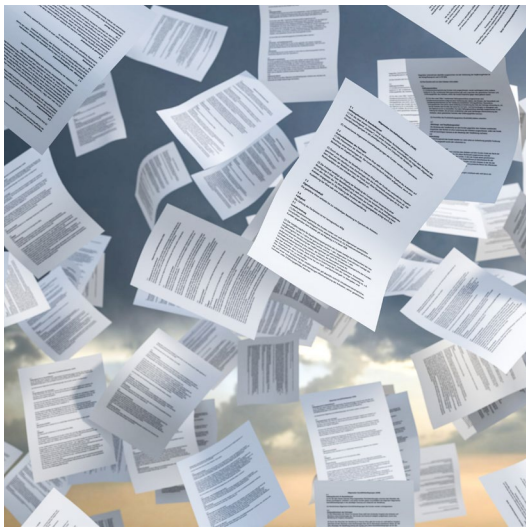
mosaic
DATA SCIENCE

http://mosaicdatascience.com/
info@mosaicdatascience.com



Data Science White Paper

Summarizing Text with Artificial Intelligence

Mosaic reveals how we built advanced language models that are great at summarizing text with AI.

Machine Learning

Mosaic data scientists collaborate with customers, digging deep into the data to inform design and deployment of custom ML tools that make a difference.

Artificial Intelligence

Mosaic integrates powerful AI tools into clients' existing technology stack to solve complex business challenges

Business Analytics

Mosaic helps corporations of all shapes and sizes take advantage of their data, transforming their decision-making processes.

THE NEED FOR AI LANGUAGE MODELS

International Data Corporation (IDC) forecasts that the total amount of digital data transmitted worldwide will jump to 180 zettabytes in 2025[1]. Couple this with the fact that 80% of the data an organization collects is estimated to be unstructured text files. Think about the sheer number of emails, resumes, text documents, research findings, legal contracts, invoices, call recording, social media posts, etc., etc., your firm possesses.

All this data presents a tremendous opportunity to benefit all stakeholders of an organization - investors, employees, processes, and the all-important customer - if the organization can find a way to sift through this data in an automated way to extract key information and solve specific challenges. In that case, they could learn about their firm and start optimizing the way they operate.

Natural Language Processing (NLP) still seems to be out of reach for many organizations. This is likely due to the speed with which the field has changed in the last few years (thanks in large part to advances in deep learning for NLP) along with the challenge to deploy these tools properly such that they can continue to generate business value over the long term.

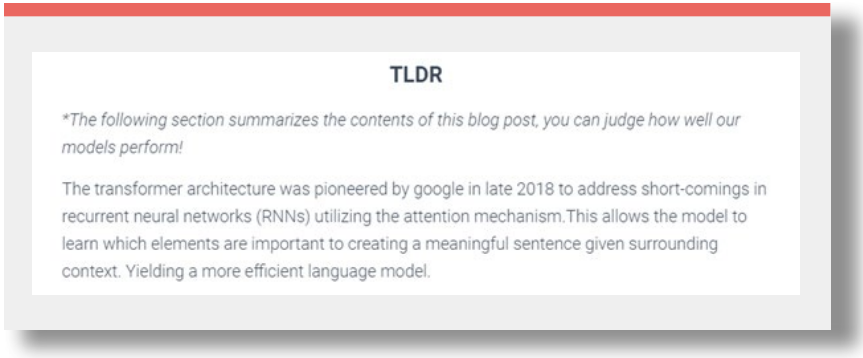
Just think about how much time you could get back, if you, the reader, could summarize lengthy text documents and even generate text with NLP models. In the following whitepaper, Mosaic will examine an approach from a recent effort in text summarization and natural language generation.

DEFINING TEXT SUMMARIZATION

Text summarization refers to synthesizing a long document or multiple documents into a concise summary, saving human review time. The intent is to create a clear overview of the main points of the document. Automating this process using deep learning-based NLP models can be challenging, which is why automated text summarization tools haven't permeated the market already. Not saying there aren't powerful NLP applications already out globally; heard of Google, but we don't think the proper definition of this technology and applicability is widely known.

To build a text summarization tool, Mosaic took advantage of recent advances in language modeling. Namely, the so-called Transformer architecture that greatly expanded the capability of NLP. For those looking for a primer on transformer-based models, we have you covered here[2] and also here[3].

Don't want to read the articles, don't worry - we turned the models loose to summarize the blog posts we wrote:



DEFINING THE PROBLEM

Mosaic's customer is an aggregator of information. They comb through a diverse set of lengthy research documents such as peer-reviewed academic papers, news articles, scientific findings, and related details to produce summaries of the current state of knowledge on various topics. Researchers can get up to speed on a new topic by reading one of these summaries before digging into individual publications. To stay ahead of the competition, saving search time and providing reliable results is incredibly important to their business.

