NEWS AGGREGATOR SERVICE FOR ASTROLOGY (NASA): A FREE RESEARCH TOOL

Renay Oshop, AyurAstro.com, renay@ayurastro.com

Abstract

Contemporary astrologers are challenged like never before to make their work scientific. This is especially true for astrology researchers who are stewards of the keystone for legitimacy in astrology. Fortunately, current strides in artificial intelligence allow new avenues to data cultivation for the public and private astrology researcher. The author presents her freely available solution to the challenge, using in her explanation basic, intermediate, and advanced concepts.

The Challenge

Contemporary astrologers the world over, whether Eastern, Western, Persian, Chinese, or Tibetan, face the same challenge. Event data of time, place, and date, including that of births, are mandatory for us to do our work.

Acquiring that data is more difficult than at first blush it may appear. For example, if one is interested in sports astrology and one wants to know the start of a basketball or cricket match from five years ago, how does one actually find that?

Perhaps there is an online news article from that time, but there are likely to be hidden problems. What I have found is that there is a wide variety among news publications as to how they treat such events.

For example, *BBC* will include the time of the match in subsequent or preceding write-ups, while *The New York Times* will not.

Another part of the challenge is the language used in an article. For example, the article may say something like “the match happened last Sunday”. *The New York Times* often uses such an approach, begging the questions of which Sunday is that and what is its date. It may be obvious to the contemporaneous online reader but not to one years thereafter.

Then, there is a question of whether those articles are even archived and hence publicly accessible for some time later. Often, only the articles subsequent to an event are kept posted online and then only for a few years.

Finally, there is the strong possibility of an astrologer not knowing that the event exists in the first place. For example, let’s switch our focus to cyclone event data. There was a cyclone in the Spring of 2019 named Cyclone Ida.

There are hundreds of cyclones per year. (National Hurricane Center, 2019) Twenty years from now, will an astrologer who is interested in cyclones know to do an online search of contemporaneous literature for that particular cyclone? Will there still be material online for that cyclone?

Doing astrological research is hard. For the cyclone researcher, he or she may need to pore over tens of thousands of webpages just to get that kernel of truth of when, where, and what time a batch of cyclone events occurred.

Reading that many documents takes time. Even if a list of ten thousand website URLs were presented instantly to the researcher, spending three minutes to scan each article would necessitate a solid 21 hours straight of such focused activity. That is not counting the search itself for the URLs or the recording of the results.

Given these daunting facts, there is no wonder as to why astrological research is relatively rare and often scanty or incompletely done. Science, by contrast, demands first and foremost the relative accuracy and completeness of data. This core difference between astrological research and conventional science represents the deepest chasm between the two world perspectives.

Bridging this chasm of data completeness would do the most to unify the work of all astrologers everywhere within the paradigm of science that could be said to define our current societies.

The Solution

A data challenge needs a data solution.

The Information Age in which we live deluges us with data. Fortunately, in these times, we are being presented with newly hatched tools to automate the search for and the processing and retaining of data in bulk. The newest and best methods use artificial intelligence (AI).

At this time, these AI tools are being used throughout every industry except astrology. These tools presently are also not for the mathematically faint of heart, so there is still a pretty steep learning curve for implementing them. Commercial industries find that the effort directed to the computational and

mathematical gymnastics is more than worth it. The results are phenomenal, world-changing, and transforming to the industries even as the computer science in use is still relatively in its infancy.

What is needed is a means of using these tools to help the future astrology researcher who is within that exponentiating global data deluge to find and record the relevant data that was previously preserved as it happened.

Just as astrologers use books for transmission of knowledge, email for communicating with clients and each other, and software for chart generation, it is time for astrology to join the rest of professional society by investigating and employing the latest machine learning methods to generate artificial intelligence within astrology.

Event data generation is the first order of business for the personal or public research astrologer. Using artificial intelligence on global public information to cultivate event data should be the first order of business for a current computer scientist who wishes to help astrology.

Basic Ideas

There are a few steps to a computer-automated processing of current event data to help a future astrologer.

A computer program would be needed to 1. **monitor** daily current events as they are published in online articles. The computer program would then need to 2. linguistically **digest** the complex language of online data to find A. the nature of the event (for example, a new cyclone) and B. the date, time, and place of the event. After digestion, the computer program, at a minimum, would need to 3. **preserve** A. and B. for future use.

