



Wydział Telekomunikacji, Informatyki i Elektrotechniki
Politechniki Bydgoskiej
im. Jana i Jędrzeja Śniadeckich

Chmury obliczeniowe: Kategoryzacja dokumentów AI

Przedmiot: Chmury obliczeniowe

Prowadzący: dr inż. Michał Kruczkowski

Autorzy sprawozdania: Anna Sarnecka, Kajetan Romanowski

Spis treści

1	Wstęp.....	3
2	Źródło danych.....	4
3	Model AI Builder	5
4	Przepływ kategoryzacji dokumentów.....	8
5	Aplikacja i działanie rozwiązania	12
6	Wnioski:	15

1 WSTĘP





Celem projektu było stworzenie rozwiązania opartego na chmurach obliczeniowych. W celu zrealizowania tematu została stworzona aplikacja pozwalająca na kategoryzację dokumentów przy pomocy sztucznej inteligencji, dodatkowo aplikacja oraz przepływy zostały wykonane z użyciem narzędzi z pakietu *Microsoft Power Platform*, które jest rozwiązaniem chmurowym typu *SaaS* – Software as a Service. Użycie narzędzi *Power Platform* pozwoliło na przyspieszenie procesu rozwoju aplikacji i skupienie się na samej kategoryzacji dokumentów.

2 ŹRÓDŁO DANYCH

Jako źródło danych zarówno dla wyników z automatyzacji jak i samych plików posłużył SharePoint. Dzięki jego możliwościom, w obrębie jednej strony (SharePoint site) można przechowywać rekordy na liście oraz pliki w bibliotece dokumentów.

Użyto domyślnej biblioteki dokumentów utworzonej strony w celu przechowywania kopii przesłanych przez użytkownika plików:

Documents

 Name	Modified	Modified By
 867377007_UserManual_cb625366-372b-4...	May 25	annsar000@o365.student.
 Balance Sheet.txt	May 24	annsar000@o365.student.
 balance-sheet.pdf	May 24	annsar000@o365.student.

Lista Kategoryzacja została utworzona w celu przechowywania wyników działania automatyzacji. Zawiera kolumny:







Title – nazwa przetwarzanego pliku

Kategoria – kategoria przydzielona przez model AI Builder

Link – URL kopii pliku utworzonej w bibliotece dokumentów strony

Podsumowanie – streszczenie dokumentu utworzone przez gotowy model AI

Created – data utworzenia rekordu na liście

 Kategoryzacja ☆				
 Title	 Kategoria	 Link	 Podsumowanie	 Created
sustainability-report-2023.pdf	Other	https://utpedupl.sh...report-2023.pdf	The given text is a title and subtitle for the Electrolux Group's Sustainability Report for the year 2023.	May 24
867377007_UserManual_cb625366-372b-4d30-b4c1-ba2e7041a92c.pdf	IT	https://utpedupl.sh...372b-4d30-b4c1-ba2e7041a92c.pdf	The given text is a user manual for an Electrolux oven model EOF6H46X2. To access the manual, users can visit the Electrolux website and register their product.	May 24
latitude-5540-rg-forwindows-en-us.pdf	IT	https://utpedupl.sh...5540-rg-forwindows-en-us.pdf	The given text is a re-imaging guide for the Dell Latitude 5540 laptop running Windows. It provides instructions on how to re-image the laptop and includes important regulatory information such as the model and type. The guide was published by Dell Technologies in March 2023, with a revision number of A00.	May 24
non-disclosure-agreement-template.pdf	Legal	https://utpedupl.sh...disclosure-agreement-template.pdf	The given text is a Non-Disclosure Agreement (NDA) between The National Archives (the "Disclosing Party") and another party (the "Receiving Party"). The purpose of the agreement is to prevent the unauthorized disclosure of Sensitive Information, as defined in the agreement, in accordance with HMG's Security Policy Framework.	May 24

3 MODEL AI BUILDER

W celach projektowych wybrano model AI Builder typu *Category Classification – Klasyfikacja w kategorii* z możliwością samodzielnego wytrenowania.