There are opportunities to automate the data collection process, classify the different text sources, and finally generate summaries that the user can trust. A user may want different levels of summarization of an 80 page paper, depending on their goals. A couple sentences of explanation may be helpful in identifying relevant papers in systematic review, whereas, a summary of a few sentences from each major section would be helpful while synthesizing the information across papers identified as relevant.

Our customer identified two potential product areas that could be improved with NLP and AI. One would be to develop a new product that summarized and synthesized academic research papers that they could sell to existing subscribers. The other would be a product enhancement to improve the quality and timeliness of their topic summaries. The models to drive these new capabilities needed to process various data sources, including academic research and news articles, to provide a multi-document summary on a specific topic. Think about all the information related to COVID-19; deploying a mechanism to summarize different aspects of the virus could save researchers a significant amount of time.

NLP PROCESSING

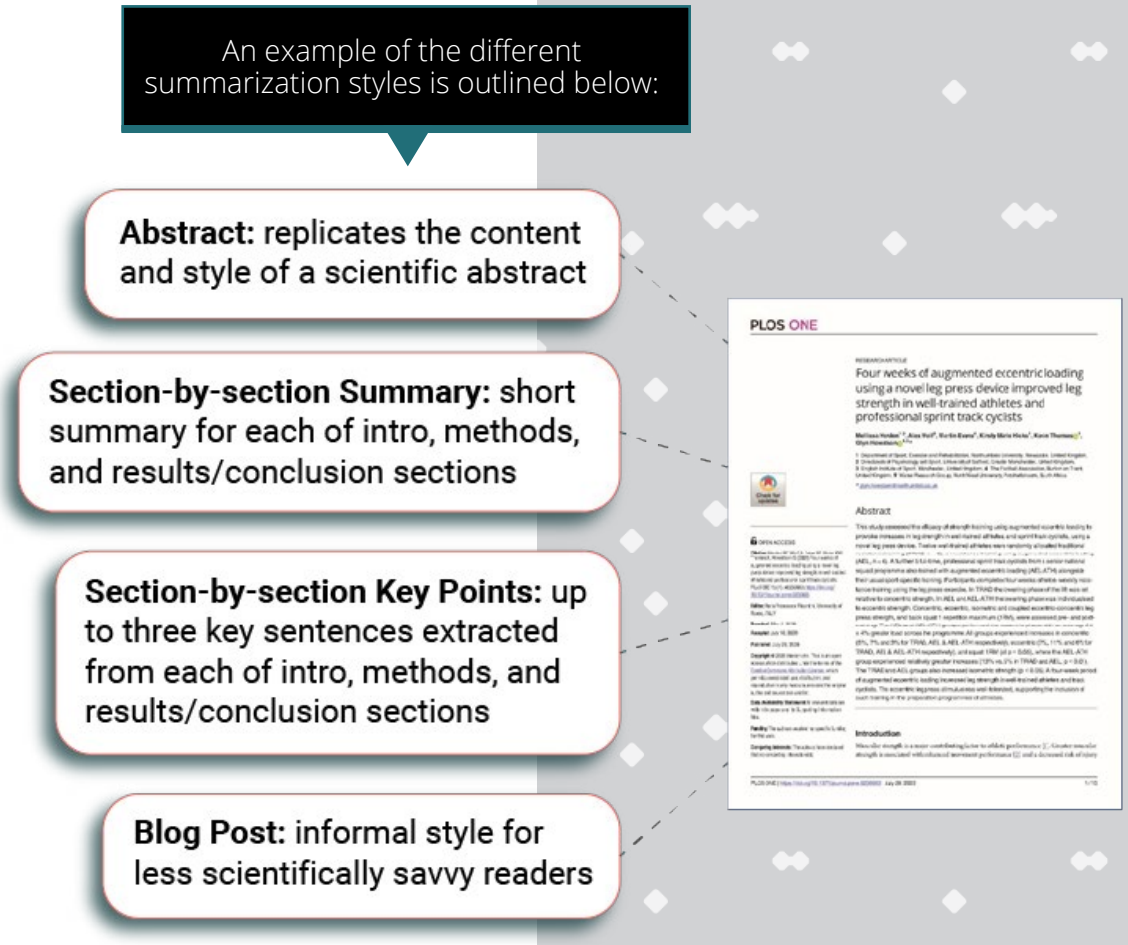
Once Mosaic had collaborated with the publisher on what sorts of information would be ideal for capturing in the topic summaries, Mosaic's data scientists laid out an actionable modeling plan to attack both problems.

