De mogelijkheden van

Textmining voor Archiefbeheer

Simon Pouwelse
Muriël Valckx

De mogelijkheden van

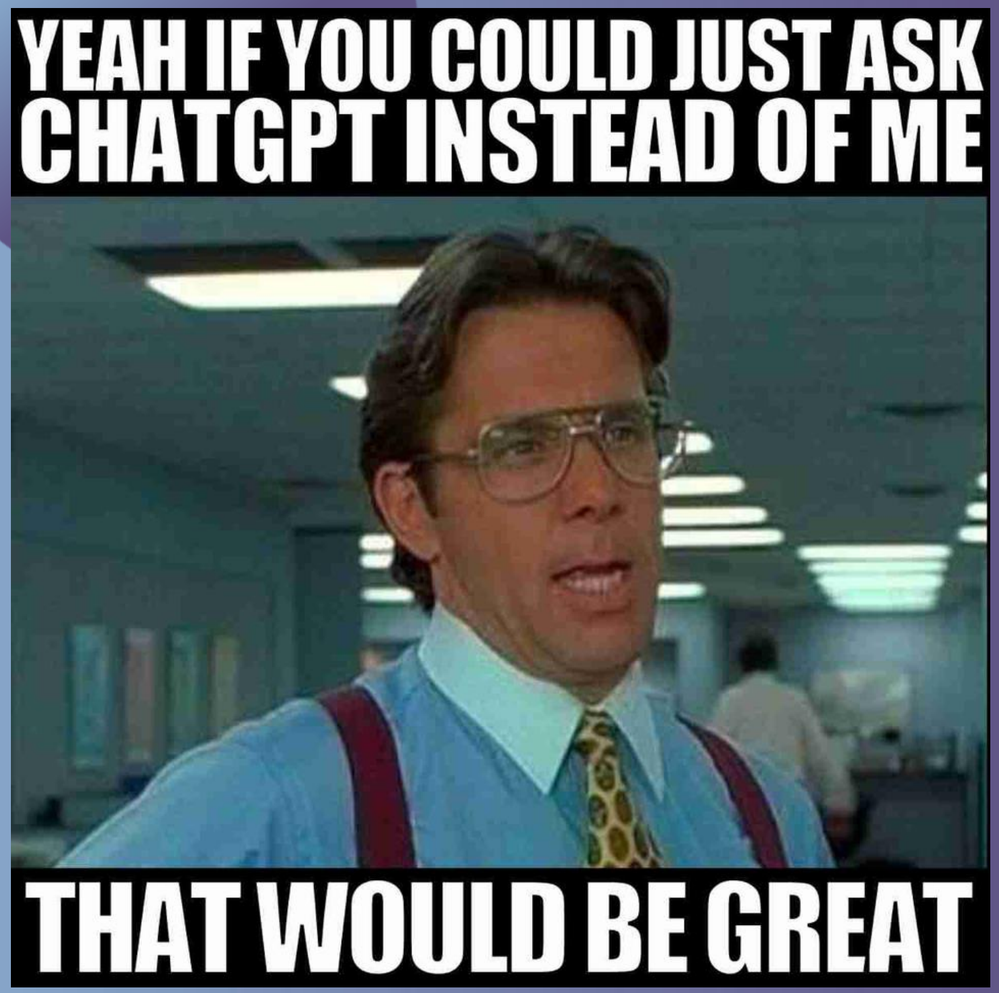
Textmining voor Archiefbeheer

Simon Pouwelse
Data Scientist &
Data Analist

Provincie Zeeland

Muriël Valckx
Data Scientist & Trainee
Informatiemanagement

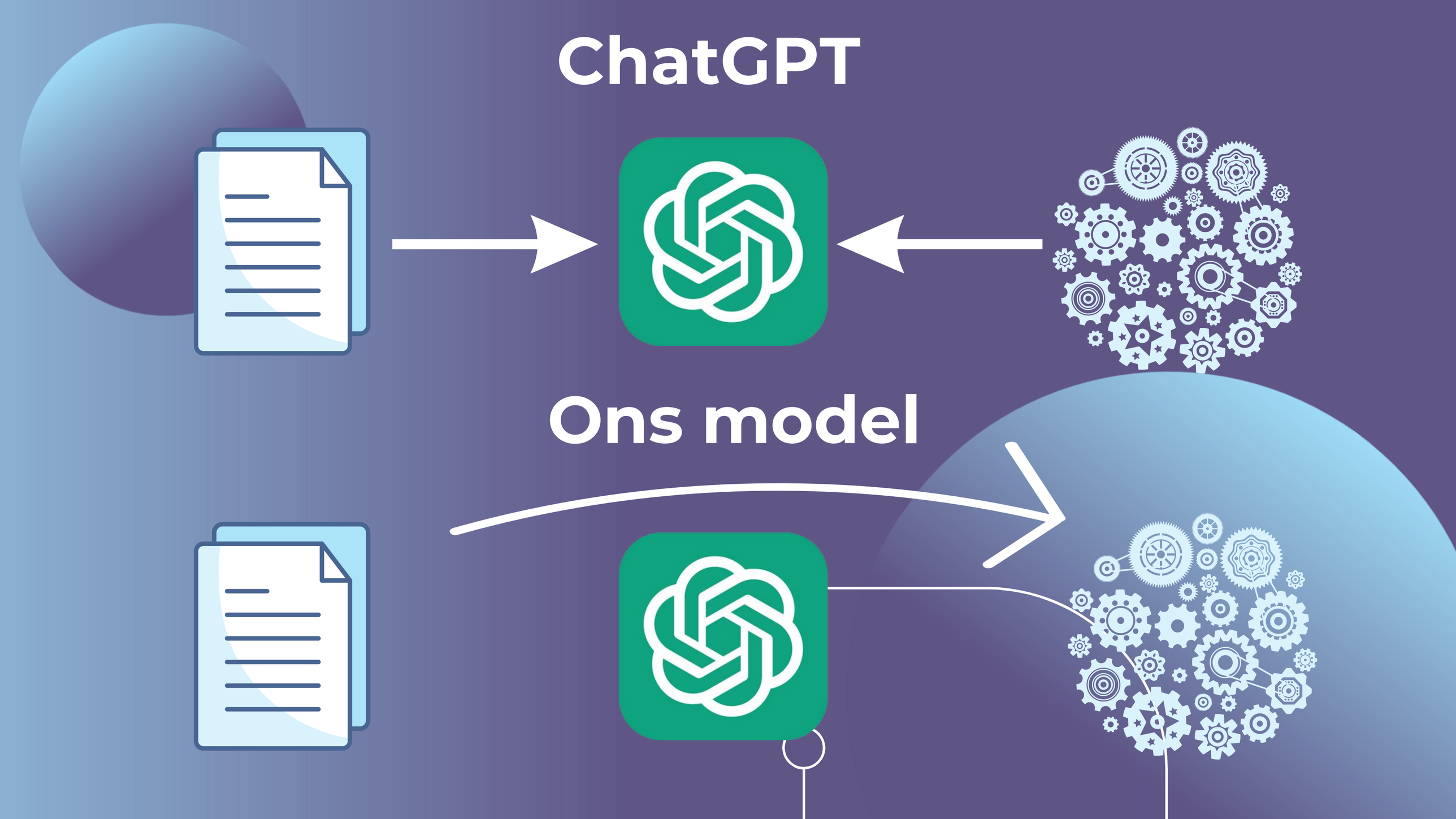
Zeeuws Archief



Bron

afbeelding

OpenAI


ChatGPT
Ons model

Agenda

Text mining

Wat is het en hoe werkt het?

Archiefbeheer

Welke mogelijkheden zijn er?

Volgende stappen

Wat heeft de toekomst te bieden?

Wat is text mining?

Waardevolle informatie uit grote hoeveelheden tekstmateriaal halen

Text mining

Tekst opschonen

Inhoud begrijpen

Nieuwe informatie

Tekst opschonen

als je binnenkomt in het spoorweg museum, word je begroet door een ENORME trein, het is niet te missen. Deze Attractie staat bovenaan de lijst en is echt een Blikvanger .het museum heeft een gevarieerde collectie locomotieven, in alle maten en vormen.%

Tekst opschonen

als je binnenkomt in het spoorweg museum, word je begroet door een ENORME trein, het is niet te missen. Deze Attractie staat bovenaan de lijst en is echt een Blikvanger .het museum heeft een gevarieerde collectie locomotieven, in alle maten en vormen.%