My AI capabilitiesPromptsAI models

Name	Owner	Permission	Status	Last modified	Last trained	Model type	Expiration
Category Classification 24.05.2024, 18:57:17	annsa000 #	Owner	Published	1 mo ago	1 mo ago	Category Classification	-

See more promptsSee more AI models

Przed konfiguracją modelu należało przygotować dane treningowe. Z racji iż model polega na analizie tekstu, a nie pliku, należało przygotować tabelę z danymi. Microsoft wymusza w tym przypadku Dataverse jako źródło danych treningowych dla modelu. Utworzono tabelę zawierającą tekst z wygenerowanych i znalezionych dokumentów oraz przypisaną kategorię:

Kategoryzacja AI columns and dataUpdate forms and viewsEdit

Name*Text

+18 more

Finance	UNITED STATES SECURITIES AND EXCHANGE COMMISSION Washington, D.C. 20549 FORM 10-K (Mark One)ANNUAL REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXC
Finance	Company: XYZ CorpDate: December 31, 2023ASSETSCurrent Assets: Cash and Cash Equivalents: \$100,000 Accounts Receivable: \$50,000 Inventory: \$20,000Total Current Assets: \$170,000N
Finance	Master your finances with business.govt.nz's Strategic finance section This balance sheet sets out sample figures for Sam, a fictional painter. It shows two years, before and after he improve
Finance	Balance sheet Current year Last year Assets Current assets: Cash (bank account) 7,500 3,000 Inventory (paints) 500 1,000 Accounts receivable (customers still to pay) 20,000 36,000 Pre-paid
Finance	Assets Current Assets 1000 Cash 1010 Checking 583,961 1020 Savings 224,600 1030 Petty Cash 89,840 Total Cash 898,402 1100 Accounts Receivable 3,593,607 1200 W.
Finance	Springfield Psychological Services 2004 2003 2004 2003 Assets Liabilities Current assets: Current liabilities: Cash.....\$ 12,597 \$ 8,173 Note payable.....\$ 4,200 \$ 4,75.
Finance	Note: This is just an example of the format 1. Your Financial Statement Account titles may differ. 2. Your chart of accounts will likely differ in the c
Finance	Company: XYZ CorpPeriod: Year Ended December 31, 2023Cash Flows from Operating Activities: Net Income: \$130,000 Adjustments for Non-Cash Items: Depreciation: \$10,000 Changes
Finance	Company: XYZ CorpPeriod: Year Ended December 31, 2023Revenue: \$500,000Cost of Goods Sold: \$200,000Gross Profit: \$300,000Operating Expenses: Salaries: \$100,000 Rent: \$20,000 Util
Finance	Springfield Psychological Services 2003 2004 Sales Client Service Revenue.....\$ 256,651 \$ 279,156 Book sales..... 3,725 3,410 Professional Consultation.....

Po przygotowaniu danych treningowych zgodnie z zaleceniami zawartymi w dokumentacji Microsoft można było przystąpić do konfiguracji modelu.

Pierwszym krokiem było wskazanie tabeli oraz kolumn z tekstem i kategoriami (tagami). Oraz mapowanie kategorii.

Select textKategoryzacja AI > Text

Select tagsKategoryzacja AI > Name

Review tags

Select text language

Model summary

Select tags

Category Classification 25.05.2024, 01:07:32Save and close

Select your tags

Kategoryzacja AI > Name

Tag separator

We have automatically detected your separator and selected it for you.

No separator (one tag per text snippet)

Finance IT Legal Sales & Marketing

Semicolon

Finance IT Legal Sales & Marketing

Comma

Finance IT Legal Sales & Marketing

Tab

Finance IT Legal Sales & Marketing

Quick tips

What table should I select?

Select a table with columns for your text and tags. Text data with associated tags will be used to train the model.

Learn more

Which tag separator should I choose?

Choosing the right tag separator will properly isolate the tags. If you only have one tag for each text snippet, choose No separator.

Correct separation

Tag one Tag two Tag three

Incorrect separation

Tag one, Tag two, Tag three

Następnie należało zweryfikować, czy dane zostały zmapowane poprawnie do modelu:

