iST®

Automated Smart Investment Commentary with Python

Richard Saldanha

richard@oxquant.com

University of York

14 September 2022

PROUD SUPPORTER OF THE
Technician Commitment

National
Technician
Development
Centre
for Higher Education

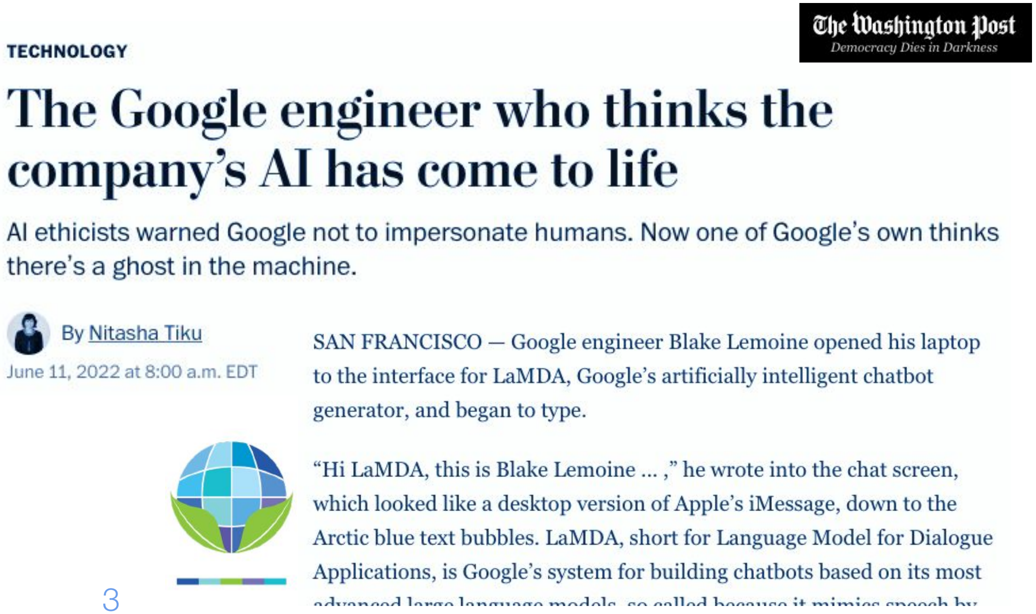
Science
Council

Natural Language Processing (NLP)

- NLP refers to range of computer science techniques aimed at processing human (natural) languages in an effective often interpretive manner
- NLP includes NL Understanding, Learning and Generation
- Terms have been around since 1950s
- Fairly recent more widespread use in real life

Unpacking NLP

- Natural Language Understanding (NLU) AI-hard problem aimed at machine comprehension
- Natural Language Learning (NLL) automatic triggering of specific responses to language using rules that define language
- Natural Language Generation (NLG) natural language from machine representation NL system



The Washington Post
Democracy Dies in Darkness

TECHNOLOGY

The Google engineer who thinks the company's AI has come to life

AI ethicists warned Google not to impersonate humans. Now one of Google's own thinks there's a ghost in the machine.

By Nitasha Tiku
June 11, 2022 at 8:00 a.m. EDT

SAN FRANCISCO — Google engineer Blake Lemoine opened his laptop to the interface for LaMDA, Google's artificially intelligent chatbot generator, and began to type.

