text_extraction

October 9, 2023

[ ]:
```
# Copyright 2023 Google LLC
#
# Licensed under the Apache License, Version 2.0 (the "License");
# you may not use this file except in compliance with the License.
# You may obtain a copy of the License at
#
#     https://www.apache.org/licenses/LICENSE-2.0
#
# Unless required by applicable law or agreed to in writing, software
# distributed under the License is distributed on an "AS IS" BASIS,
# WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
# See the License for the specific language governing permissions and
# limitations under the License.
```

1 Text Extraction with Generative Models on Vertex AI

Run in Colab

View on GitHub

Open in Vertex AI Workbench

1.1 Overview

Text extraction is a process of extracting text from a document. This can be done manually or automatically. Manual text extraction is the process of reading the document and copying the text into a new document. Automatic text extraction is the process of using software to extract the text from the document.

Text extraction can be used for a variety of purposes. One common purpose is to convert documents into a machine-readable format. This can be useful for storing documents in a database or for processing documents with software. Another common purpose is to extract information from documents. This can be useful for finding specific information in a document or for summarizing the content of a document.

Large language models (LLMs) are good for text extraction because they are trained on massive datasets of text and code, which allows them to learn the relationships between words and phrases. They can also understand the context of text and generate text, which allows them to extract information that is not explicitly stated or fill in the gaps in text that is missing information. The answers from LLMs can also be further improved through methods like few-shot prompting.

Learn more about extraction prompts in the official documentation.

1.1.1 Objective

In this tutorial, you will learn how to use generative models to extract the information from text by working through the following examples: - Google Pixel technical specifications extraction - WiFi troubleshooting with constraints - Respond to inquiries in character - Converting an ingredients list to JSON format - Organizing the results of a text extraction

1.1.2 Costs

This tutorial uses billable components of Google Cloud:

- Vertex AI Generative AI Studio

Learn about Vertex AI pricing, and use the Pricing Calculator to generate a cost estimate based on your projected usage.

1.2 Getting Started

1.2.1 Install Vertex AI SDK

[ ]:
```
!pip install google-cloud-aiplatform --upgrade --user
```

Colab only: Uncomment the following cell to restart the kernel or use the button to restart the kernel. For Vertex AI Workbench you can restart the terminal using the button on top.

[ ]:
```
# # Automatically restart kernel after installs so that your environment can␣
↪access the new packages
# import IPython

# app = IPython.Application.instance()
# app.kernel.do_shutdown(True)
```

1.2.2 Authenticating your notebook environment

- If you are using **Colab** to run this notebook, uncomment the cell below and continue.
- If you are using **Vertex AI Workbench**, check out the setup instructions here.

[ ]:
```
# from google.colab import auth
# auth.authenticate_user()
```

1.2.3 Import libraries

Colab only: Uncomment the following cell to initialize the Vertex AI SDK. For Vertex AI Workbench, you don't need to run this.

[ ]:
```
# import vertexai

# PROJECT_ID = "[your-project-id]"  # @param {type:"string"}
# vertexai.init(project=PROJECT_ID, location="us-central1")
```

[ ]:
```
from vertexai.language_models import TextGenerationModel
```

1.2.4 Import models

[ ]:
```
generation_model = TextGenerationModel.from_pretrained("text-bison@001")
```

1.3 Text Extraction

1.3.1 Google Pixel technical specifications extraction

In this example, you try to extract the technical specifications of a Pixel phone from text in JSON format using the PaLM API.

[ ]:
```
prompt = """
Extract the technical specifications from the text below in JSON format.

Text: Google Nest WiFi, network speed up to 1200Mpbs, 2.4GHz and 5GHz␣
 ↪frequencies, WP3 protocol
JSON: {
  "product":"Google Nest WiFi",
  "speed":"1200Mpbs",
  "frequencies": ["2.4GHz", "5GHz"],
  "protocol":"WP3"
}

Text: Google Pixel 7, 5G network, 8GB RAM, Tensor G2 processor, 128GB of␣
 ↪storage, Lemongrass
JSON:
"""

print(
    generation_model.predict(
        prompt, temperature=0.2, max_output_tokens=1024, top_k=40, top_p=0.8
    ).text
)
```

1.3.2 WiFi troubleshooting with constraints

In this example, you ask the generative model to answer a question about troubleshooting a Google WiFi router based on the description of the different status lights on the router. The model will only be able to respond with the text that was provided, which helps to prevent it from generating potentially harmful or incorrect answers. Here is how you can do this using the PaLM API.