Tokenization

["als", "je", "binnenkomt", "in", "het", "spoorwegmuseum", "word", "je", "begroet", "door", "een", "enorme", "trein", "het", "is", "niet", "te", "missen", "deze", "attractie", "staat", "bovenaan", "de", "lijst", "en", "is", "echt", "een", "blikvanger", "het", "museum", "heeft", "een", "gevarieerde", "collectie", "locomotieven", "in", "alle", "maten", "en", "vormen"]

Lemmatiseren

["binnenkomen", "spoorwegmuseum", "begroeten", "enorm", "trein", "missen",

"attractie", "bovenaan", "lijst", "echt", "blikvanger", "museum", "gevarieerd",

"collectie", "locomotief", "maat", "vorm"]

Stemming

["binnenkom", "spoorwegmuseum", "begroet", "enorm", "trein", "miss",

"attractie", "bovenaan", "lijst", "echt", "blikvang", "museum", "gevarieerd",

"collectie", "locomotief", "maat", "vorm"]

Vectoriseren

TF-IDF

Term Frequency-Inverse Document Frequency

Relevantie van woorden in een document, ten opzichte van een verzameling documenten.

- "binnenkomen" wordt 0.34987439
- "spoorwegmuseum" wordt 0.23940439
- "begroeten" wordt 0.30207439
- "enorm" wordt 0.2438042
- "trein" wordt 0.3256239