The basic triplicate of requirements of monitoring, digestion, and preservation of astrologically relevant data in a timely manner is in truth a cascade of separate algorithms, or sub-programs, each of which only has been developed recently, sometimes as recently as just a few months ago.

Monitoring can be done through a process called web scraping, wherein the texts of new articles are programmatically culled. However, of the millions of new articles that get published per day, which have data relevant to events of interest to the future astrologer? There is a computational cost to aggregating such monitored data. Even if the program takes but one second per article to seek out, acquire, and analyze for relevancy the text of the article, only 86400 such articles could be processed per day for this step alone. Thus, for current network, hardware, and software constraints, a necessary restriction in the numbers of internet sources and research topics is required.

Digestion is the most technically challenging part of the three-step process. The computer program would need to do nothing less than understand language itself. Take this sentence as an example: “Last Sunday night, three hours after kickoff, the Chicago Bears won the Super Bowl championship against the Denver Broncos.” The program would need to know that A. the article is about a match in American football and B. what was last Sunday’s date. The location and time of kickoff would hopefully similarly also be determinable through other sentences of the article. Digestion makes heavy use of something called natural language processing, a field of artificial intelligence that is still rapidly improving. The following two sections explore this step in some detail.

Finally, **preservation** of the data in a central, publicly accessible repository is the last, necessary step to the program. This is the easiest step of the three and typically just requires publishing the data file online.

Intermediate Ideas

Natural language processing (NLP) across different languages and across different subjects is a wide computer science field with many compartments. The main ones of relevance here are **automatic summarization**, **natural language understanding**, and **question answering**.

Automatic summarization is needed to determine whether an astrology research topic is indeed the topic of the article. Typically, this is done by condensing the article into a small summary and seeing if the topic is present therein. How does one determine the summary of an article? The non-trivial prospect of text extraction and synthesis based on statistical significance metrics and pattern-matching is a relatively straight-forward type of summarization. These days, abstraction of language content is also often required as well. Machine learning based on linguistic feature extraction is at the core of this memory- and computation-heavy abstraction process. As a sub-field of NLP, automatic summarization benefits as being one of the longest studied topics of it. (Hahn & Mani, 2000)