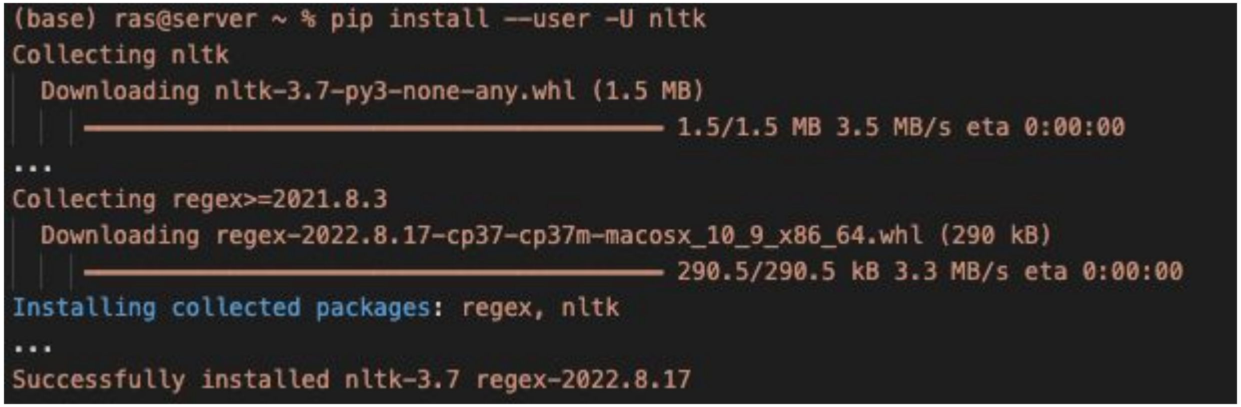
"Hi LaMDA, this is Blake Lemoine ... ," he wrote into the chat screen, which looked like a desktop version of Apple's iMessage, down to the Arctic blue text bubbles. LaMDA, short for Language Model for Dialogue Applications, is Google's system for building chatbots based on its most advanced large language models, so called because it mimics speech by

PROUD SUPPORTER OF THE
Technician Commitment

Natural Language Toolkit (NLTK)

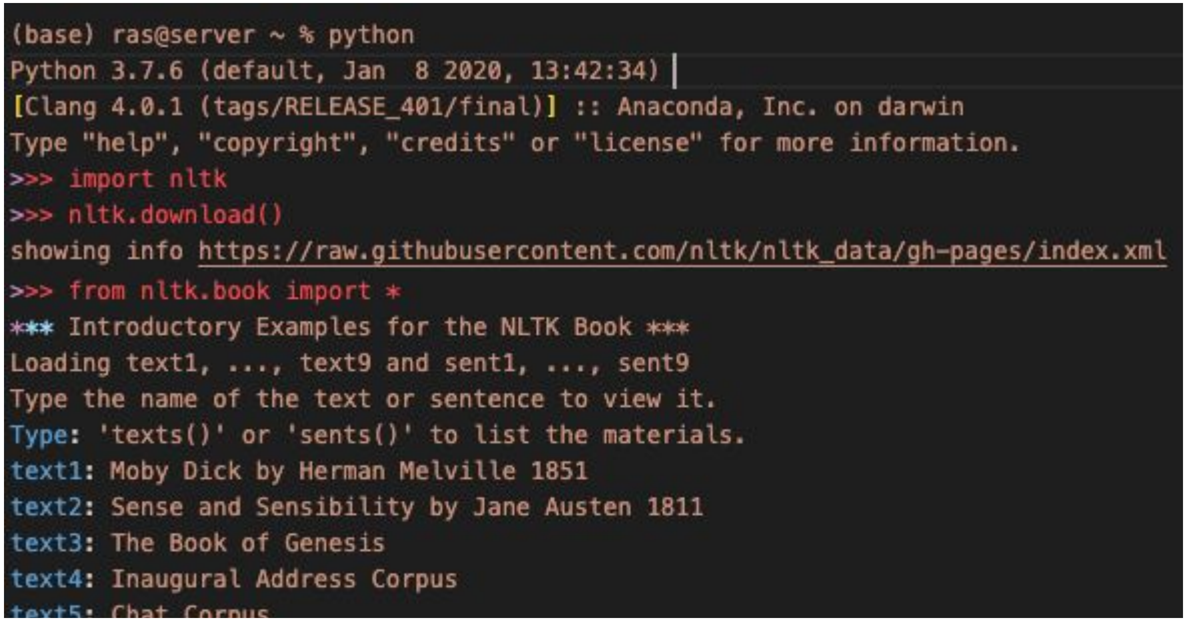
- NLTK allows Python programs to work with human language data
- Provides text processing libraries for:
  - **classification** choosing the correct class label for a given input
  - **tokenization** separating a chunk of text into smaller units (words, subwords individual characters) called *tokens*
  - **stemming** producing morphological variants of root/base word, e.g. reducing “chocolates”, “chocolatey” and “choco” to “chocolate”
  - **tagging** classifying words into parts of speech
  - **parsing** generating tree structures that depict internal organization of text
  - **semantic reasoning** representing the meaning of text in suitable form