[ ]:
```
prompt = """
Answer the question using the text below. Respond with only the text provided.
Question: What should I do to fix my disconnected WiFi? The light on my Google␣
 ↪WiFi router is yellow and blinking slowly.
```

```
Text:
Color: No light
What it means: Router has no power or the light was dimmed in the app.
What to do:
Check that the power cable is properly connected to your router and to a␣
 ↪working wall outlet.
If your device is already set up and the light appears off, check your light␣
 ↪brightness settings in the app.
If there's still no light, contact WiFi customer support.

Color: Solid white, no light, solid white
What it means: Device is booting up.
What to do:
Wait for the device to boot up. This takes about a minute. When it's done, it␣
 ↪will slowly pulse white, letting you know it's ready for setup.

Color: Slow-pulsing white
What it means: Device is ready for set up.
What to do:
Use the Google Home app to set up your router.

Color: Solid white
What it means: Router is online and all is well.
What to do:
You're online. Enjoy!

Color: Slowly pulsing yellow
What it means: There is a network error.
What to do:
Check that the Ethernet cable is connected to both your router and your modem␣
 ↪and both devices are turned on. You might need to unplug and plug in each␣
 ↪device again.

Color: Fast blinking yellow
What it means: You are holding down the reset button and are factory resetting␣
 ↪this device.
What to do:
If you keep holding down the reset button, after about 12 seconds, the light␣
 ↪will turn solid yellow. Once it is solid yellow, let go of the factory reset␣
 ↪button.

Color: Solid yellow
What it means: Router is factory resetting.
What to do:
This can take up to 10 minutes. When it's done, the device will reset itself␣
 ↪and start pulsing white, letting you know it's ready for setup.
```

```
Image Solid red light Solid red Something is wrong. Critical failure. Factory␣
 ↪reset the router. If the light stays red, contact WiFi customer support.
"""

print(
    generation_model.predict(
        prompt, temperature=0.2, max_output_tokens=256, top_k=1, top_p=0.8
    ).text
)
```

1.3.3 Respond to inquiries in character

Now, you instruct the generative model to pretend to be Klara, a person. You will also tell the model about Klara's personality traits. Then, you will ask the model to answer a question as Klara would answer it.