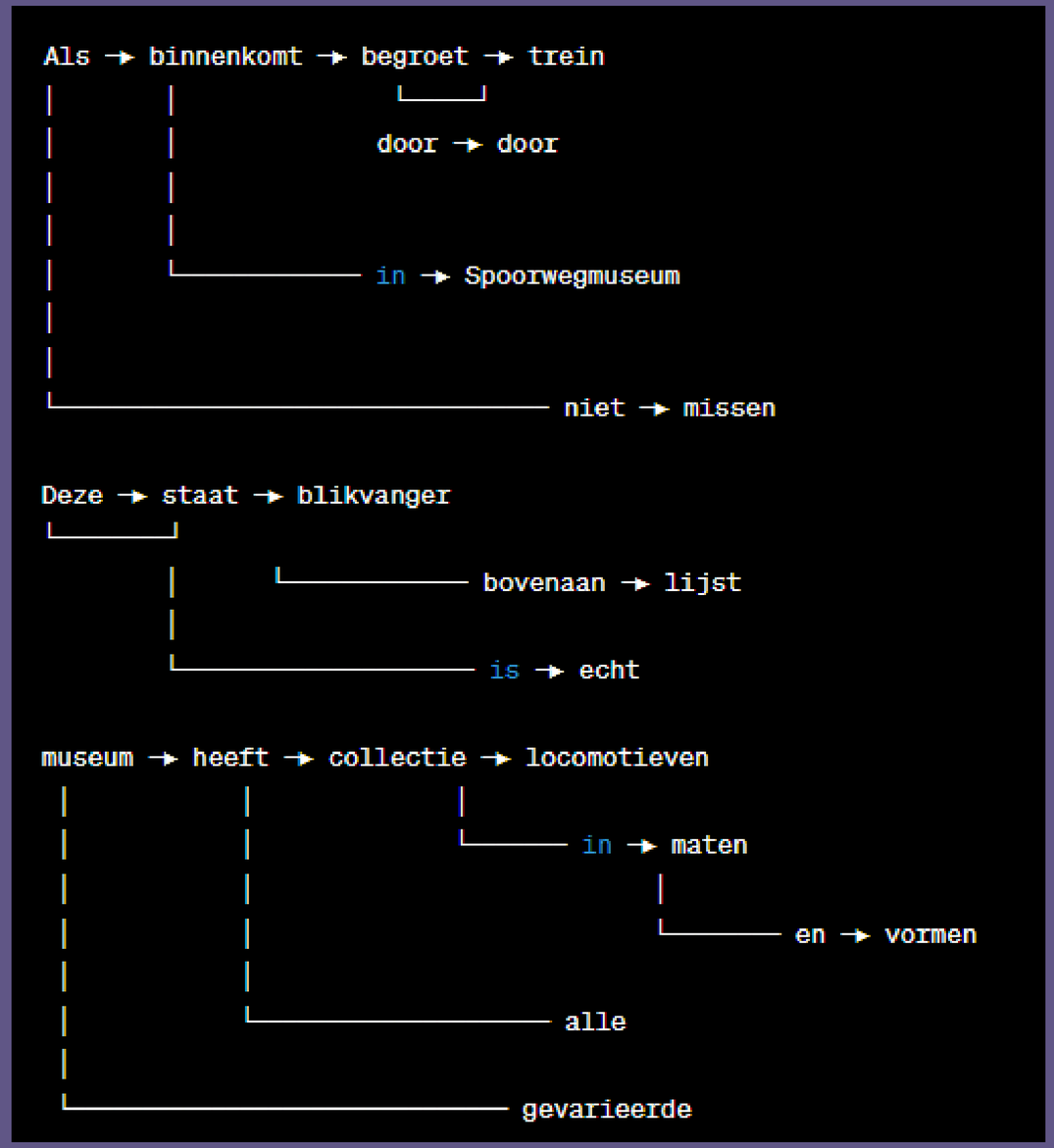
Tekst begrijpen

- **"Als"** - Voegwoord
- **"je"** - Pronomina (persoonlijk voornaamwoord)
- **"binnenkomt"** - Werkwoord
- **"in"** - Voorzetsel
- **"het"** - Lidwoord
- **"Spoorwegmuseum"** - Zelfstandig naamwoord (eigennaam)