✓ Select text
Kategoryzacja AI > Text

✓ Select tags
Kategoryzacja AI > Name

● Review tags

○ Select text language

○ Model summary

Review your text and tags

Text	Tags
UNITED STATES SECURITIES AND EXCHANGE COMMISSION Washington, D.C. 20549 FORM 10-K (Mark One)ANNUAL REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934 For the fiscal year ended September 24, 2022 orTRANSITION REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934 For the transition period from to Commission File Number: 001-36743 Apple Inc. (Exact name of Registrant as specified in its charter) California (State or other jurisdiction of incorporation or organization) 94-2404110 (I.R.S. Employer Identification No.) One Apple Park Way Cupertino, California (Address of principal executive offices) (Registrant's telephone number, including area code) Securities registered pursuant to Section 12(b) of the Act: Title of each class Common Stock, \$0.00001 par value per share 1.000% Notes due 2022 1.375% Notes due 2024 0.000% Notes due 2025 0.875% Notes due 2025 1.625% Notes due 2026 2.000% Notes due 2027 1.375% Notes due 2029 3.050% Notes due 2029 0.500% Notes due 2031 3.600% Notes due 2042	Finance
Company: XYZ CorpDate: December 31, 2023ASSETSCurrent Assets: Cash and Cash Equivalents: \$100,000 Accounts Receivable: \$50,000 Inventory: \$20,000Total Current Assets: \$170,000Non-Current Assets: Property, Plant, and Equipment: \$200,000 Intangible Assets: \$30,000Total Non-Current Assets: \$230,000TOTAL ASSETS: \$400,000LIABILITIESCurrent Liabilities: Accounts Payable: \$30,000 Short-Term Debt: \$10,000Total Current Liabilities: \$40,000Non-Current Liabilities: Long-Term Debt: \$100,000Total Non-Current Liabilities: \$100,000TOTAL LIABILITIES: \$140,000EQUITYShareholder Equity: \$260,000TOTAL LIABILITIES AND EQUITY: \$400,000	Finance
Master your finances with business.govt.nz's Strategic finance section This balance sheet sets out sample figures for Sam, a	Finance

Na końcu wybrano język:

✓ Select text
Kategoryzacja AI > Text

✓ Select tags
Kategoryzacja AI > Name

✓ Review tags
Done

● Select text language

○ Model summary

Select your text language

Language

English

English

German

French

Italian

Portuguese

Spanish

Po zweryfikowaniu podsumowania modelu potwierdzono rozpoczęcie jego trenowania.

Model summary


Review your model's details below. If something is missing you can go back to the previous steps. If everything looks good, select Train. Learn more about training. [Learn more about training](#)

Overview

Owner
annsar000 #

Model type

Information to extract


Data source
 Dataaverse

Text language

English

Input

[Klasyfikacja AI > Name](#)



Your model is training

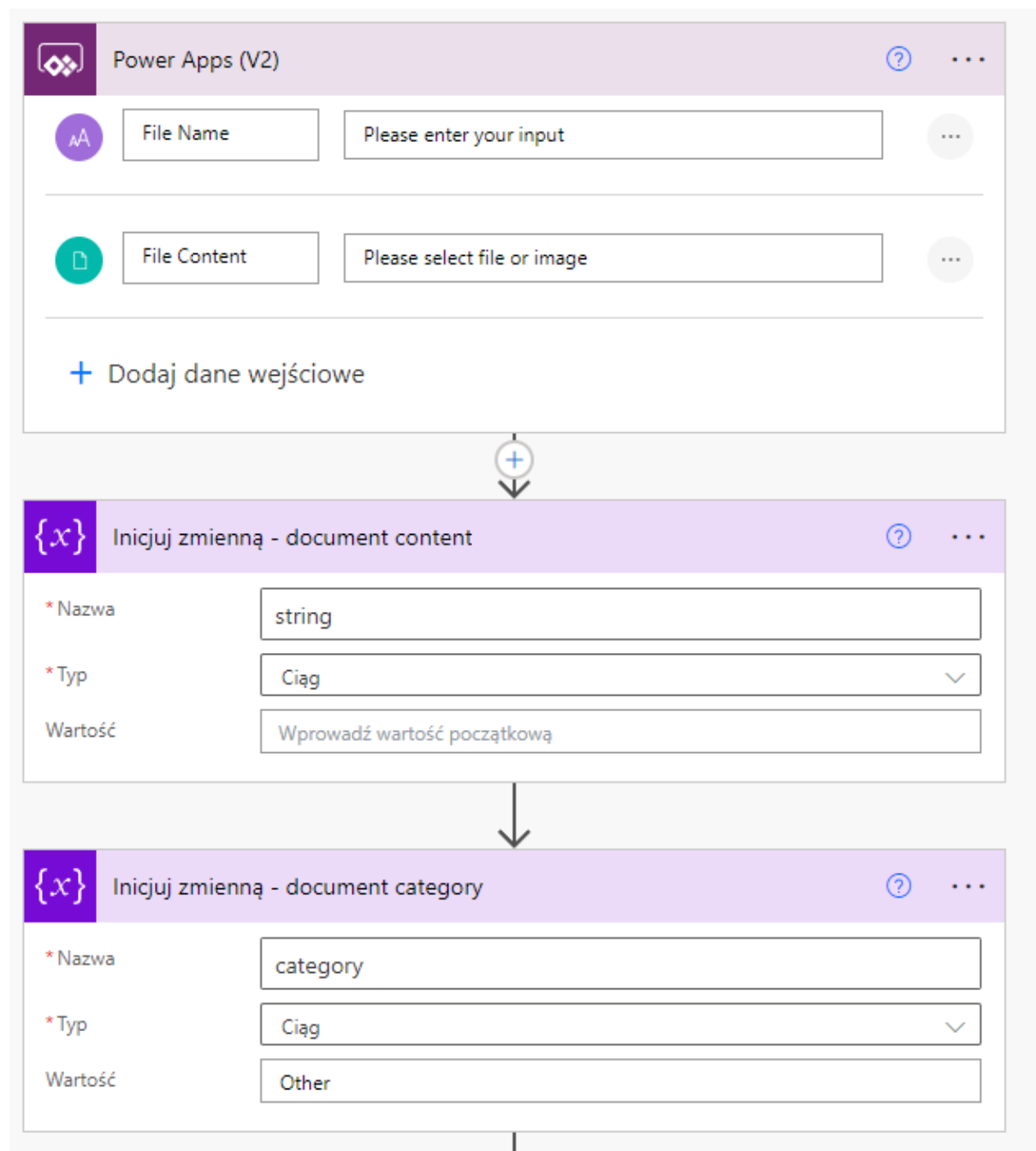
This may take a few minutes depending on the size of your training data.
You can close this window and come back later.

