Rosoka

Quote Analysis Using the Rosoka Python API

A Case Study

by Gregory F. Roberts, Jason James, and Dr. Kemp Williams
groberts@rosoka.com , jjames@rosoka.com, kwilliams@rosoka.com

Introduction

In this paper we describe how Rosoka's Python API can be used to easily extract quotes and attribute them to their originator. The Rosoka Python API also exposes metrics that allow users to analyze quotes in terms of their readability and their persuasiveness. A test set of quotations on the Covid-19 pandemic from prominent individuals is used for analysis. The results demonstrate a wide range of readability ease despite the role of the speaker, while also illustrating that much of the discourse surrounding Covid-19 is less than persuasive.

NLP

Natural Language Processing (NLP) is a branch of linguistics concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data. For the purposes of this experiment, we are using the industry-standard, enterprise-scale entity extraction engine by Rosoka Software that employs a recursive, finite state automata. The Rosoka extraction engine is written in Java. The engine is controlled by the Rosoka LxBase, a data component that contains all of the entity and relationship definitions, language- and domain-specific dictionaries, entity extraction rules, and relationship expansion details. With a single pass on input, Rosoka provides language identification, entity extraction, relationship extraction, document-level multivector sentiment analysis metrics, and entity-level multivector sentiment analysis metrics. A user is guaranteed the same results if using the same engine and same LxBase on multiple passes on the same data, regardless of operating system or hardware configuration.

RFO Output

The Rosoka Full Object (RFO) format is the native output for Rosoka. This object can be represented as a Java POJO, XML, or JSON object. All extracted information is contained in this object, from the language identification, entity extraction, pronominal anaphora resolution, relationship extraction, non-English language gloss, and sentiment analysis metrics.

A user does not have to use the entire RFO, instead exploiting just the portions of the RFO that are necessary for their use case. For instance, if the user is only interested in PersonToQuote relationships, all other information in the RFO can be filtered or ignored.

The RFO can be used directly as a collection of files or the RFO can be mapped into different persistent data storage applications, such as RDBMS like MySQL, MariaDB, and Postgres, RDF graph servers like Allegrograph, Neo4J, DGraph, Giraph, or StarDog, or Full Text Search Engines like Lucene, Solr, or ElasticSearch. Most modern tools can ingest XML or JSON formats.

In particular, JSON is an easy data format to exploit with Python, so for the purposes of this case study, we are using the RFO in a JSON format. The RFO can be used directly and readily consumed by Python with no additional third-party libraries.

Quote Extraction

A quote is a textual representation of what an individual or group said. Explicit quotes are those that are word-for-word recitations of what was said. For the purposes of this study, we deal with explicit quotes even when we just use the more ambiguous terms “quote,” “quotation,” or “utterance.” Quote attribution is the process of connecting the originator of the quote to the quote. Most readers should be familiar with the typical format of a quote with an attribution to a speaker. The following example illustrates an utterance from the character *Lord Henry* in Oscar Wilde’s novel, *The Picture of Dorian Gray:*

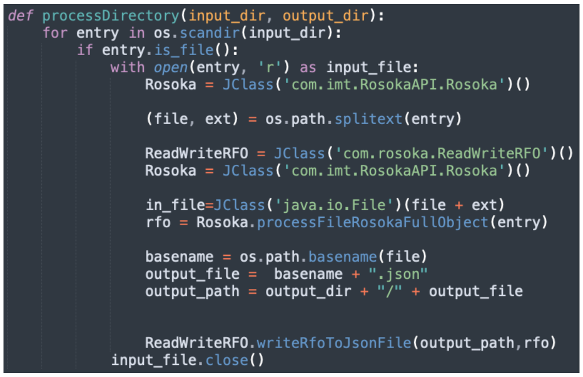
“It is your best work, Basil, the best thing you have ever done,” said Lord Henry languidly.

The out-of-the-box Rosoka NLP capabilities have been extended to recognize direct quotes and create a PersonToQuote relationship to connect the originator to the utterance. From here, additional Rosoka and other text analytical metrics can be applied to the quote or a series of quotes to analyze trends.

Rosoka Python API

Python is a general purpose, high-level programming language that has recently surged in popularity, particularly among data scientists. It is often cited as one of the top five most popular programming languages in the world.

