Task 4:

Natural Language Processing (Artificial Intelligence). An introduction to Natural Language Processing, which is one of the biggest current fields of research in Artificial Intelligence. First, let's get you started with some background knowledge. By the end of this task you'll be able to build your own AI program that can automatically classific tweets using the same technology that Google use to filter spam from your email inbox!

Introduction:

The image above is from the movie Ironman. In this movie, the fictional main character Tony Stark is a billionaire genius engineer. He creates a suit of armour that is so powerful and advanced that he becomes a superhero nicknamed 'Ironman'.

Tony has a robot AI inside his mansion called Jarvis. He talks to this robot many times during the film and it helps him perform certain tasks. The picture above is from a scene where Tony has just spoken to Jarvis. Let's think about this a bit..

What is the most advanced technology or crazy idea in the movie 'Ironman'?

- Is it the fact that Tony has built a suit made out of metal that he can fly around in?
- Is it the fact that the suit makes Tony so strong that he can shoot missiles from it, reflect bullets and fly around?
- Is it the fact that Tony is a billionaire engineering that is smart enough to do this by himself?
- Or is it the seemingly small and insignificant fact that Tony can talk to the Artificial Intelligence 'Jarvis' and Jarvis can understand exactly what he says?

If you didn't know better, and had no background in Artificial Intelligence, you may think that flying around in a suit shooting missiles is more advanced than a robot understanding the few simple things that Tony says to it. But you'd be wrong.

The fact that Jarvis understands Tony's simple words '*Wake up, daddy's home*' and can reply correctly with '*Welcome home, sir*' is a **massive technological feat** of artificial intelligence and problems of creating superhero suits and flying around in one is actually nothing compared to the huge field of **Natural Language Processing** which is the main area of research in the field of **Artificial Intelligence** today.

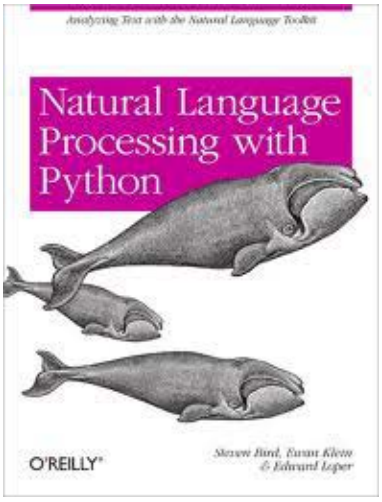
Even before Ironman there have been movies about space travel with people traveling on spaceships that have AI that can understand the crew speaking and reply to them. Today we have travelled space, gone to the moon, yet we have still failed to produce artificial intelligence systems that can do what is shown in these movies. How can this be the case? How can Natural Language Processing be harder than going to the moon! How can it still be a totally unsolved area in Artificial Intelligence?

How would we even start?

In order to start thinking about creating Jarvis in real life (ie a robot that we can talk to, can understand what we say and act on it or even just reply correctly) we'd need many things. We call programs like Jarvis that converse with humans in natural language **conversational agents** or dialogue systems. **Natural languages** are languages humans talk to each other in, **formal languages** are programming languages like Python or Java.

Jarvis must be able to recognize words from an audio signal and to generate an audio signal from a sequence of words. These tasks of **speech recognition** and **speech synthesis** require knowledge about **phonetics** and **phonology**: how words are pronounced in terms of sequences of sounds and how each of these sounds is created acoustically. Pronouncing variations of words correctly (such as plurals, contractions) requires knowledge about **morphology** – the way words break down into component parts that carry meanings .

What if we asked Jarvis '*How many University of KwaZulu-Natal university students are in the Math130 class by the end of the day?*'. Jarvis needs to know something about lexical semantics – the meaning of all the words (eg 'class' or 'students') and compositional semantics (what exactly makes a student a 'University of KwaZulu-Natal' student and not another type of student?). What does 'end' mean when combined with 'the day'? Jarvis needs to know about the relationship of the words to each other – how does Jarvis know that '*by the end of the day?*' refers to a time and doesn't refer to something like 'the book is written **by** the author JK Rowling'. Humans know this automatically, but how can computers learn this?