[Go to models](#)

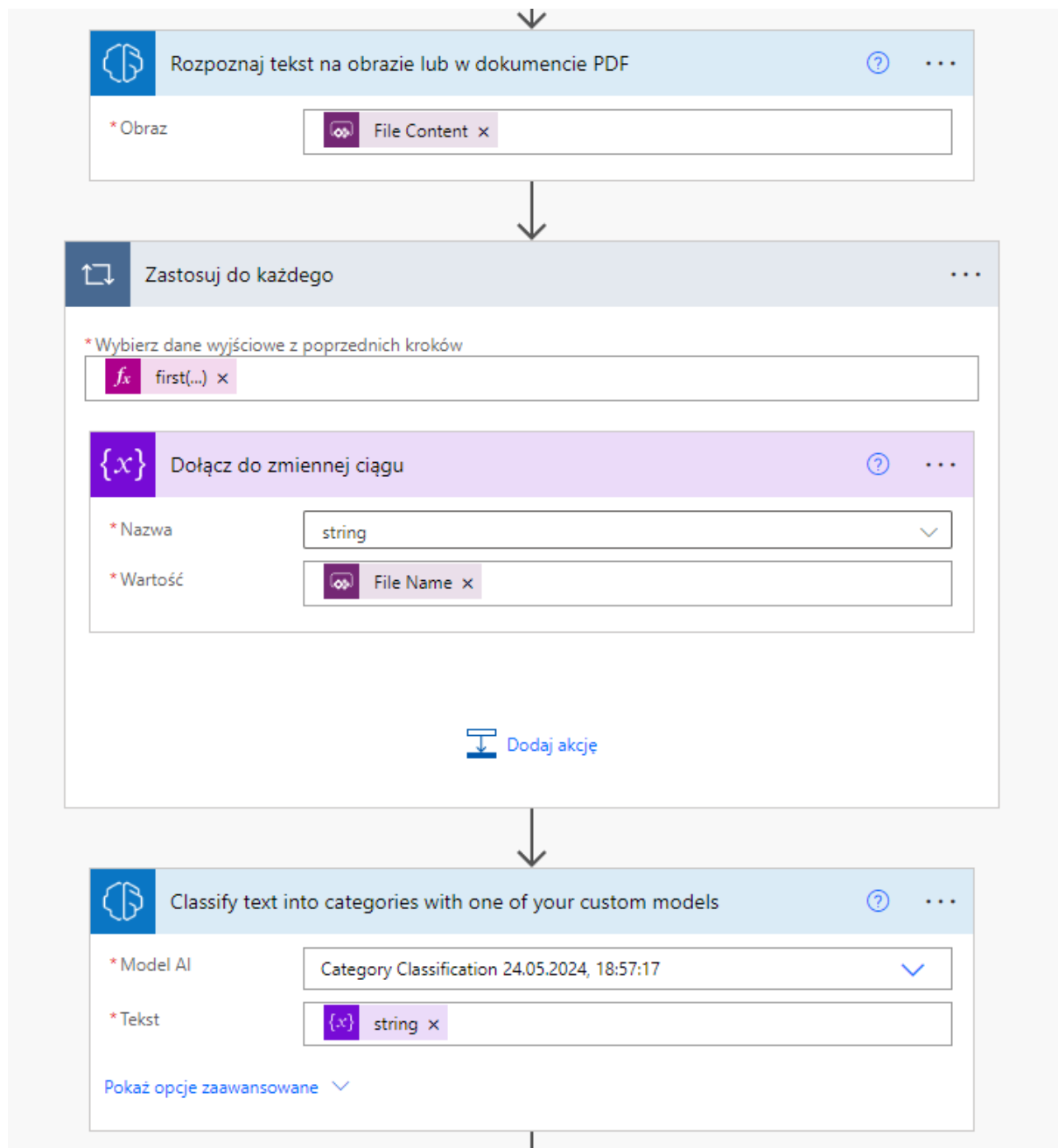
Aby móc używać modelu w przepływach Power Automate i aplikacjach Power Apps należy przetrenowany model opublikować.