[ ]:
```
prompt = """
You are Klara.
Klara is an investment manager.
Klara only answers if Klara is sure it is correct.
Klara answers the user question based on the summaries of the pages below.
Klara outputs the Reference ID where Klara found the answer for each sentence␣
 ↪in the answer.
Klara also summarizes the part where the information is found.
```

```
Summaries of the pages: ['Reference ID 1. UBS wants to be seen as a global bank␣
 ↪with Swiss roots, not just a European bank . New board chairman Colm␣
 ↪Kelleher and CEO Ralph Hamers have held a series of meetings with␣
 ↪influential U.S. fund managers to increase their stakes in the bank . UBS is␣
 ↪one of the most valuable European banks with a price-to-book ratio of 1 .',␣
 ↪'Reference ID 2. Credit Suisse sells 30 percent stake in Swiss asset manager␣
 ↪Energy Infrastructure Partners . EIP specializes in long-term equity␣
 ↪investments for large-scale renewable and energy infrastructure assets . The␣
 ↪transaction may be part of an ongoing effort by the Swiss bank to further␣
 ↪close the capital gap .', 'Reference ID 3. Saudi Arabias Crown Prince␣
 ↪Mohammed bin Salman is preparing to invest in Credit Suisse Group AGs␣
 ↪investment bank . Prince Mohammed would inject about $500 million in the␣
 ↪spinoff of CS First Boston . Other investors could include former Barclays␣
 ↪chief Bob Diamonds Atlas Merchant Capital . Saudi National Bank already has␣
 ↪a 9. 9 percent stake in the troubled Swiss institution .', 'Reference ID 4.␣
 ↪US-based Apollo Global Management is among a group of financial firms in␣
 ↪talks with Credit Suisse about a stake in the revamped investment bank . The␣
 ↪Wall Street Journal has reported that Apollo is also said to be interested␣
 ↪in investing in CS First Boston . The investment banks investment banking␣
 ↪division is set to be spun off into a new unit .', 'Reference ID 5.␣
 ↪Geneva-based private bank Pictet has signed up a new team in its␣
 ↪billion-dollar private market investment business . Edmund Buckley, Nikolaus␣
 ↪Hubmann, Sean Howard, Jan Dreesen and Hugo Hickson will work in the area of␣
 ↪direct investments in private equity in the future . The commitment of␣
 ↪Buckleys team can be seen as a coup for Pictet partner Elif Aktuğ .',␣
 ↪'Reference ID 6. UBS wants to be seen as a global bank with Swiss roots, not␣
 ↪just a European bank . New board chairman Colm Kelleher and CEO Ralph Hamers␣
 ↪have held a series of meetings with influential U.S. fund managers to␣
 ↪increase their stakes in the bank . UBS is one of the most valuable European␣
 ↪banks with a price-to-book ratio of 1 .', 'Reference ID 7. US-based Apollo␣
 ↪Global Management is among a group of financial firms in talks with Credit␣
 ↪Suisse about a stake in the revamped investment bank . The Wall Street␣
 ↪Journal has reported that Apollo is also said to be interested in investing␣
 ↪in CS First Boston . The investment banks investment banking division is set␣
 ↪to be spun off into a new unit .', 'Reference ID 8. Luxembourg-based Apex␣
 ↪subsidiary European Depositary Bank (EDB) and securitization specialist␣
 ↪Gentwo enter into partnership . EDB and Gentwo will offer paying agent and␣
 ↪banking services to third-party investors around the world . Gentwo develops␣
 ↪platforms for asset managers, banks, family offices, and venture capitalists␣
 ↪.', 'Reference ID 9. SNB has been pursuing plans to acquire stakes in␣
 ↪European and American financial institutions for some time . Credit Suisse,␣
 ↪Julius Baer, Standard Chartered and the Asian DBS Group have also been␣
 ↪mentioned as possible targets .', 'Reference ID 10. Zuercher Kantonalbank is␣
 ↪aiming to expand its private banking business, including abroad . CEO Urs␣
 ↪Baumann sees Credit Suisse as a reliable partner for the state-owned bank .␣
 ↪Baumann will continue to focus on reducing the banks dependence on the␣
 ↪interest differential business in an attempt to diversify its business .']
```

```
User question: Are global banks investing into ESG initiatives?
Klara's answer:
"""

print(
    generation_model.predict(
        prompt, temperature=0.2, max_output_tokens=256, top_k=40, top_p=0.8
    ).text
)
```

1.3.4 Converting an ingredients list to JSON format

Suppose that you want to itemize ingredients in recipes to enter into a database, which requires a well-formatted output like JSON. This can be done using a generative model in the following way:

[ ]:
```
prompt = """
Extract the ingredients from the following recipe. Return the ingredients in␣
 ↪JSON format with keys: ingredient, quantity, type.

Ingredients:
* 1 tablespoon olive oil
* 1 onion, chopped
* 2 carrots, chopped
* 2 celery stalks, chopped
* 1 teaspoon ground cumin
* 1/2 teaspoon ground coriander
* 1/4 teaspoon turmeric powder
* 1/4 teaspoon cayenne pepper (optional)
* Salt and pepper to taste
* 1 (15 ounce) can black beans, rinsed and drained
* 1 (15 ounce) can kidney beans, rinsed and drained
* 1 (14.5 ounce) can diced tomatoes, undrained
* 1 (10 ounce) can diced tomatoes with green chilies, undrained
* 4 cups vegetable broth
* 1 cup chopped fresh cilantro
"""

print(
    generation_model.predict(
        prompt, temperature=0.2, max_output_tokens=1024, top_k=40, top_p=0.8
    ).text
)
```

