ONDERZOEKSVOORSTEL

Een applicatie die dagelijks een AI-gegenereerde-kunstwerk deelt op basis van een automatische analyse van nieuwswebsites of socialmediaplatformen: Toegepast onderzoek.

Bachelorproef, 2022-2023

Dario Bronders

E-mail: dario.bronders@student.hogent.be
Co-promotor: M. De Buck (We-Are, manu@we-are.be)

Samenvatting

In dit toegepast onderzoek wordt er in de eerste fase onderzocht op welke manier we nieuwswebsites of social media platformen kunnen scrapen aan de hand van een python library BeautifulSoup. Deze zal dan ook worden geïmplementeerd. In de tweede fase gaan we onderzoeken hoe we deze data kunnen rangschikken en filteren, op deze manier achterhalen we de kernzaak van de dag. Eenmaal we dit achterhaald hebben, kunnen we dit vervolgens gebruiken om een kunstwerk te genereren op basis van de tekstuele input. Hiervoor zullen we deep learning models op toepassen die in 2022 publiek beschikbaar zijn gesteld.

Keuzerichting: Mobile & Enterprise development
Sleutelwoorden: Data scrapen en analyseren, Stable Diffusion en DALL-E 2, AI-gegenereerd kunstwerk

Inhoudsopgave

1. Introductie

De meestgekende nieuwsbronnen proberen al jaren objectief en feitelijk te blijven om informatie vanop eenzelfde standpunt en met een gelijkaardig boodschap over te brengen.
Binnen mijn toegepast onderzoek zal ik een toepassing maken die een kunstwerk zal genereren met behulp van één of meerdere deep learning modellen. Dit zou er voor zorgen dat het dagelijkse hoogtepunt geabstraheerd kan worden tot een unieke AI-gegenereerde kunstwerk.

2. Literatuurstudie

2.1. Wat is webscraping?

Webscraping is een term die gebruikt wordt voor het extraheren van inhoud van websites om het te importeren in lokale opslag zoals een database of CSV bestand. (Salem & Mazzara, 2020)
Websites kunnen ervoor kiezen om een *robots.txt* in de root van hun filesystem te plaatsen. Binnen deze tekstfile kunnen ze beschrijven welke routes gescraped mogen worden. (Google, 2022)

```
# Alle auteurs-, naburige en databankrechten die op de inhoud en opmaak van de DPG Media websites
# en DPG Media apps rusten, worden door DPG Media BV uitdrukkelijk voorbehouden. De inhoud van de
# DPG Media websites en apps is uitsluitend voor persoonlijk, niet-commercieel gebruik en het is
# niet toegestaan om gegevens van de website of uit de apps door middel van screen scraping
# (of een andere geautomatiseerde werkwijze) te vergaren.
# Zie ook de Gebruikersvoorwaarden van DPG Media op www.dpgmedia.be/gebruiksvoorwaarden

# All copyrights, neighbouring rights and database rights in the content and layout of the
# DPG Media websites and DPG Media apps are explicitly reserved by DPG Media BV. The content of the DPG Media
# websites and DPG Media apps is for personal, non-commercial use only and it is not allowed to
# collect data from the website or from the apps by means of screen scraping (or any other
# automated method).
# See also the terms of use of DPG Media at www.dpgmedia.be/gebruiksvoorwaarden

# Tell robots which pages are not very interesting
User-agent: *
Disallow: /*webview
Disallow: /auth
Disallow: /*widget*
Disallow: /*?*otag=
Disallow: /*?*abo_type=
Disallow: /*?*utm_source=
Disallow: /zoeken?query=
Disallow: /inloggen?*
# Tell robots not to crawl redirect urls
Disallow: *?*redirect_urls*

User-agent: Twitterbot
Allow: /

Sitemap: https://www.hln.be/sitemap.xml
Sitemap: https://www.hln.be/sitemap-news.xml
```