4 PRZEPŁYW KATEGORYZACJI DOKUMENTÓW

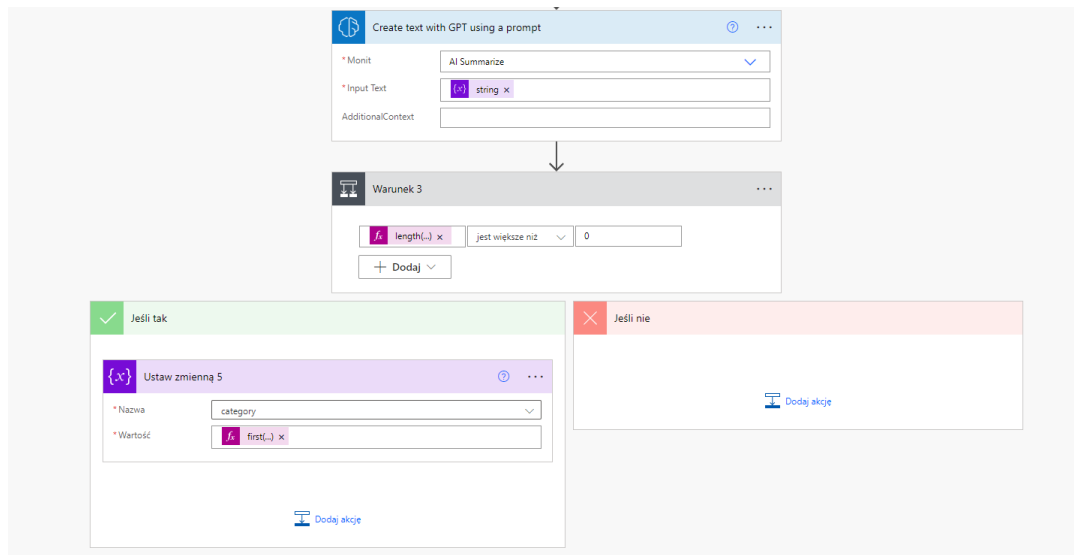
Pierwsze trzy akcje obejmują odebranie parametrów wejściowych czyli nazwy dokumentu oraz jego zawartości. Następne dwie akcje inicjują zmienne typu string, które będą przechowywać informacje o treści dokumentu oraz przydzielonej mu kategorii.