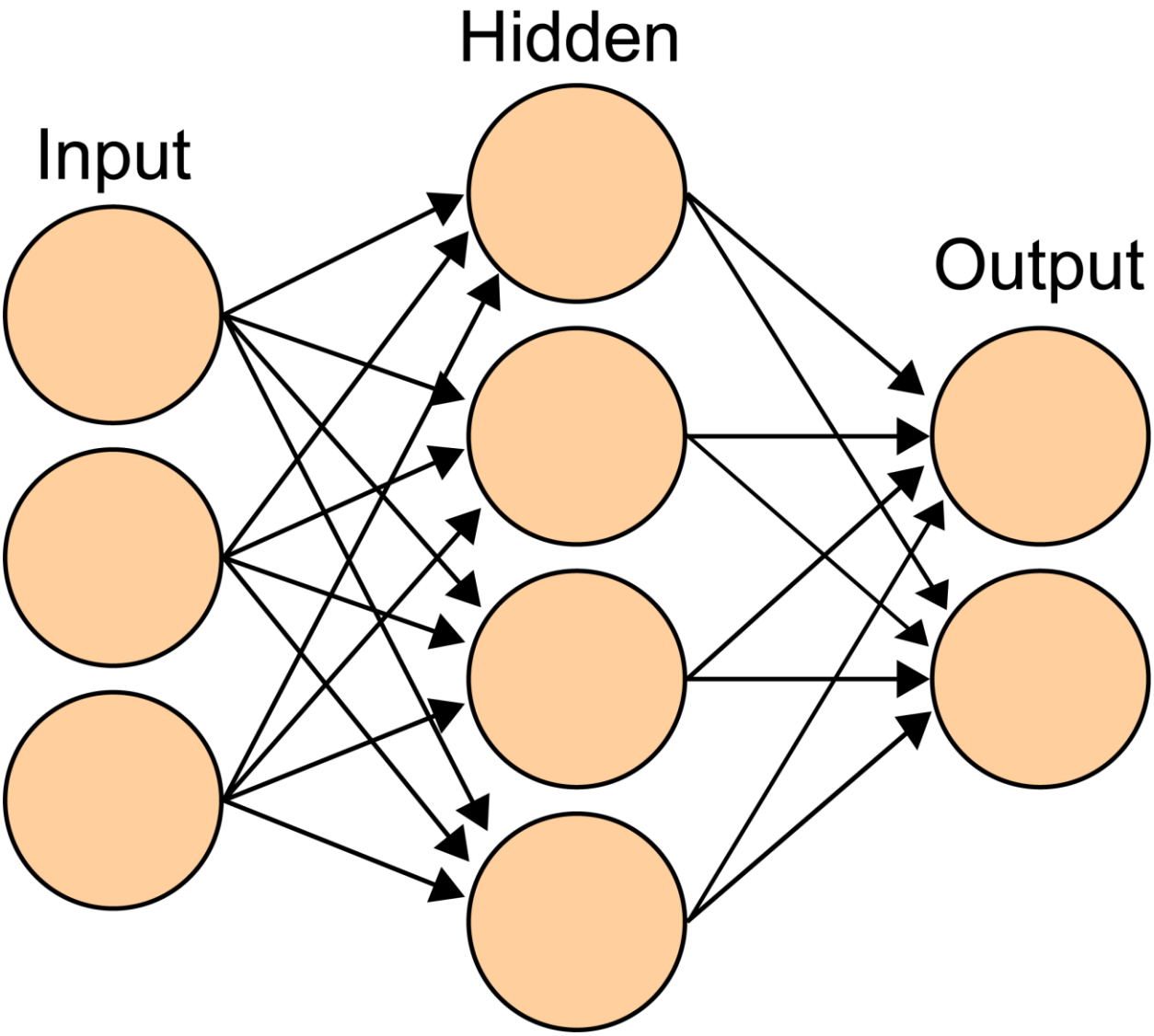
Natural language understanding is the study of how to improve *comprehension* of text by a computer. It is a necessary middle step to answering questions from the text. The reader may be familiar with interfaces for language understanding from *Siri, Bixby*, or *Alexa* apps on their cell phones. Such commonness belies the extraordinary technical achievement that they represent. For example, one may say “The man started on a new novel,” and “The man started on a sandwich”. The formal structure is identical even though the meanings are wildly different.

Finally, **question answering** is critical to the astrologer’s needs. Even with the well-understood text of an article whose topic was deemed relevant, asking questions of the article is also non-trivial. Fortunately for astrologers, there are only three main questions. On what day did the event occur? At what time of day did the event occur? In what town did the event occur? However, the news article may be ambiguous on these matters or it may not include these details at all. For example, for a cyclone, what counts as the birth event? The initial early formation far off at sea? Landfall? Landfall only reaching a major population center? Such delicate domain-specific answering to a question is only lately, in the past few years or so, reaching maturity.

Advanced Ideas

The implementations of automatic summarization, natural language understanding, and question answering that are useful to know about for my project involve **neural nets**.

Neural nets, a.k.a. artificial neural networks (ANN), are a way for the computer to solve problems of NLP that is lightly based on how biological animal brains work. Their loose depiction is as follows.