The core Rosoka extraction engine is written in Java. The Rosoka Python API uses the Python JPype library to provide full Java access to Python programs. With about 15 lines of Python code, one is able to process a directory of text files and create JSON output files in the Rosoka Full Object format (RFO).



```
def processDirectory(input_dir, output_dir):
    for entry in os.scandir(input_dir):
        if entry.is_file():
            with open(entry, 'r') as input_file:
                Rosoka = JClass('com.imt.RosokaAPI.Rosoka')()

                (file, ext) = os.path.splitext(entry)

                ReadWriteRFO = JClass('com.rosoka.ReadWriteRFO')()
                Rosoka = JClass('com.imt.RosokaAPI.Rosoka')()

                in_file=JClass('java.io.File')(file + ext)
                rfo = Rosoka.processFileRosokaFullObject(entry)

                basename = os.path.basename(file)
                output_file =  basename + ".json"
                output_path = output_dir + "/" + output_file


                ReadWriteRFO.writeRfoToJsonFile(output_path,rfo)
            input_file.close()
```

Image 1. Use Rosoka’s Python API to easily process a directory of documents and generate JSON Rosoka Full Object files.

The Rosoka Extraction Engine can process virtually any document format, such as DOCX, PDF, HTML, RTF, and plain text.

Quote Analytics

For the purposes of this study, we focused on two metrics, readability and conviction. To measure the relative ease of readability, we used the Flesch readability score, a metric designed to calculate the relative ease or difficulty in understanding a passage of text. Although it is possible for there to be outliers, the Flesch metric typically returns scores in a range from 1-100, moving from harder-to-understand to easier-to-understand as scores increase. A book intended for young adults might score above 80, while an academic treatise might score less than 40.

Conviction is a measure of the speaker's attitude regarding their quote insofar as they believe or do not believe what they are saying. Quotes with higher conviction scores are therefore more likely to be evaluated as persuasive. We derived a conviction score using Rosoka's aspect sentiment metric, one of the multi-vector sentiment scores computed by the Rosoka engine. Aspect measures the degree to which an author or speaker controls a narrative. By normalizing this score as applied to a quote into a range of *strong, neutral*, and *weak,* we estimate the degree to which the speaker is committed to what they are saying.

Case Study

In our experiment, we analyzed 6,068 news articles related to "Covid-19" from online newswire documents collected with the Rosoka News online news aggregation reference architecture. Rosoka News has a set of over 200 seed URLs that collect news from approximately 30 different languages from around the world. A document set of this size represents an average day of news aggregation for the Rosoka News reference architecture on a single topic. This document set contained 21,551 unique quotes attributed to 2,314 unique individuals.

For the purposes of our experiment, we limited our attention to 223 quotes attributed to four widely known individuals often quoted regarding the topic of "Covid-19." Restricting the study to four prominent speakers allowed us to have a sample size consisting of more than just one or two quotes from any single individual. The quotes we chose were from President Joe Biden (59 quotes), White House Press Secretary Jen Psaki (59 quotes), Director of NIAID Dr. Anthony Fauci (75 quotes), and Florida Governor Ron DeSantis (30 quotes).

Table 1 below indicates that more than half of each speaker's quotes received Flesch readability scores within the "Fairly Difficult" to "Very Confusing" range.

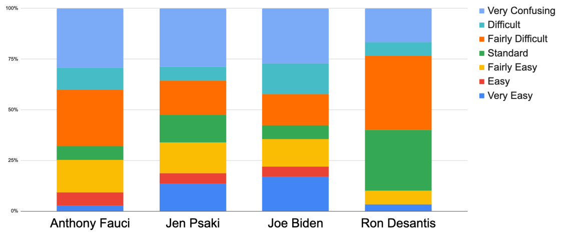


Table 1. Flesch Readability Scores for quotes about COVID-19

Given his role as the chief scientific spokesperson during the Covid-19 pandemic, it is not surprising that the individual with the highest percentage of quotes (68%) in the fairly difficult to very confusing range was Dr. Fauci. The following is an example of one of Dr. Fauci's quotes algorithmically determined to be very confusing:

"if you look at the data, the data are strongly suggestive in this country, and more than just suggestive in israel, that you have a waning of immunity among people across age groups, not just the very very elderly, you have clearly waning of immunity against infection and clear cut indication of waning of immunity against severe disease."