1.3.5 Organizing the results of a text extraction

In this section, you extract the information you want from a block of text and organize it in a structured way, such as separating it by commas. Here you use few-shot prompting to guide the

model to format your outputs to be separated by commas.

```
[ ]: prompt = """
Message: Rachel Green (Jennifer Aniston), a sheltered but friendly woman, flees␣
 ↪her wedding day and wealthy yet unfulfilling life and finds childhood friend␣
 ↪Monica Geller (Courteney Cox), a tightly wound but caring chef.
Rachel becomes a waitress at West Village coffee house Central Perk after she␣
 ↪moves into Monica\'s apartment above Central Perk and joins Monica\'s group␣
 ↪of single friends in their mid-20s:
previous roommate Phoebe Buffay (Lisa Kudrow), an odd masseuse and musician;␣
 ↪neighbor Joey Tribbiani (Matt LeBlanc), a dim-witted yet loyal struggling␣
 ↪actor; Joey\'s roommate Chandler Bing (Matthew Perry),
a sarcastic, self-deprecating data processor; and Monica\'s older brother and␣
 ↪Chandler\'s college roommate Ross Geller (David Schwimmer), a sweet-natured␣
 ↪but insecure paleontologist.

Extract the characters and the actors who played them from above message:
Rachel Green - Jennifer Aniston, Monica Geller - Courteney Cox, Phoebe Buffay -␣
 ↪Lisa Kudrow, Joey Tribbiani - Matt LeBlanc, Chandler Bing - Matthew Perry,␣
 ↪Ross Geller - David Schwimmer

Message: Games such as chess, poker, Go, and many video games have always been␣
 ↪fertile ground for AI research. Diplomacy is a seven-player game of␣
 ↪negotiation and alliance formation, played on an old map of Europe␣
 ↪partitioned
into provinces, where each player controls multiple units (rules of Diplomacy).␣
 ↪In the standard version of the game, called Press Diplomacy, each turn␣
 ↪includes a negotiation phase, after which all players reveal their
chosen moves simultaneously. The heart of Diplomacy is the negotiation phase,␣
 ↪where players try to agree on their next moves. For example, one unit may␣
 ↪support another unit, allowing it to overcome resistance by other units,
as illustrated here: Computational approaches to Diplomacy have been researched␣
 ↪since the 1980s, many of which were explored on a simpler version of the␣
 ↪game called No-Press Diplomacy, where strategic communication between
players is not allowed. Researchers have also proposed computer-friendly␣
 ↪negotiation protocols, sometimes called␣
 ↪\342\200\234Restricted-Press\342\200\235.

Extract the deinition of Diplomacy:
A seven-player game of negotiation and alliance formation


Message: Back in 2016, when we weren\'t using simulation and were using a small␣
 ↪lab-configuration of industrial robots to learn how to grasp small objects␣
 ↪like toys, keys and everyday household items, it took the equivalent of
```

```
four months for one robot to learn how to perform a simple grasp with a 75%%␣
 ↪success rate. Today, a single robot learns how to perform a complex task␣
 ↪such as opening doors with a 90%% success rate with less than a day
of real-world learning. Even more excitingly, we\'ve shown that we can build on␣
 ↪the algorithms and learnings from door opening and apply them to a new task:␣
 ↪straightening up chairs in our cafes. This progress gives us
hope that our moonshot for building general purpose learning robots might just␣
 ↪be possible.

Extract the success rates of the robots in 2016 and today, respectively:
75%, 90%

Message: CapitalG was founded a decade ago to empower entrepreneurs with␣
 ↪Alphabet and Google\'s unparalleled expertise in growth.
We are privileged to share the lessons learned from helping to scale Google,␣
 ↪Stripe, Airbnb, CrowdStrike, Databricks, and Zscaler with the next wave of␣
 ↪generational tech companies-perhaps including yours.
Alphabet is our sole LP and provides patient, long-term capital. As an␣
 ↪independent growth fund, our priorities align with our entrepreneurs.␣
 ↪CapitalG companies have achieved product-market fit and are ready to scale.␣
 ↪We maintain a small, concentrated portfolio so every company receives␣
 ↪substantial capital and hands-on support.

Extract the companies funded by CapitalG:
"""

print(
    generation_model.predict(
        prompt, temperature=0.2, max_output_tokens=256, top_k=1, top_p=0.8
    ).text
)
```

As you can see in the output above, based on the few-shot prompt, you should see the names of companies funded by CapitalG.