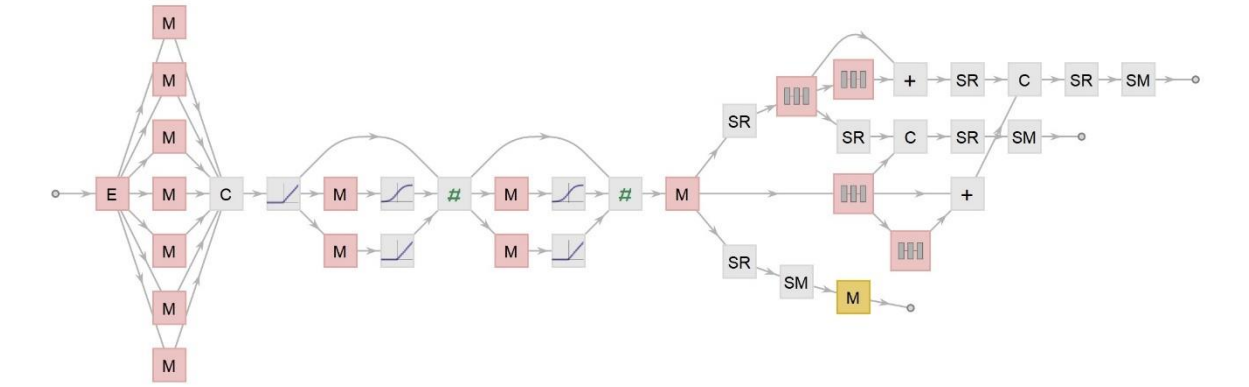


“In common ANN implementations, the signal at a connection between artificial neurons is a real number, and the output of each artificial neuron is computed by some non-linear function of the sum of its inputs. The connections between artificial neurons are called 'edges'. Artificial neurons and edges typically have a weight that adjusts as learning proceeds. The weight increases or decreases the strength of the signal at a connection. Artificial neurons may have a threshold such that the signal is only sent if the aggregate signal crosses that threshold. Typically, artificial neurons are aggregated into layers. Different layers may perform different kinds of transformations on their inputs. Signals travel from the first layer (the input layer), to the last layer (the output layer), possibly after traversing the layers multiple times.” (Wikipedia, 2019)

There are three neural network systems at play in the News Aggregator for Astrology (NASA) project, each decidedly more complex than the above simplistic diagram.

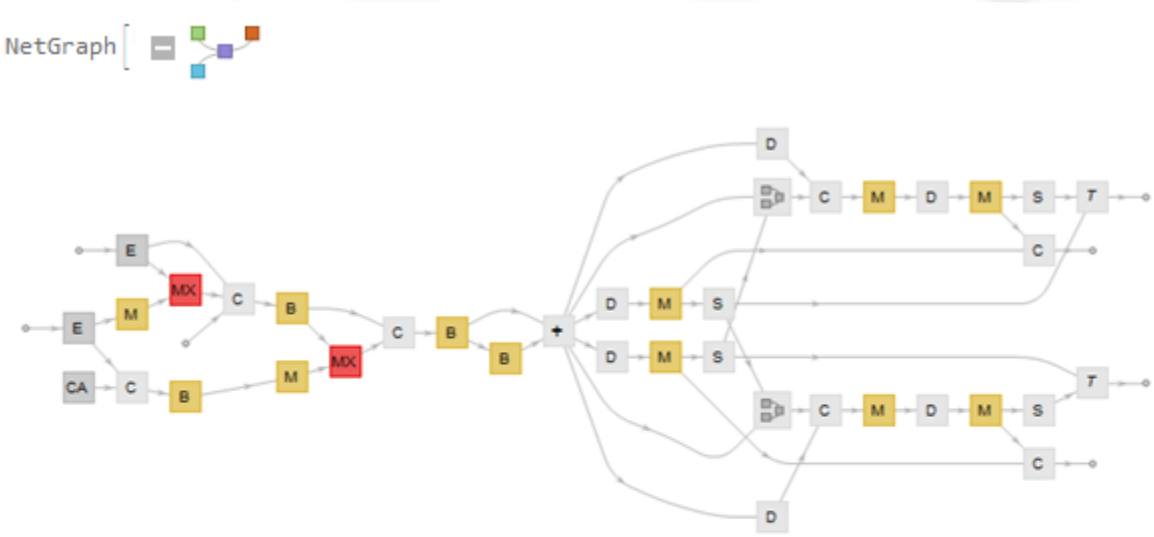
The first is for article summarization. The documentation here is still sparse, with the author promising it as a “to-do”. (Ou-Yang, 2013)

The second is for natural language understanding of semantics via ELMo-encoding that is strong in deriving context of words. (Anon., 2018)