How does Jarvis know that when Tony says '*Daddy's home?*' Tony is actually talking about himself? Jarvis knows this because he says 'Welcome home sir' so clearly he understood that somehow. This knowledge about the king of actions that speakers intend by their use of sentences is **pragmatic or dialogue** knowledge. To summarise, Jarvis needs the following knowledge of language:

- Phonetics and Phonology – knowledge about linguistic sounds
- Morphology – knowledge of the meaningful components of words
- Syntax – knowledge of the structural relationships between words
- Semantics – knowledge of meaningful
- Pragmatics – knowledge of the relationship of meaning to the goals and intensions of the speaker
- Discourse – knowledge about linguistic units larger than a single utterance

Ambiguity:

What if Tony was telling Jarvis a story about his female assistant who had annoyed him. So Tony threw a piece of paper at her and she ducked to avoid it. Tony tells Jarvis the following sentence:

'*I made her duck*'

This simple sentence has the following meanings. The correct meaning in Tony's story is in bold:

- I cooked waterfowl for her.
- I cooked waterfowl belonging to her.
- I created the (plaster?) duck she owns
- **I caused her to quickly lower her head or body**
- I waved my magic wand and turned her into undifferentiated waterfowl

But how can we make Jarvis smart enough to know this? There are many **ambiguities** in this sentence because just the word 'duck' can be a **verb** (move your head down) or a **noun** (an animal). 'her' can mean the women or can refer to the fact that the duck **belongs** to her.

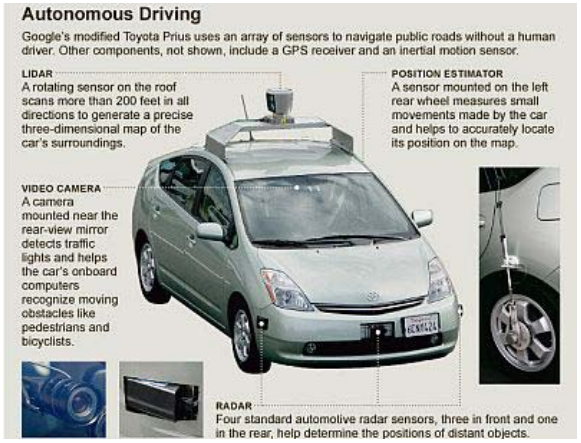
What about hearing? Say the word 'I' out loud. How does Jarvis know that this word isn't actually 'eye'. What about 'made', it sounds just like 'maid'! Poor Jarvis!

We must use complicated models and algorithms as ways to resolve or disambiguate (remove ambiguities) these ambiguities.

- **Part of speech tagging**: Deciding whether 'duck' is a verb or a noun is known as part of speech tagging. Verb and noun are each different 'parts of speech' and we 'tag' a word in a sentence by assigning it one part of speech that we think is correct for the context or sentence it has been used in. Probabilities are used to decide this sometimes, for example the word 'man' is more probable to mean the noun 'an older male person' than the verb 'put someone there' (eg 'Man the cannons!')

- **Word sense disambiguation**: Deciding whether 'make' means 'create' or 'cook'.

As you can see, Natural Language Processing (NLP) is a huge field and extremely important in the field of **Artificial Intelligence**. NLP also includes the study in things like Machine Translation which is the same as the technology behind **Google Translate** – automatically translating between two languages. **Google Search** also uses NLP techniques in order to understand your searches faster and this is what makes Google Search more accurate, reliable and faster than any other search method on the internet. NLP is used in **Gmail** in order to identify spam mail and delete it.



Natural Language Processing research is one of the main reasons Google is so successful. The probabilistic, machine learning techniques that Google apply to NLP can be applied to other Artificial Intelligence tasks such as creating driverless cars (currently in operation in California) and also surprisingly the field of Bioinformatics.

The task of tagging a sentence with the correct parts of speech for example:

The old man the boat

The : tag as 'determiner'
old: tag as 'adjective'
man: tag as 'verb'
the: tag as 'determiner'
boat: tag as 'noun'

Is very similar to the task of tagging a sequence of DNA with the correct amino acids when the DNA structure isn't always clear and the start and end points are muddled up:

AC?GTA??CATA?

DNA isn't always as clearly defined as it was given to you Task 3 and just like the ambiguity that arises in sentences, we must apply probabilities in algorithms to solve these problems.

I hope you can see how many fields in Artificial Intelligence and Bioinformaics have to do with probability. This is because it is the only way we can deal with ambiguity to try make robots/ AI programs to make the best decisions!

Machine Learning:

Natural Language Processing is a big part of Artificial Intelligence because a large amount of it has to do with training computers or AI programs to identify certain patterns and use probabilities to make good decisions. This is known as Machine Learning and is a massive field of research right now. Facebook uses Machine Learning to try recommend friends to you, Amazon uses it to recommend items to buy, Google Image search uses machine learning techniques to identify patterns in pixels to try find similar images, Google has designed driver less cars to act according to their environment by integrating many different machine learning techniques....the list is endless.