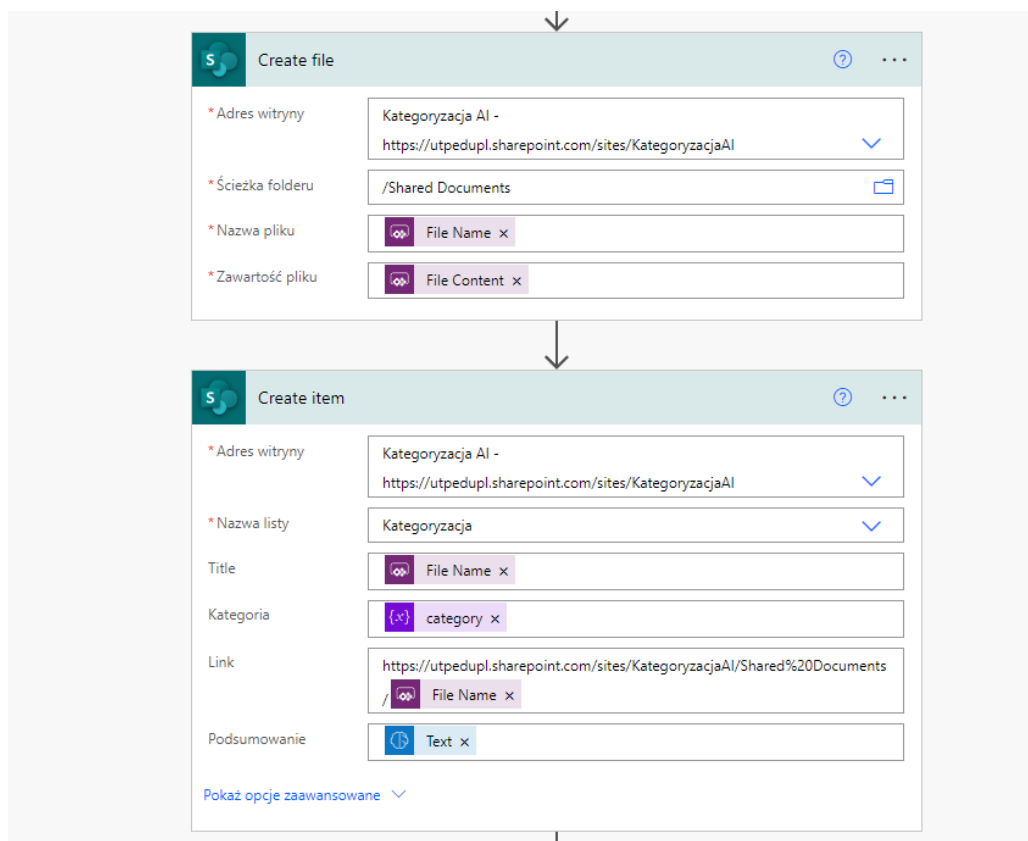
Kolejną akcją jest rozpoznanie tekstu na obrazie lub w dokumencie PDF, jest to model AI wytrenowany do odczytywania tekstu z obrazu lub pliku PDF. Akcja ta na wyjściu przekazuje tablicę z odczytanym tekstem podzielonym na linie, dlatego w kolejnym kroku użyta jest akcja *Zastosuj do każdego* czyli odpowiednik pętli *for each*. Iterując wers po wersji tworzony jest pojedynczy ciąg znaków za pomocą akcji *Dołącz do zmiennej ciągu*. Następny krok to użycie kolejnego modelu sztucznej inteligencji, tym razem w celu przeprowadzenia kategoryzacji przesłanego dokumentu. Parametrami wejściowymi tej akcji są: uprzednio wytrenowany model oraz treść do kategoryzacji.



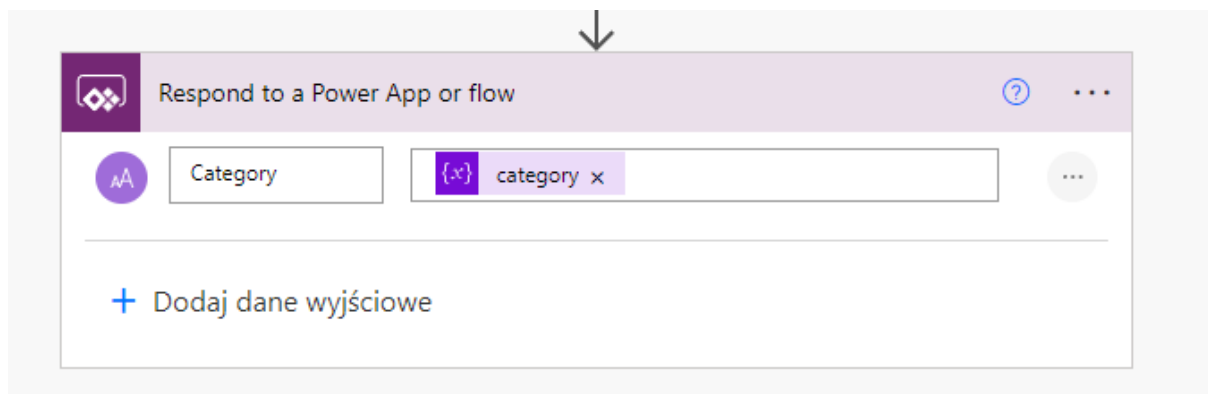
Ostatnim akcją w przepływie związaną ze sztuczną inteligencją jest stworzenie tekstu na podstawie promptów w oparciu o GPT. Parametrami wejściowymi tej akcji są: do wyboru jeden z przygotowanych modeli, tekst wejściowy oraz dodatkowy kontekst. Wybrany modelem jest *AI Sumarize*, który pozwala na stworzenie podsumowania tekstu na podstawie jego treści. Kolejną akcją jest warunek, w tej akcji sprawdzane jest czy model kategoryzujący dokumenty znalazł podobieństwa z jedną z wytrenowanych kategorii. Jeśli co najmniej jedna kategoria została przydzielona to dokumentowi zostaje przydzielona kategoria z największym % podobieństwa.