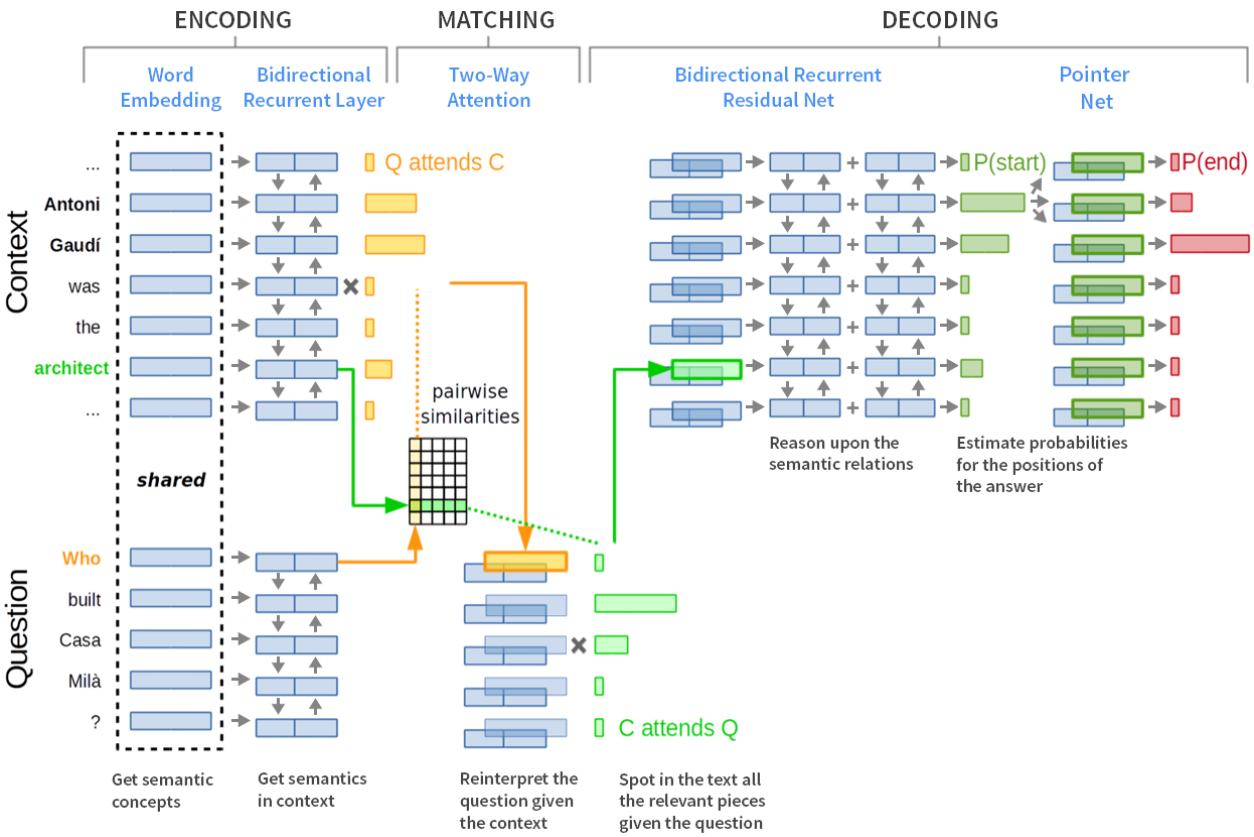
Inputs		Outputs	
Input:	expression	ContextualEmbedding/1:	matrix (size: $n_3 \times 1024$)
		ContextualEmbedding/2:	matrix (size: $n_5 \times 1024$)
		Embedding:	matrix (size: $n_7 \times 1024$)

The last neural net model that follows is used for both natural language understanding and question and answering. It is based on GLoVE-encoding that is strong in deriving meaning of words. (Anon., 2017)



Inputs		Outputs	
Question:	string	End:	matrix (size: $n \times 1$)
WordMatch:	matrix (size: $n \times 3$)	StartActivation:	matrix (size: $n \times 2$)
Context:	string	EndActivation:	matrix (size: $n \times 2$)
		Start:	matrix (size: $n \times 1$)

The rubric for the question and answering part of the net is the following. (Louradour, 2018)



NASA

I have operationalized the above neural nets in code form. Their results are available for public use for free at ayurastro.com/nasa.html. No coding is required by the user. The results are in downloadable *csv* format.

The basic coding rubric for my contribution is the following.

1. Import the RSS feed of each news source.
2. Extract URLs of new articles from each RSS feed.
3. Import the summary of each new article.
4. Extract keywords from each summary.
5. Seek match of keywords to acceptable topics.
6. For each match, encode the sentences of the article to form word contexts and meanings.
7. Use the neural net encodings to answer:
    a. Where did the event occur?
    b. On what day did the event occur?
    c. At what time did the event occur?
8. Semantically interpret the answers as location, date, and time.
9. If two or more of the above answers exist, append data to database.