Interacting directly with NLTK



```
(base) ras@server ~ % pip install --user -U nltk
Collecting nltk
  Downloading nltk-3.7-py3-none-any.whl (1.5 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 1.5/1.5 MB 3.5 MB/s eta 0:00:00
...
Collecting regex>=2021.8.3
  Downloading regex-2022.8.17-cp37-cp37m-macosx_10_9_x86_64.whl (290 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 290.5/290.5 kB 3.3 MB/s eta 0:00:00
Installing collected packages: regex, nltk
...
Successfully installed nltk-3.7 regex-2022.8.17
```

← installation



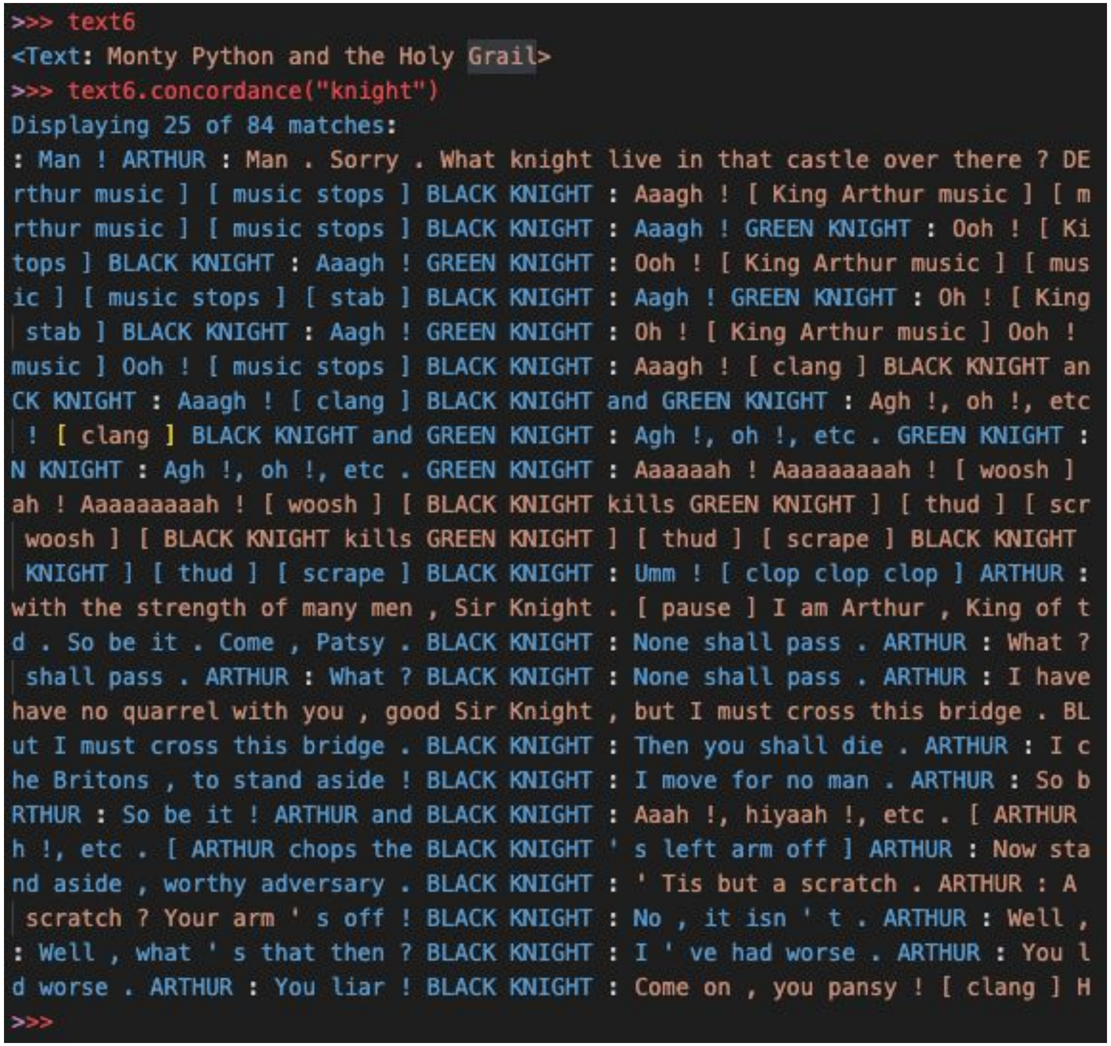
```
(base) ras@server ~ % python
Python 3.7.6 (default, Jan  8 2020, 13:42:34)
[Clang 4.0.1 (tags/RELEASE_401/final)] :: Anaconda, Inc. on darwin
Type "help", "copyright", "credits" or "license" for more information.
>>> import nltk
>>> nltk.download()
showing info https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/index.xml
>>> from nltk.book import *
*** Introductory Examples for the NLTK Book ***
Loading text1, ..., text9 and sent1, ..., sent9
Type the name of the text or sentence to view it.
Type: 'texts()' or 'sents()' to list the materials.
text1: Moby Dick by Herman Melville 1851
text2: Sense and Sensibility by Jane Austen 1811
text3: The Book of Genesis
text4: Inaugural Address Corpus
text5: Chat Corpus
```

