*Plan van Aanpak*

*Daan Eising, Mark Olieman, Jesse van der Westen*

# Probleemstelling

## Title:

S&P 500 (VUSA.AS) voorspellingen

## Bedrijf:

Vanguard, een bedrijf dat vele aandelen en fondsen beheert

Vanguard is een bedrijf dat vele aandelen en ETF’s beheert hieronder vallen ook ETF’s die de S&P 500 volgen. S&P 500 staat voor standards and poors 500 en probeert de 500 grootste bedrijven uit de verenigde staten te volgen. Als mensen het over de Amerikaanse markt hebben refereren ze vaak naar deze ETF. Vanguard wilt nu beter begrijpen hoe de ETF beweegt.

De missie van het machine learning team is om de data van de bedrijven in de S&P 500, de S&P500 zelf en artikelen over de bedrijven en de S&P500 te analyseren en vervolgens doormiddel van een machine learning model een voorspellend model te creëren die de ETF VUSA.AS weet te voorspellen. Dit is een ETF die de S&P500 volgt.

## Vereisten van het data-engineeringteam:

1. Een database met prijsgeschiedenis, gegevens over de bedrijven en nieuwsartikelen over de bedrijven die in de SP500 zitten

2. belangrijke features zijn onder andere prijsgeschiedenis, aandelen in de ETF, jaar- en rekeningverslagen, positiviteit/negativiteit van aandelen in de ETF in het nieuws

3. De tekstdata moet worden verwerkt naar feature-data die geschikt is voor het trainen van machine learning modellen

4. een robuuste pipeline die artikelen uit verschillende bronnen weet te halen, deze analyseert op hoedanig positief het artikel is en dit omzetten naar bruikbare feature-data. Ook moet deze pipeline data over de bedrijven en de S&P500 ophalen uit api’s en databases en dit samenvoegen, verwerken zodat het bruikbaar is voor een machine learning model en vervolgens laden.

# Data extraheren

Er zal onder meer gebruik worden gemaakt van de volgende databronnen:

1. <https://github.com/ranaroussi/yfinance> (API)

2. <https://datahub.io/core/s-and-p-500> (Publieke Database)

3. <https://finance.yahoo.com/> (Artikelen scrapen)

4. <https://www.investopedia.com/markets-news-4427704> (Artikelen scrapen)

# Data transformeren

1. Sentiment analysis op de tekst zodat deze gebruikt kan worden als feature.

2. Verder moeten er uit alle artikelen dezelfde overeenkomstige lijst met features hebben.

3. Alle data moet passend zijn, consistent zijn, en bij elkaar staan om bruikbaar te worden.

# Data laden

Het data-engineeringteam zal één CSV-bestand opleveren, met alle verzamelde data en sentiment analyses van de artikelen.

# Planning

|  |  |
| --- | --- |
| *Week nummer* | *Taak* |
| 3 | Beginnen aan: PvA, domeinonderzoek  ruwe data verzamelen |
| 4 | Afronden: PvA, domeinonderzoek  PvA Inleveren.  Beginnen aan: EDA |
| 5 | Afronden: EDA  EDA uitvoeren op data |
| 6 | Beginnen aan: data extraherende pipeline |
| 7 | Verder werken aan: data extraherende pipeline |
| 8 | Afronden: data extraherende pipeline |
| 9 | Beginnen aan: transformeren-onderdeel pipeline, data opschonen/preprocessen |
| 10 | Verder werken aan: transformeren-onderdeel pipeline, data opschonen/preprocessen |
| 11 | Afronden: transformeren-onderdeel pipeline, data opschonen/preprocessen |
| 12 | Beginnen aan: data laden pipeline, database [creëren](https://www.google.com/search?client=firefox-b-d&sca_esv=566330112&sxsrf=AM9HkKkqHUvaj_ludndCeCPAw98P4stB2Q:1695061145569&q=cre%C3%ABren&spell=1&sa=X&ved=2ahUKEwjazdDr4rSBAxXDhv0HHYTkAYsQkeECKAB6BAgLEAE) |
| 13 | Verder werken aan: data laden pipeline, database [creëren](https://www.google.com/search?client=firefox-b-d&sca_esv=566330112&sxsrf=AM9HkKkqHUvaj_ludndCeCPAw98P4stB2Q:1695061145569&q=cre%C3%ABren&spell=1&sa=X&ved=2ahUKEwjazdDr4rSBAxXDhv0HHYTkAYsQkeECKAB6BAgLEAE) |
| 14 | Afronden: data laden pipeline, database [creëren](https://www.google.com/search?client=firefox-b-d&sca_esv=566330112&sxsrf=AM9HkKkqHUvaj_ludndCeCPAw98P4stB2Q:1695061145569&q=cre%C3%ABren&spell=1&sa=X&ved=2ahUKEwjazdDr4rSBAxXDhv0HHYTkAYsQkeECKAB6BAgLEAE) |
| 15 | Beginnen aan: explanatory dashboard |
| 16 | Verder werken aan: explanatory dashboard  Beginnen met: model testen |
| 17 | Verder met: model testen  Afronden: explanatory dashboard  Notebook laatste aanpassingen  Notebook controleren op fouten  Individuele reflectie schrijven |
| 18 | Afronden: model testen  Inleveren: notebook & individuele reflectie  Voorbereiden presentatie: Notebook code doornemen en begrijpen |
| 19 | Rapport presenteren |