AI Classifier

Rob Wells and his pal Claude

04-18-2025

This exercise is an introduction to using AI to classify content. We're using the ellmer package: https://ellmer.tidyverse.org/

Background on ellmer

ellmer supports a wide variety of model providers:

Anthropic's Claude: chat_claude(). AWS Bedrock: chat_bedrock(). Azure OpenAI: chat_azure(). Databricks: chat_databricks(). DeepSeek: chat_deepseek(). GitHub model marketplace: chat_github(). Google Gemini: chat_gemini(). Groq: chat_groq(). Ollama: chat_ollama(). OpenAI: chat_openai(). OpenRouter: chat_openrouter(). perplexity.ai: chat_perplexity(). Snowflake Cortex: chat_snowflake() and chat_cortex_analyst(). VLLM: chat_vllm()

Google's chat_gemini() is great for large prompts because it has a much larger context window than other models. It allows up to 1 million tokens (about 8 average length English novels) and has a generous free tier. By contrast, other LLMS allow 32,000 tokens or 128,000 tokens in their context windows. One warning: your data is used to improve the model.

For Gemini models, a token is equivalent to about 4 characters. 100 tokens is equal to about 60-80 English words.

And so we will use chat_gemini()

API Key

Google Studio, create API Key

–You may need to use your personal gmail account. –Do not activate billing. We are just using the free tier of Gemini –Create API key here: https://aistudio.google.com/app/apikey

Important! DO NOT STORE YOUR API Key in GitHub!

Important! after verifying that the chat_gemini is working, I would delete the API key from Line 37 so you don't mistakenly save it to GitHub.

Tokens

On average an English word needs ~1.5 tokens so a page might require 375-400 tokens.

For Gemini models, a token is equivalent to about 4 characters. 100 tokens is equal to about 60-80 English words. A 5,000 word essay would be about 8,000 tokens (1.6 tokens per word. . .).

From the ellmer documentation:

An LLM is a model, and like all models needs some way to represent its inputs numerically. For LLMs, that means we need some way to convert words to numbers. This is the goal of the tokenizer. For example, using the GPT 4o tokenizer, the string "When was R created?" is converted to 5 tokens: 5958 ("When"), 673 (" was"), 460 (" R"), 5371 (" created"), 30 ("?"). As you can see, many simple strings can be represented by a single token. But more complex strings require multiple tokens. For example, the string "counterrevolutionary" requires 4 tokens: 32128 ("counter"), 264 ("re"), 9477 ("volution"), 815 ("ary"). (You can see how various strings are tokenized at http://tiktokenizer.vercel.app/).

Price

State of the art models (like GPT-4o or Claude 3.5 sonnet) cost $2-3 per million input tokens, and $10-15 per million output tokens. Cheaper models can cost much less, e.g. GPT-4o mini costs $0.15 per million input tokens and $0.60 per million output tokens.

We are using the free tier of Gemini. Do not activate billing.

Check token usage

In ellmer, you can see how many tokens a conversations has used by printing it, and you can see total usage for a session with token_usage()

Load a very small subset: 10 articles

Industrial strength cleaning and processing

–Articles put into a single string of text, no punctuation or spaces

The AI prompt

#Process first response –We're taking the chat response and putting it into a dataframe

validate results

process the results to a single file, separated by file name

Now, open a Google Sheet, import nixon_ai_verification.csv, read the articles and rate the responses

Second AI prompt

#Process first response –We're taking the chat response and putting it into a dataframe

validate results

–democrat critique

process the results to a single file, separated by file name

Now, open a Google Sheet, import dem_ai_verification.csv, read the articles and rate the responses

A better AI prompt

–Spoiler alert. –This prompt seeks to improve on the Democrat_critique : " If an article contains two or more mentions of Democrats or the Democratic Party and if adjectives modifying Democrats are critical, classify it as 'democrat_critique"'

#vietnam

process the results to a single file, separated by file name

Token usage and pricing

From Google: https://ai.google.dev/gemini-api/docs/billing

Billing for the Gemini API is based on two pricing tiers: free of charge (or free) and pay-as-you-go (or paid). Pricing and rate limits differ between these tiers and also vary by model.

We are not using the paid tier. Do not activate billing.

More from Google on counting tokens, using python: https://ai.google.dev/gemini-api/docs/tokens?lang=python

Learn more about the possibilities of AI analysis of video, audio: https://ai.google.dev/gemini-api/docs/long-context