← usage

PROUD SUPPORTER OF THE
Technician Commitment

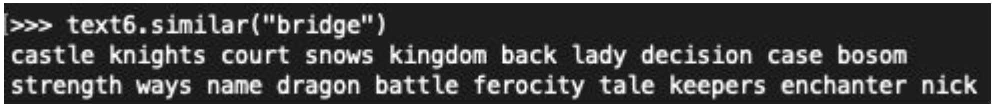
National Technician Development Centre
for Higher Education

Science Council



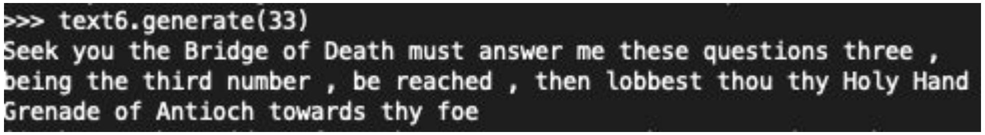
```
>>> text6
<Text: Monty Python and the Holy Grail>
>>> text6.concordance("knight")
Displaying 25 of 84 matches:
: Man ! ARTHUR : Man . Sorry . What knight live in that castle over there ? DE
rthur music ] [ music stops ] BLACK KNIGHT : Aaagh ! [ King Arthur music ] [ m
rthur music ] [ music stops ] BLACK KNIGHT : Aaagh ! GREEN KNIGHT : Ooh ! [ Ki
tops ] BLACK KNIGHT : Aaagh ! GREEN KNIGHT : Ooh ! [ King Arthur music ] [ mus
ic ] [ music stops ] [ stab ] BLACK KNIGHT : Aagh ! GREEN KNIGHT : Oh ! [ King
 stab ] BLACK KNIGHT : Aagh ! GREEN KNIGHT : Oh ! [ King Arthur music ] Ooh !
music ] Ooh ! [ music stops ] BLACK KNIGHT : Aaagh ! [ clang ] BLACK KNIGHT an
CK KNIGHT : Aaagh ! [ clang ] BLACK KNIGHT and GREEN KNIGHT : Agh !, oh !, etc
 ! [ clang ] BLACK KNIGHT and GREEN KNIGHT : Agh !, oh !, etc . GREEN KNIGHT :
N KNIGHT : Agh !, oh !, etc . GREEN KNIGHT : Aaaaaah ! Aaaaaaaaah ! [ woosh ]
ah ! Aaaaaaaaah ! [ woosh ] [ BLACK KNIGHT kills GREEN KNIGHT ] [ thud ] [ scr
 woosh ] [ BLACK KNIGHT kills GREEN KNIGHT ] [ thud ] [ scrape ] BLACK KNIGHT
 KNIGHT ] [ thud ] [ scrape ] BLACK KNIGHT : Umm ! [ clop clop clop ] ARTHUR :
with the strength of many men , Sir Knight . [ pause ] I am Arthur , King of t
d . So be it . Come , Patsy . BLACK KNIGHT : None shall pass . ARTHUR : What ?
 shall pass . ARTHUR : What ? BLACK KNIGHT : None shall pass . ARTHUR : I have
have no quarrel with you , good Sir Knight , but I must cross this bridge . BL
ut I must cross this bridge . BLACK KNIGHT : Then you shall die . ARTHUR : I c
he Britons , to stand aside ! BLACK KNIGHT : I move for no man . ARTHUR : So b
RTHUR : So be it ! ARTHUR and BLACK KNIGHT : Aaah !, hiyaah !, etc . [ ARTHUR
h !, etc . [ ARTHUR chops the BLACK KNIGHT ' s left arm off ] ARTHUR : Now sta
nd aside , worthy adversary . BLACK KNIGHT : ' Tis but a scratch . ARTHUR : A
 scratch ? Your arm ' s off ! BLACK KNIGHT : No , it isn ' t . ARTHUR : Well ,
: Well , what ' s that then ? BLACK KNIGHT : I ' ve had worse . ARTHUR : You l
d worse . ARTHUR : You liar ! BLACK KNIGHT : Come on , you pansy ! [ clang ] H
>>>
```