Następnymi krokami przepływu są: utworzenie kopii dokumentu na platformie SharePoint, dodanie do listy SharePoint wpisu o przeprowadzonej kategoryzacji wraz z odnośnikiem do utworzonego dokumentu, podsumowaniem oraz kategorią.



Ostatnim krokiem jest zwrócenie odpowiedzi do aplikacji, w której przekazana zostaje przydzielona kategoria.



Przepływ po zakończeniu pracy zwraca informację do aplikacji, która prezentuje wynik użytkownikowi, w aplikacji zostaną wyświetlone takie informacje jak nazwa dokumentu, przydzielona kategoria, streszczenie dokumentu oraz URL do kopii dokumentu na platformie SharePoint.

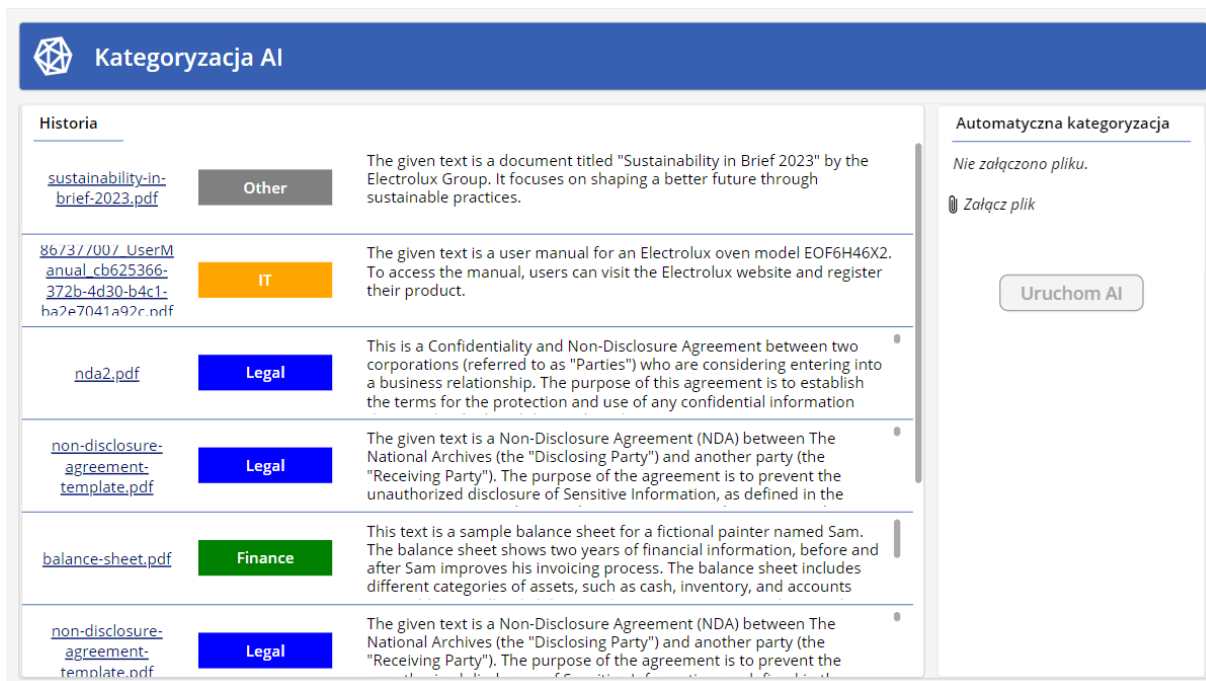
5 APLIKACJA I DZIAŁANIE ROZWIĄZANIA

Aby udostępnić potencjalnym użytkownikom przyjazny interfejs umożliwiający korzystanie z automatyzacji, przygotowano aplikację z wykorzystaniem narzędzia Power Apps Canvas.

