raters and then do a majority vote.

**Interviewer 1:** [00:21:48] Yeah.

**Interviewee:** [00:21:48] So to give you an example, suppose we're getting back to that news article problem that I was telling you about and whether a news article is for a topic or not. It's not a deterministic.. There is no deterministic answer, right? So you can ask three different humans. Do you like these? Do you like this answer more than you like the previous answer? Or is it worse than the previous answer? So that's how you compare models, you decide: well, you know, previously I was giving you these answers now I'm proposed to give you another set of answers, which ones are better. And if your users, fake users, which.. You have to give them a set of instructions. They tell you yeah, the new ones are better. It's never like all the new ones are bad [inaydible] a little bit. Some are better, some are worse. So you have to then build a metric based on that and that can be both expensive and very time consuming. So unsupervised problems have that issue.

**Interviewer 1:** [00:23:01] Okay, so I have one more question regarding stackoverflow. Could you like give us a very brief summary of the errors that people do like very often you think?

**Interviewee:** [00:23:17] Bugs people do very often.. Well, I don't know, about ninety percent of the questions are about people not having experience in being beginners in this space, right? Then the biggest error typically is, you know, when you deal with your input data and not testing your code. And I guess that the other pattern is with a lot of deep learning libraries, especially Keras, Tensorflow. There's this concept of an execution graph which is not the same thing as invoking functions one at a time and that is a very confusing concept for most people. So people people would expect, you know, they have an instruction, have a line and you execute that and git ives a result but when you're doing.. So a lot of the ML libraries don't do that, they build an execution graph and after the fact the graft runs and yeah, most people don't understand that. That's what's going on.

**Interviewer 1:** [00:24:50] Okay. Yeah, thank you. Yeah, well we're talking maybe you remembered about any other problems that you think are specific for this kind of system?

**Interviewee:** [00:25:01] No, not that I can think of.

**Interviewer 1:** [00:25:04] Okay, so Interviewer 2 if you don't have any questions..

**Interviewer 2:** [00:25:12] Maybe you can remember some more problems related to GPUs?

**Interviewee:** [00:25:23] Yeah, so like, for instance, in it's very easy.. In the same Hardware it's very easy for two different.. So you have two different notebooks, they interfere with each other, they can cause one to create errors for the other, so it's hard to do anything in parallel if you have GPUs so, that's a that's a very typical problem at least with the setup I have at home. If you do something in one notebook and then you change to some other task the two things will interfere and you just get errors and sometimes those errors are silent and it just means your result are invalid.

**Interviewer 2:** [00:26:17] Thank you. That's interesting.

**Interviewer 1:** [00:26:20] So, thanks a lot for your time.

**Interviewee:** [00:26:22] Welcome.

**Interviewer 2:** [00:26:23] Thank you.

**Interviewer 1:** [00:26:24] Bye.

**Interviewee:** [00:26:25] Thank you. Bye.