← **concordance** shows every occurrence of a given word (e.g. ”knight”), together with context



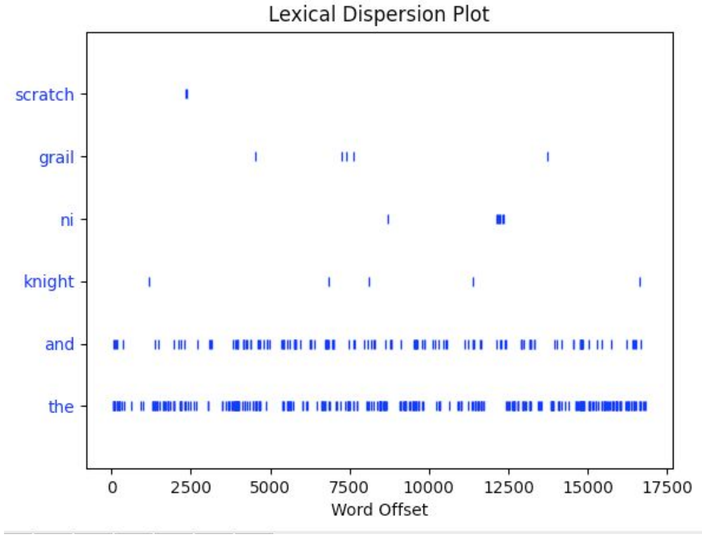
```
>>> text6.similar("bridge")
castle knights court snows kingdom back lady decision case bosom
strength ways name dragon battle ferocity tale keepers enchanter nick
```

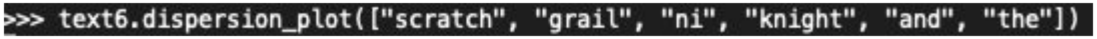
← **similar** gives other words in a related range of contexts



```
>>> text6.generate(33)
Seek you the Bridge of Death must answer me these questions three ,
being the third number , be reached , then lobbest thou thy Holy Hand
Grenade of Antioch towards thy foe
```

← **generate** random text in the same writing style as the author (33 = 30 words plus 3 commas)





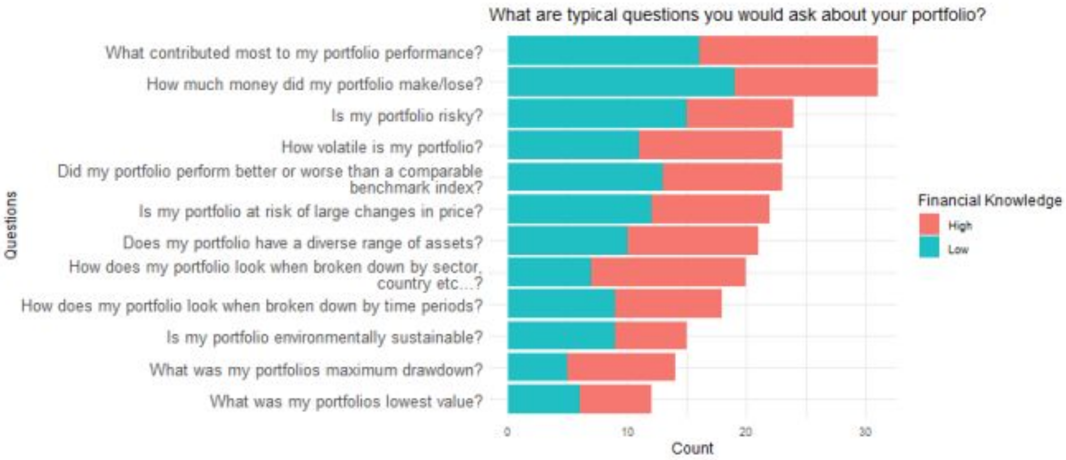
```
>>> text6.dispersion_plot(["scratch", "grail", "ni", "knight", "and", "the"])
```