The language I employed to get the results is principally *Mathematica* with a sprinkling of *Python* for the summarization step.

For each event, the following is included:

- Category (for example, cyclone)
- Keywords (for example, Ida)
- Day
- Time
- Town
- Originating URL
- URL Publication Date

Due to resource constraints, currently only the following separate categories of data are being monitored:

- accident
- assassination
- birth
- coup
- crash
- cyclone
- death
- earthquake
- explosion
- fire
- game
- hospitalization
- hurricane
- IPO
- match
- merger
- premiere
- shooting
- tsunami
- volcano
- wedding.

Similarly, the news publications being monitored are presently limited to:

- *The New York Times*
- *The Guardian*
- *Der Spiegel* (English edition)
- *Beijing Bulletin*
- *U. S. News & World Report*
- *BBC*
- *Wikipedia Current Events*.

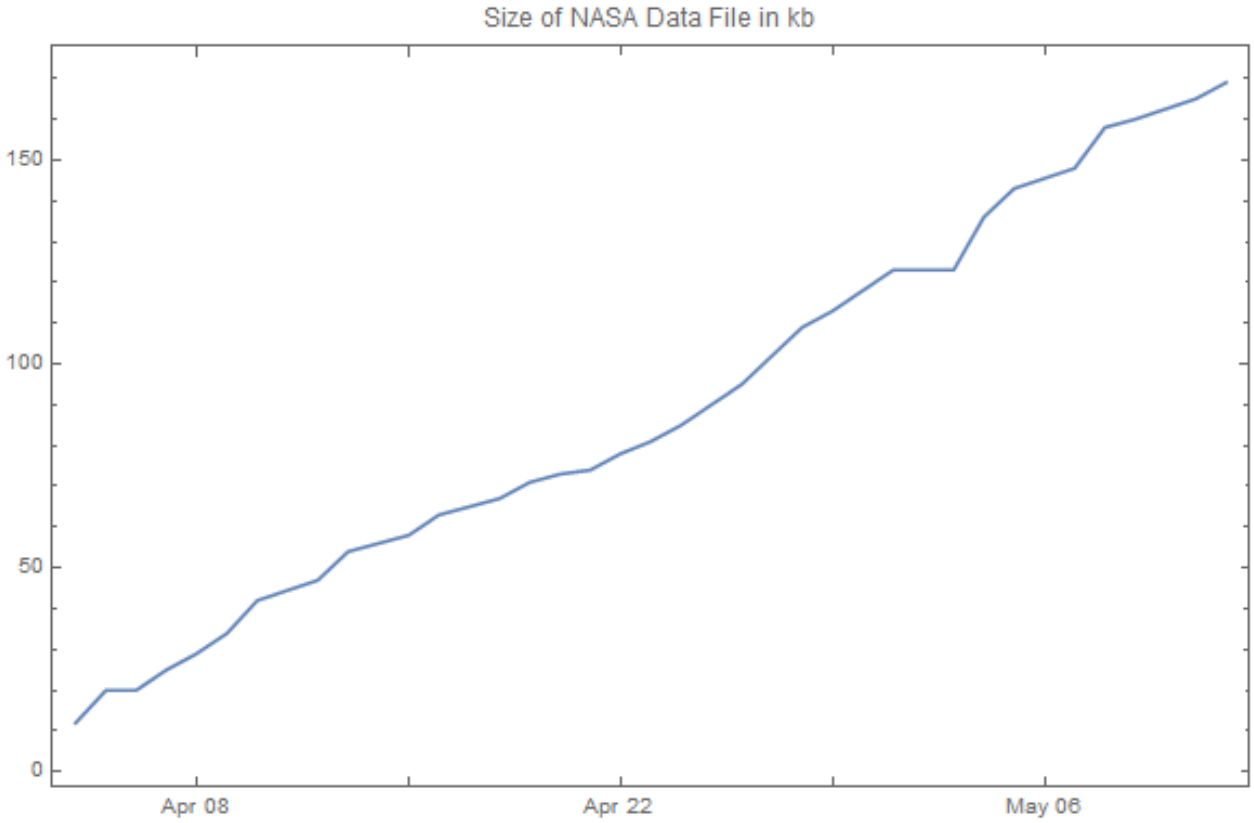
On the NASA web page are entry fields through which the user can submit new topics or new publication sources for consideration for inclusion in scans of future events.