For this proof of value to be successful, Mosaic needed to demonstrate a flexible approach that would provide multiple summarization structures and styles for different objectives and end users. Modern NLP tools can be tuned to stylistic preferences based on examples of similarly styled summaries. The tool also needed to be general enough to apply to various research fields and serve as a source for a variety of downstream applications.

Iterative Model Prototyping

There are many cutting-edge transformer-based models to select. Mosaic identified which are better at text generation, incorporating semantic information, or fully leveraged transfer learning. The ERNIE[4] and T5[5] models proved to be quite valuable in meeting the publisher's dynamic needs. Mosaic developed bespoke document summarization models to match the publishers' scaling needs using various model architectures and transfer learning methods to tune models to each style. The advantages of including multiple summarization styles are:

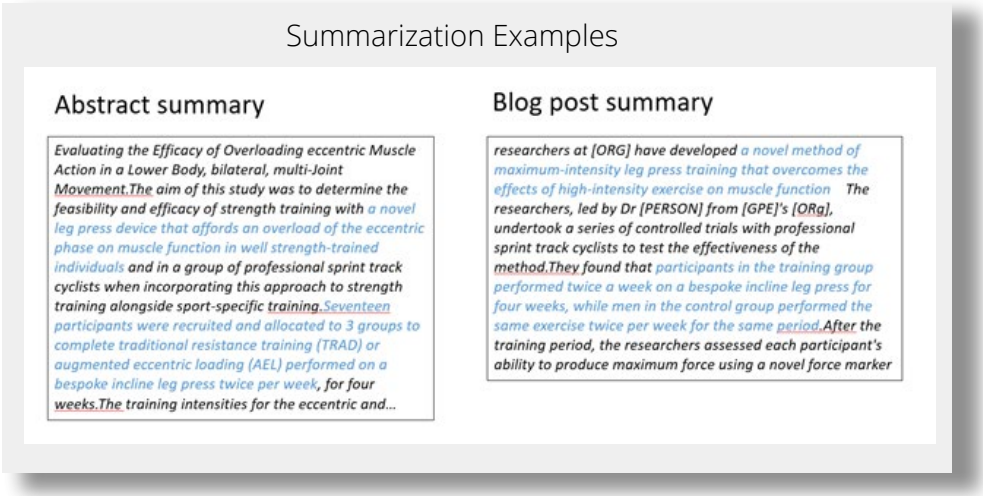
- Built to be customized, the user can tune to various style, level of detail, etc., that is desire for different use across the publisher's diverse product suite.
- Effective across fields of research: from particle physics to anthropology
- Built from the ground up with open-source tools: the publisher can own all IP.
- Scalable, extensible, and easy to integrate, not limited by the fixed architecture of a commercial summarization tool.



To see the model in action, Mosaic demoed the capability for the customer and turned it loose on the following research paper[6]. The exercise science paper examined the effectiveness of a new leg press device for high-level cyclists.

Sample shot of the Research Paper & Abstract

Mosaic's NLP models were able to extract the following abstract and blog post summary.



Abstracts were important to the customer because so many of their papers did not have them. In addition, the models were able to summarize different sections of the paper, as outlined above.

NLP Model Examination

This work leverages the latest in deep learning language models. Heuristic rules could be implemented to enforce higher quality of model outputs. For example, decisions need to be standardized to place higher importance on more recent information or resolve conflicting information.

OUTCOMES OF PROOF OF VALUE

The offline models demonstrated automation capabilities for target research topic areas using static inputs sourced from target papers. A thorough evaluation of model performance, gaps, and opportunities for continued improvement will be conducted. Mosaic also defined steps and timelines for productionalizing models, refactoring analysis code for production, extension to other research topic areas and source types, connection to live data inputs.

In short, the models did what they needed to do to prove value in this effort. The publisher can use these models to generate four styles of summaries to meet their customers' needs. As stated previously, this effort is ongoing. It warrants further validation, but by partnering with Mosaic, the publisher can now investigate how best to deploy AI to automate their text review process.

WHAT CAN THIS TECH DO FOR ME?

So much of an organization's information lives in text data. AI, specifically NLP, helps organizations leverage their data to make better decisions. These models were able to accurately summarize complicated research papers into clear, concise summaries, saving countless human-review time. If you and your organization find yourself spending time reading through complex documents, there is a good chance partnering with Mosaic and deploying custom language models can help you automate that review process.

Endnotes

1. https://www.networkworld.com/article/3325397/idc-expect-175-zettabytes-of-data-worldwide-by-2025.html

2. https://www.mosaicdatascience.com/2020/08/25/document-summarization-techniques-nlp-blog/

3. https://www.mosaicdatascience.com/2020/08/25/language-model-review-gpt3/

4. https://arxiv.org/pdf/1904.09223.pdf

5. https://github.com/google-research/text-to-text-transfer-transformer

6. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0236663