lexical (word) **dispersion_plot** requires NumPy and Matplotlib ← packages to work

and now for something more practical …

Henry Shepherd (MSc Student, Economics for Business Intelligence and Systems at the *University of Bath* hcshepherd98@gmail.com) has been working on automated translation of investment portfolio performance into natural language for investors.

Q. What do investors want to know about their portfolio?

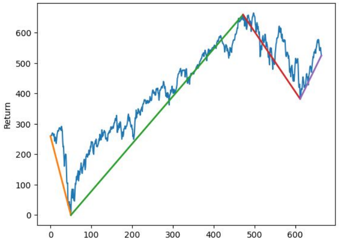


iST

Python pipeline

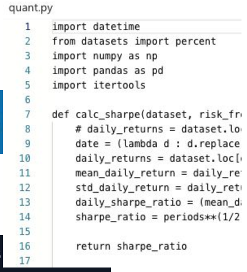
yfinance

1. Build portfolio using Yahoo finance data



2. Note portfolio performance over time

pyportfolio



quant.py

```
import datetime
from datasets import percent
import numpy as np
import pandas as pd
import itertools

def calc_sharpe(dataset, risk_fr
    # daily_returns = dataset.lo
    date = (lambda d : d.replace
    daily_returns = dataset.loc[
    mean_daily_return = daily_re
    std_daily_return = daily_ret
    daily_sharpe_ratio = (mean_d
    sharpe_ratio = periods**(1/2

    return sharpe_ratio

```

QuantStats

Portfolio analytics for quants, written in Python

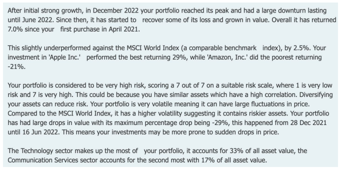
3. Calculate portfolio performance statistics

4. Create commentary via template, e.g. "*Your portfolio has performed* `<sentiment>` *returning* `<value>`"

🤗 Transformers

BART (from Facebook)

5. Order and decide on sentence merging



After initial strong growth, in December 2022 your portfolio reached its peak and had a large downturn lasting until June 2022. Since then, it has started to recover some of its loss and grown in value. Overall it has returned 7.0% since your first purchase in April 2021.

This slightly underperformed against the MSCI World Index (a comparable benchmark index), by 2.5%. Your investment in 'Apple Inc.' performed the best returning 29%, while 'Amazon, Inc.' did the poorest returning -21%.

Your portfolio is considered to be very high risk, scoring a 7 out of 7 on a suitable risk scale, where 1 is very low risk and 7 is very high. This could be because you have similar assets which have a high correlation. Diversifying your assets can reduce risk. Your portfolio is very volatile meaning it can have large fluctuations in price. Compared to the MSCI World Index, it has a higher volatility suggesting it contains riskier assets. Your portfolio has had large drops in value with its maximum percentage drop being -29%, this happened from 28 Dec 2021 until 16 Jun 2022. This means your investments may be more prone to sudden drops in price.

The Technology sector makes up the most of your portfolio, it accounts for 33% of all asset value, the Communication Services sector accounts for the second most with 17% of all asset value.