Figuur 1: voorbeeld: www.hln.be/robots.txt

2.2. Wat is DALL-E (2)?

DALL-E is een kunstmatig intelligentieprogramma ontwikkeld door openAI dat beelden creëert uit tekstuele beschrijvingen, ook wel *prompts* genoemd. Het gebruikt een versie met 12 miljard parameters van het GPT-3 Transformer-model om natuurlijke taalinvoer te interpreteren en overeenkomstige beelden te genereren. In april 2022 heeft OpenAI DALL-E 2 gelanceerd, ontwikkeld om meer realistische foto's met hogere resolutie te kunnen genereren. (nl.wikipedia.org, 2022) (en.wikipedia.org, 2022a)

DALL-E 2 is bovendien getrained met behulp van 650 milioen tekstinputs gescraped van het internet. (Borji, 2022)

Deze code is niet open source maar kun je gebruiken aan de hand van de openAI API.

2.3. Wat is Stable Diffusion?

Stable Diffusion is een deep learning, tekst-naar-beeld model uitgebracht in 2022. In tegenstelling tot DALL-E (2) is Stable Diffusion getrained aan de hand van een diepe generatieve neurale netwerk. Deze code is opensource en kun je lokaal draaien op een computer met een GPU. (en.wikipedia.org, 2022b)

3. Methodologie

Inleiding
Het toegepast onderzoek begint 2 maart 2023 en zal beëindigd worden voor 28 mei 2023.

Fase 1: Realiseren van een scraper
Om de data te bekomen van de verschillende soorten websites of social-media platformen zal er een web scraper worden gemaakt. Deze scraper zal ontwikkeld worden in python met behulp van een externe library *BeautifulSoup*.
Doordat de presentatie van de verschillende artikelen kunnen verschillen in taal en structuur, zal de scraper een algoritme implementeren die het mogelijk maakt om op een uniforme manier verschillende websites te scrapen.

Fase 2: Data verwerken en analyseren
Tijdens de tweede fase zullen we onderzoeken op welke manier we de bekomen data uit voorgaande fase kunnen analyseren en sorteren.

Het zal belangrijk zijn om rekening te houden met de volgende vragen:

- Wat zijn de te extraheren kernzaken?
- Wat is het sentiment van de dag?
- Welke topic komt het vaakst voor?
- Op basis van welke gegevens kunnen we de artikels sorteren?

Nadat er een gepaste methode wordt gevonden om dit te realiseren, zal deze ook geïmplementeerd worden. Op deze manier kunnen we steeds het belangrijkste artikel van de dag eruit halen.

Fase 3: Kunstwerk genereren
Nu dat we weten uit de vorige fase wat het hoogtepunt van de dag was. Kunnen we hierop een kunstwerk laten genereren.
Hiervoor zal er gebruik gemaakt worden van een of meerdere deep learning modellen DALL-E 2 en/of Stable Diffusion die de kerntekst van een artikel zal omvormen tot een foto.

Één of beide technologieën zullen gebruikt worden.

4. Verwacht resultaten

Een applicatie ontwerpen die dagelijks een kunstwerk kan genereren op basis van het hoogtepunt van de dag.

Op 27 oktober, toen Marokko won van België tijdens de WK kon het hoogtepunt in België *'Riots in Brussels after soccer game, painting'* geweest zijn. Hieronder vindt u enkele voorbeelden die gegenereerd zijn met behulp van DALL-E 2 op basis van deze tekstinput.

Referenties

Borji, A. (2022). Generated Faces in the Wild: Quantitative Comparison of Stable Diffusion, Midjourney and DALL-E 2. *Computer Vision and Pattern Recognition*.

en.wikipedia.org. (2022a). *DALL-E*. Wikipedia. https://en.wikipedia.org/wiki/DALL-E

en.wikipedia.org. (2022b). *Stable Diffusion*. Wikipedia. https://en.wikipedia.org/wiki/Stable_Diffusion

Google. (2022). *Introduction to robots.txt*. Google. https://developers.google.com/search/docs/crawling-indexing/robots/intro

nl.wikipedia.org. (2022). *DALL-E*. nl.wikipedia.org: DALL-E. https://nl.wikipedia.org/wiki/DALL-E

Salem, H., & Mazzara, M. (2020). Pattern Matching-based scraping of news websites. *Journal of Physics: Conference Series*, *1694*, 6.