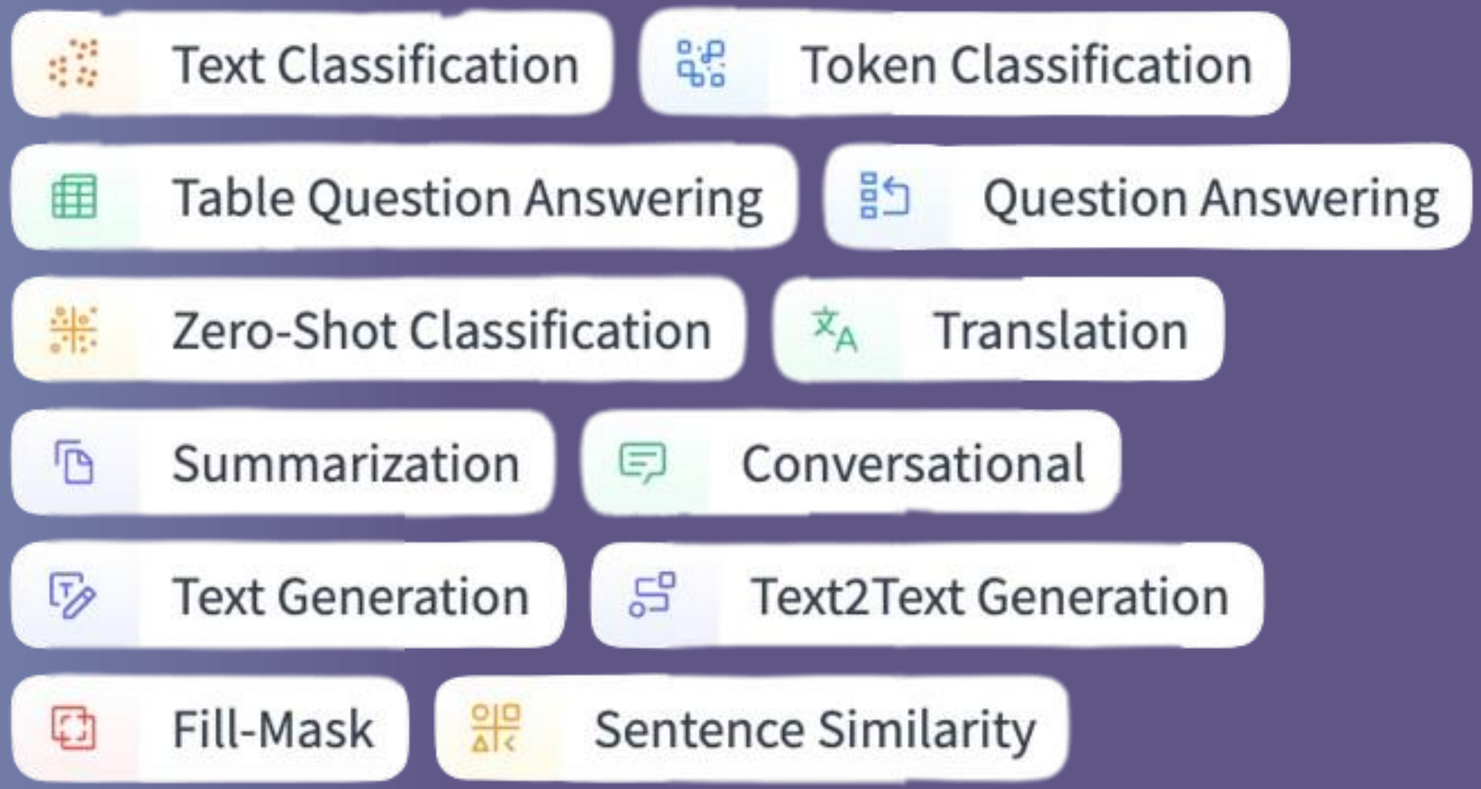
Tekst begrijpen

Als je binnenkomt in het **Spoorwegmuseum**, word je begroet door een enorme **trein**, het is niet te missen. Deze **attractie** staat bovenaan de lijst en is echt een blikvanger. Het **museum** heeft een gevarieerde **collectie** **locomotieven**, in alle maten en vormen.