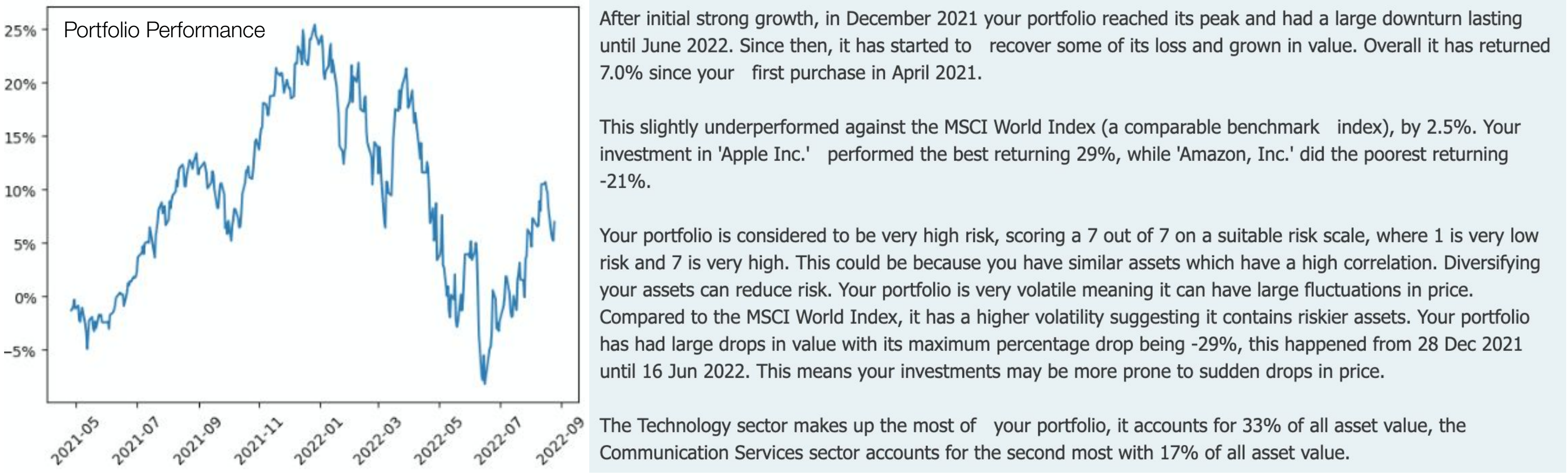
6. Final output

PROUD SUPPORTER OF THE Technician Commitment

National Technician Development Centre for Higher Education

Science Council

Example output



After initial strong growth, in December 2021 your portfolio reached its peak and had a large downturn lasting until June 2022. Since then, it has started to recover some of its loss and grown in value. Overall it has returned 7.0% since your first purchase in April 2021.

This slightly underperformed against the MSCI World Index (a comparable benchmark index), by 2.5%. Your investment in 'Apple Inc.' performed the best returning 29%, while 'Amazon, Inc.' did the poorest returning -21%.

Your portfolio is considered to be very high risk, scoring a 7 out of 7 on a suitable risk scale, where 1 is very low risk and 7 is very high. This could be because you have similar assets which have a high correlation. Diversifying your assets can reduce risk. Your portfolio is very volatile meaning it can have large fluctuations in price. Compared to the MSCI World Index, it has a higher volatility suggesting it contains riskier assets. Your portfolio has had large drops in value with its maximum percentage drop being -29%, this happened from 28 Dec 2021 until 16 Jun 2022. This means your investments may be more prone to sudden drops in price.

The Technology sector makes up the most of your portfolio, it accounts for 33% of all asset value, the Communication Services sector accounts for the second most with 17% of all asset value.

PROUD SUPPORTER OF THE
Technician Commitment

National
Technician
Development
Centre
for Higher Education

Science
Council

See also spaCy

NLTK and spaCy (https://spacy.io/usage/spacy-101) are two of the most popular NLP toolkits available in Python

- NLTK aimed at researchers who want to build something from the ground up or test thesis
- spaCy aimed more at app developers
- NLTK provides many algorithms to get something done, spaCy makes choices of the 'best'
- NLTK implements a string-based approach to NLP – input strings return output strings
- spaCy is more object oriented – each function returns objects instead of strings or arrays

iST®

RSciTech
Registered Science Technician

RSci
Registered Scientist

CSci
Chartered Scientist

FIScT(Reg)
Registered Practitioner

MIScT(Reg)
Registered Practitioner

DALL·E 2 generated graphic based on language input *"robot and python in photorealistic dance scene"*; see https://openai.com/dall-e-2/

office@istonline.org.uk

https://istonline.org.uk/ist/conference/

PROUD SUPPORTER OF THE
Technician Commitment

National
Technician
Development
Centre
for Higher Education

Science Council