POS tagging:

One example is to solve the problem of part of speech (POS) tagging as explained above. We can give a program a big set of tagged words (training data) and then give it a new sentence it must try tag with the information learnt from the training data. The University of Pennsylvania had the first NLP research program to ever take a large corpus (body) of words and tag each and every one by hand (see **http://en.wikipedia.org/wiki/Treebank**), A program was then activated and slowly learnt how to tag words and the probability that a certain word appeared with a certain tag in a certain context. Ever since then, AI runs on huge sets of training data to be more accurate. In POS tagging we try to tag words with the correct part of speech tag so that we can then PARSE the sentence correctly.

PARSING is the formal term for 'putting a sentence together in the right way' so that it can be 'understood' . We will talk about this later.

Text classification:

We also use Machine learning to try classify a text. For example, gmail classifies emails as either 'spam' or 'not spam'. It uses machine learning by having trained an AI program on a set of already 'classified' emails (examples of non spam and spam emails). Then when the AI sees a new incoming email, it can use its prior knowledge or 'training' to classify the new email correctly.

This task will have an example of how we can use Machine Learning to get a program to identify positive or negative tweets – similar to the problem of identifying spam mail through machine learning.

Instructions:

In this task we will use the Natural Language Toolkit, which is an external Python module that must be downloaded and installed.

1. Open the folder 'Installers' inside this Task 4 folder.
   There is an exe file in this folder called 'NLTK installer.exe'. Run it and install.

   If you get an error 'Python cannot be found in the registry' , please email students@hyperiondev.com and we will email you a file you need to run to fix this error.
   You cannot continue with the other installations or this task if you get this error and we need to fix it now.

2. After you have installed this NLTK installer, run either the 'Numpy 32 bit windows installer' or Numpty 64 bit windows installer' depending on the type of windows you have on your PC.

   To find out if your computer is running a 32-bit or 64-bit version of Windows in Windows 7 or Windows Vista open System by clicking the Start button , right-clicking Computer, and then clicking Properties. Under System, you can view the system type.

3. Now run the PyYaml installer.exe file.

Check that everything is working:

Open your Windows command line (Start->Search 'cmd'->Open)
In the black box type 'python' without the quotes and hit enter

If this doesn't work, you didn't setup python on the command line correctly and should email us ASAP.

Now type 'import nltk'
If nothing happens and your cursor goes to the next line - Hooray!
If you get any type of error, please make sure you followed the three steps listed above.
There are *3* seperate installers you should have run - NLTK installer, Numpy installer (either 32bit or 64) and PyYaml installer.
Please email us ASAP if you can't the 'import nltk' statement to work!

If everything is working:

- Read example.py and see some examples of using the Natural Language Toolkit in Python.

- Find the instructions on the compulsory task in the example.py file. Follow the instructions in the comments to complete the task.

Acknowledgements:

This task was made with permission from the University if Edinburgh Informatics department – the leading Natural Language Processing Research Department in the world. Examples, slides and code was used with permission from this department on the condition that all lessons we give continue to be free. Text, code and slides were adapted from the 2nd year Edinburgh course – Informatics 2A: Processing Formal and Natural Languages. All course content can be seen here: http://www.inf.ed.ac.uk/teaching/courses/inf2a/

Feedback:

Please email all feedback about this task and the Artificial Intelligence explanation above to students@hyperiondev.com – especially if you feel the explanation is not clear. Please also complete our short 5 question survey at http://www.surveymonkey.com/s/PSLGN6F to tell us how you think our course is going so far.

Need some help?

Firstly, make sure you have installed and setup all programs correctly.

Please refer to the pdf file **PythonReference.pdf** if you would like more examples of Python coding and explanations.

If you having problems understanding example.py or how to complete Task 4, please contact students@hyperiondev.com. One on one help sessions are available over the internet or in person in Westville, Durban or UKZN (Westville Campus) and these can be arranged by contacting us. **We employ paid teachers who are here to help you!** A full list of ways to get help can be seen on: www.rmoola.com/advice.html

If there are any specific areas that are unclear or areas that require additional information:

Please add to 'What do you want to learn.txt' and one of our teachers will assist you once they read your request.

A peek ahead:

Task 5: Python on the web, extracting statistical data and using 2 dimension lists in python to generate Javascript graphs (such as the one seen on http://ec2-175-41-179-225.ap-southeast-1.compute.amazonaws.com/files/KNGBF175VER5ZAWJ8DM7/Visualize.html. A task based on work done for the South African Medial Research Council (Biomedical Research Division).