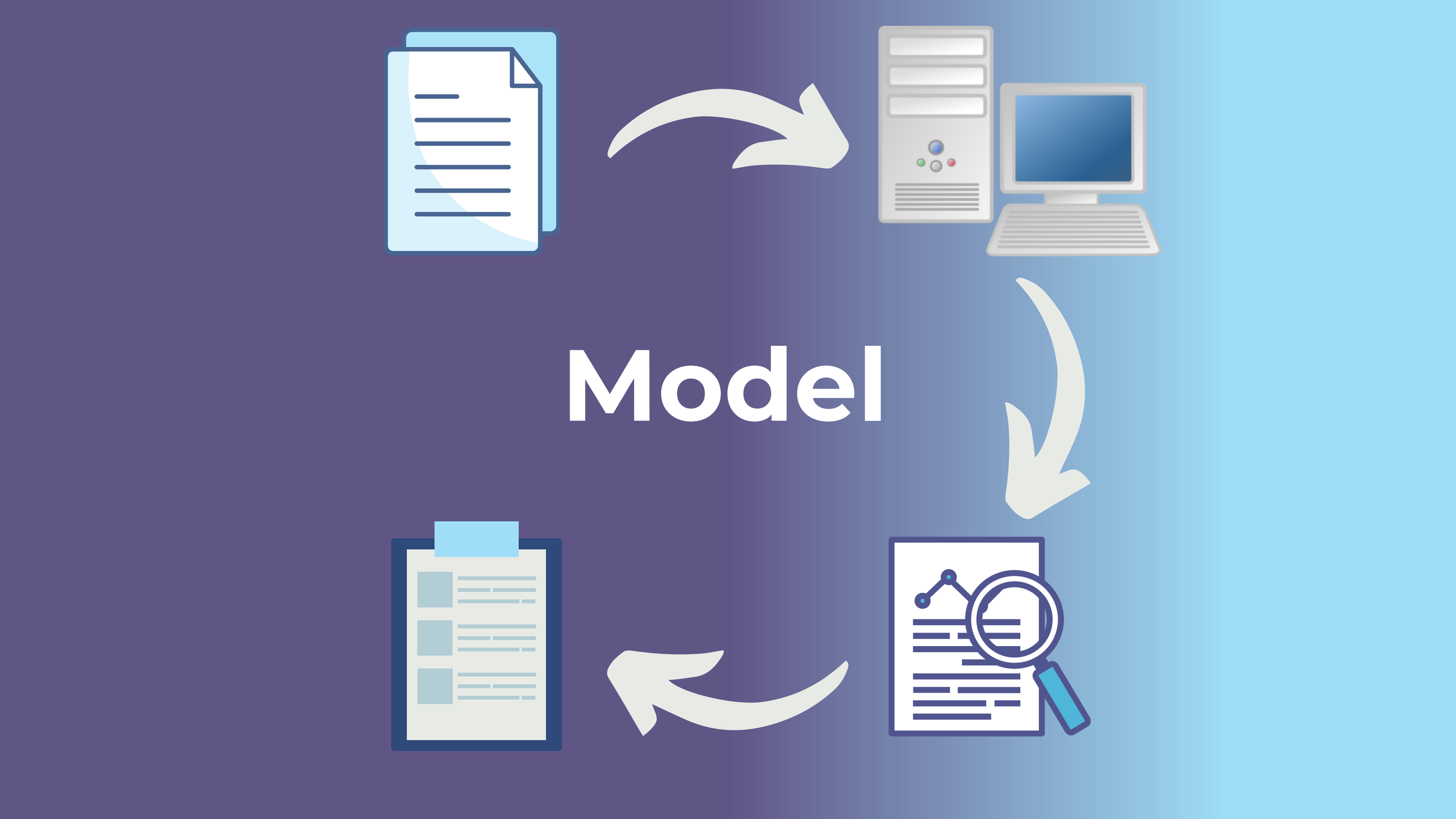
Tekst begrijpen



Toepassingen voor text mining



Bron: Huggingface.co


Model

```python
class pdf_metadata:
  def __init__(self, path: str, clean_text: bool = True):
    self.path = self.check_path(path)
    if self.path != None:
      self.text = self.get_text(clean_text)
      self.summary = self.get_summary()
      self.keywords = self.get_keywords()
      self.title = " ".join((self.keywords)[0:4])


  # Function to check if the given path is correctly specified and gives a PDF-file
  def check_path(self, path):
      """
      :param path: the path to the pdf-file
      :return:
      """
      # Check if the pdf path exists and if it leads to a PDF-file
      if not path.endswith('.pdf'):
        print('The given path doesn\'t give a PDF-file. Use the format: \'C\...\filename.pdf\'.')
        return None
      if not os.path.exists(path):
        print('The given path doesn\'t exist or can\'t be found. Use the format: \'C\...\filename.pdf\'.')
```

Standaard informatie	Geavanceerde informatie
Titel	Samenvatting
Taal	Steekwoorden
?	?

Resultaat metadataverrijking	
Originele bestandsnaam	20032474_S_20032474_4_A_43578_tds.pdf
Titel	December mensenrechtenvlag 10 mensenrechten
Samenvatting	De provincie Zeeland heeft een aanvraag ingediend bij het ministerie van Sociale Zaken en Werkgelegenheid om mee te doen aan de landelijke lancering van de Mensenrechtenvlag.
Steekwoorden	december, mensenrechtenvlag, 10, mensenrechten, provincie, nederland, lancering, vlag, hijsen, mee
Bestandsextensie	PDF document, version 1.4, 4 pages
Taal bestand	nl

Resultaat metadataverrijking	
Originele bestandsnaam	GSnota ICL-IP_S_20032653_9_A_26978_tds.pdf
Titel	Iclip pva dcmr gs
Samenvatting	De provincie Zeeland (DCP) en de provincie Zuid-Holland (VVS) hebben in een vergadering van het ministerie van Veiligheid en Justitie een overkoepelend Plan van aanpak (PvA) van ICL-IP goedgekeurd.
Steekwoorden	iclip, pva, dcmr, gs, overtredingen, plan, bedrijf, situatie, toezicht, instanties
Bestandsextensie	PDF document, version 1.4, 4 pages
Taal bestand	nl

Resultaat metadataverrijking	
Originele bestandsnaam	20031924_S_20031924_1_A_50192_tds.pdf
Titel	Zeeland refinery mwe support
Samenvatting	Het ministerie van Economische Zaken en Waterstaat heeft de provincie Zeeland gevraagd om een 'letter of support' van de Provincie Zeeland voor het bedrijf Zeeland Refinery.
Steekwoorden	zeeland, refinery, mwe, support, letter, gs, waterstof, ', ', consequenties
Bestandsextensie	PDF document, version 1.4, 3 pages
Taal bestand	nl

Resultaat metadataverrijking	
Originele bestandsnaam	20032018_S_20032018_3_A_74348_tds.pdf
Titel	2020 begroting wijziging najaarsnota
Samenvatting	De effecten die betrekking hebben op het begrotingsjaar 2020 en 2021 zijn in het Statenvoorstel 10e wijziging begroting 2020, 2e wijziging begroting 2021 en 1e wijziging begroting 2021 vastgesteld.
Steekwoorden	2020, begroting, wijziging, najaarsnota, 2021, statenvoorstel, gs, zeeland, provincie, consequenties
Bestandsextensie	PDF document, version 1.4, 3 pages
Taal bestand	nl

Volgende versie

- Uitbreiden van functionaliteiten
- Ontwikkelen van een *package*
- Metadata verrijking

Blijf op de hoogte