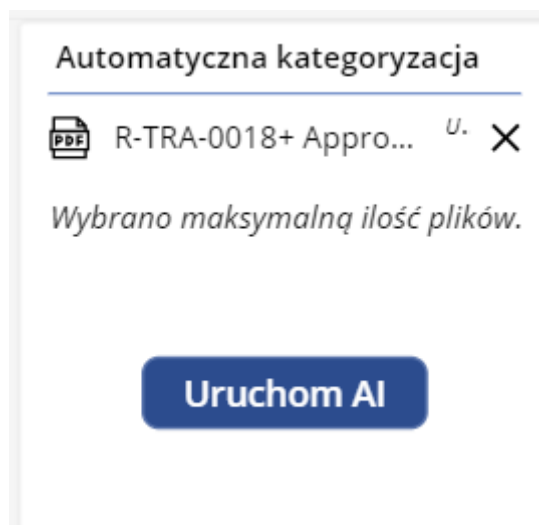
Zaprojektowano ekran początkowy (powitalny).



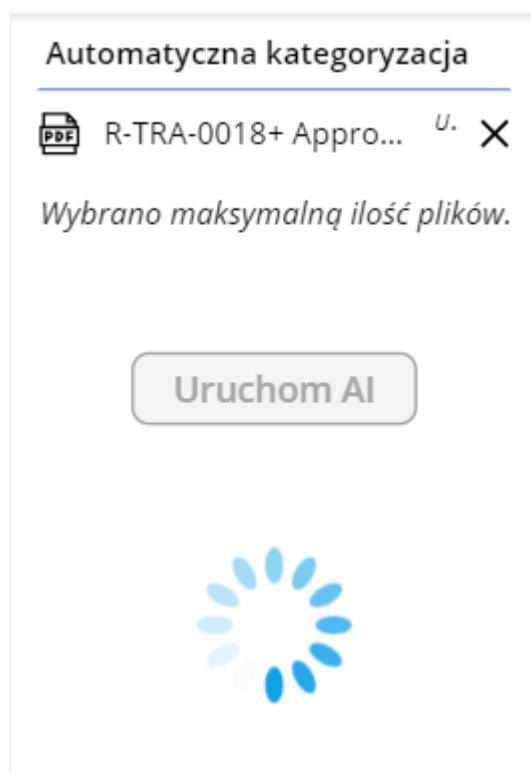
Po naciśnięciu przycisku użytkownikowi prezentuje się ekran zawierający listę przetworzonych plików. Widoczność opiera się na dostępach zdefiniowanych w SharePoint.



Użytkownik może rozpocząć proces poprzez wybranie pliku ze swojego urządzenia, a następnie naciśnięcie przycisku „Uruchom AI”




W celach testowych wybrano plik zawierający potwierdzenie akceptacji wniosku o podróż służbową. Według przesłanych danych testowy pliki zawierające wnioski powinny być kategoryzowane jako *Legal*. Po naciśnięciu przycisku rozpocznie się działanie utworzonego wcześniej przepływu Power Automate.



Jeśli przepływ zakończy się sukcesem, w aplikacji ukaże się jego wynik:

Automatyczna kategoryzacja

Nie załączono pliku.

 Załącz plik

Uruchom AI

Kategoria przesłanego pliku to:
Legal

Testowy plik został poprawnie skategoryzowany. Lista w aplikacji także odświeży się automatycznie, użytkownik na niej może zobaczyć podsumowanie przesłanego dokumentu:

R-TRA-0018+ Approve Travel Request for PredicaDev01.pdf	Legal	On June 18, 2024, the Approvals Report shows that the travel request for PredicaDev01 was approved under reference R-TRA-0018. The request was made by PredicaDev01 and was approved without any further response.
--	--------------	--

Test zakończył się powodzeniem.

6 WNIOSKI:

- Modele AI zużywają kredyty, które odnawiają się co miesiąc
- Wytrenowany model kategoryzujący wykazał się wysoką skutecznością pomimo niewielkiej ilości danych uczących
- Użycie Power Apps znacznie przyspiesza proces rozwoju aplikacji
- Wytrenowanie modelu wymagało dostarczenia minimum 10 dokumentów z każdej kategorii