Results: The Good

The program works!

As of this writing (on May 12, 2019, after about a month of operation), the tally for each topic is "accident": 9, "birth": 8, "coup": 7, "crash": 75, "cyclone": 17, "death": 111, "earthquake":10, "explosion": 14, "game": 123, "hurricane": 8, "match": 18, "merger": 8, "premiere": 9, "shooting": 69, "tsunami": 1, "volcano": 3, "wedding": 14.

The following is a graph of the growth of data over time.



At this rather linear rate of growth, in twenty years the size of the data file will be about 36 MB, containing approximately 110000 unique event records.

In comparison, Astro-databank has been in operation for longer and currently represents approximately 63000 records. (Anon., 2019)

Specificity in the current iteration is relatively high. In other words, if the article does indeed have a date, time, or place for an event, the neural nets are able to detect it.

Results: The Not-So-Good

Sensitivity is sometimes poor. For example, an article on a political cyclone of Mr. Trump might still register as a cyclone. The user of the data would then need to read each cyclone-related news article to be sure of false positives. This is nonetheless apt to take less time then doing a broad Google search of all cyclone news articles from across decades.

While entirely automated, the program takes up a lot of computing time and resources – even with the limitations described above. It can only run once a day and uses up a whole computer and software license and so, some entries may be missed.

What to Do Next

In November of 2018, a revolution in public NLP resources was ushered in when the Google Research Group's published the Bidirectional Encoder Representations from Transformers (BERT). (Google Research, 2018)

Unfortunately, presently it takes about ten times the amount of time to process a news article via BERT versus ELMo or GLoVE. So, it is not presently feasible for use in a server for a public operation. With time and increased computer power, BERT should be implementable in NASA within a few years.

Conclusion

As a proof of concept, NASA breaks new ground for automating the hardest and most necessary part of astrology research: the acquisition of data.

However, it is a very young technology and needs subsequent human review of entries due to occasionally poor sensitivity.

Nonetheless, even with only being a month old, and even with the necessary human review of entries, the technology gives the current and future astrology researcher substantial savings in terms of time and energy in finding possible events to review.

NASA should benefit from the inception of new tools that are emerging each month in hardware and artificial intelligence. The author is committed to maintaining the database over the decades and improving it over time.

References

Anon., 2017. *GloVe 100-Dimensional Word Vectors Trained on Wikipedia and Gigaword 5 Data.* [Online] Available at: https://resources.wolframcloud.com/NeuralNetRepository/resources/GloVe-100-Dimensional-Word-Vectors-Trained-on-Wikipedia-and-Gigaword-5-Data
[Accessed 12 May 2019].

Anon., 2018. *ELMo Contextual Word Representations Trained on 1B Word Benchmark.* [Online] Available at: https://resources.wolframcloud.com/NeuralNetRepository/resources/ELMo-Contextual-

Word-Representations-Trained-on-1B-Word-Benchmark
[Accessed 12 May 2019].

Anon., 2019. *Astro-databank.* [Online]
Available at: https://www.astro.com/astro-databank/Main_Page
[Accessed 12 May 2019].

Google Research, 2018. *BERT.* [Online]
Available at: https://github.com/google-research/bert
[Accessed 12 May 2019].

Hahn, U. & Mani, I., 2000. The Challenges of Automatic Summarization. *Computer,* 33(11), pp. 29-36.

Louradour, J., 2018. *New in the Wolfram Language: FindTextualAnswer.* [Online]
Available at: https://blog.wolfram.com/2018/02/15/new-in-the-wolfram-language-findtextualanswer/
[Accessed 12 May 2019].

National Hurricane Center, 2019. *Tropical Cyclone Climatology.* [Online]
Available at: https://www.nhc.noaa.gov/climo/
[Accessed 12 May 2019].

Ou-Yang, L., 2013. *Newspaper3k: Article scraping & curation.* [Online]
Available at: https://newspaper.readthedocs.io/en/latest/index.html
[Accessed 12 May 2019].

Wikipedia, 2019. *Artificial Neural Network.* [Online]
Available at: https://en.wikipedia.org/wiki/Artificial_neural_network
[Accessed 12 May 2019].