The Flesch readability score for this quote is 8, placing it firmly within the realm of academic or professional speech. Also not surprising, the individual with the lowest percentage of quotes (53%) in the fairly difficult to very confusing range was Secretary Psaki, who has the job of communicating to a wide, non-scientific audience what can sometimes be complex or nuanced information. A quote from Secretary Psaki that is more typical of someone in such a role is the following one:

"We are continuing to look for ways for the U.S. government to support districts and schools as they try to follow the science."

This quote has a Flesch score of 73, indicating that it is fairly easy to understand.

Table 1 makes it clear that there is a broad range of readability values among the quotes attributed to each of the four speakers in the quotation test set. All speakers but one have quotes across all seven categories on the readability scale. (The exception is Governor DeSantis, none of whose quotes fell within the "Easy" readability range. This may be due to his having the fewest number of quotes in the test set.) Despite the broad range of readability scores, however, there is not a significant correlation between these scores and a measure of the speakers' convictions as evinced in the quotes.

Table 2 below shows that of the sampled quotes, each speaker communicated with weak conviction approximately 75% of the time. In many cases, a quote with a weak conviction score is likely due to a speaker using vague or unspecific language, such as in these quotes from Governor DeSantis and Secretary Psaki, respectively:

"I think protecting the vulnerable has been the right way to go."

"The president's position is that every company should take a look at how to protect their workforces, and there are going to be different carrots and sticks that can be used by different private sector entities."

Whatever the reason for the use of such imprecise language, it is clear that no official, regardless of point of view, spoke persuasively much of the time.

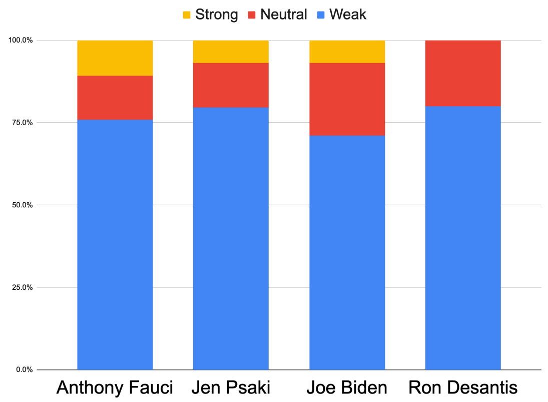


Table 2. Conviction measures as a percentage of each speaker's quotes.

Conclusion

The examination of quotes in this study suggests that the polarization surrounding the topic of Covid-19 should not be unexpected. These quotes come from individuals who all play prominent roles in communicating the urgency of the Covid-19 pandemic to the public, yet much of what they have to say is at a readability level that the general population would find difficult or confusing to understand. Furthermore, analysis of the speakers' convictions regarding the information they are communicating generally lacks persuasiveness. A recent article on *The Readability Blog* summarizing studies of the information available to the public on Covid-19 concluded:

> **"Quite simply, if the general public can't understand healthcare information, they don't know where to turn--resources seem to conflict, and it's hard to trust medical advice, which is alienating."**

This analysis of what public officials are saying about Covid-19 suggests that much of the time their speech is above a level the general public finds easy to understand. Furthermore, even when their speech is understandable, it is not delivered in such a way that it is perceived as convincing.

Further Research

There are two areas of further research that we would like to explore. The first is to determine whether we can correlate any metrics to ascertain if a quote attributed to a particular individual, such as a political candidate, was actually produced by the individual or was in reality created by a ghost- or speech-writer. The second area of research we would like to explore is examining the interaction of the Rooska salience metric and how it interacts with the Flesch readability score. The Rosoka engine produces a salience measure to indicate how important a given entity is to the document (or in this case, quote). Assessing this measure in conjunction with readability and conviction can provide an indication of the speaker's perspective with regard to the most salient entity in the quote.

References

Readability Formulas. *The Flesch Reading Ease Readability Formula.* https://readabilityformulas.com/flesch-reading-ease-readability-formula.php. Accessed 10-01-2021.

The Readable Blog. *Readability and Coronovirus: Scoring Top-Ranked Advice.* https://readable.com/blog/readability-and-coronavrius-healthcare-advice. Accessed 10-01-2021.