THE SUM(MARY) OF ALL THINGS

Alex Sloan

May 27, 20202

There are two main approaches to summarizing text documents; they are:

1. **Extractive Methods**
2. **Abstractive Methods**

… **there are two different approaches for automatic summarization: extraction and abstraction. Extractive summarization methods work by identifying important sections of the text and generating them verbatim; […]**

abstractive summarization methods aim at producing important material in a new way. In other words, they interpret and examine the text using advanced natural language techniques in order to generate a new shorter text that conveys the most critical information from the original text

Text Summarization Techniques: A Brief Survey, 2017.

OBJECTIVE:

Build a bert based domain/business specific abstractive summarizer

The domain, in this model -- COVID-19

In other words, build CovidBert

I WILL BE USING FOUR DIFFERENT EXTRACTIVE SUMMARIZER MODELS:

COUNT VECTORIZE

TF_IDF VECTORIZE

GENSIM

BERT

STAKEHOLDERS

The stakeholders are just about anybody.

1.7MB of data is created every second for every person on earth.

No one has time to read even an infinitesimal fraction of it.

A domain specific summarier lets you scan the summary of articles that are germane to your discipline or allows you to create summarizers that you can use in your business -- either internally or client facing.

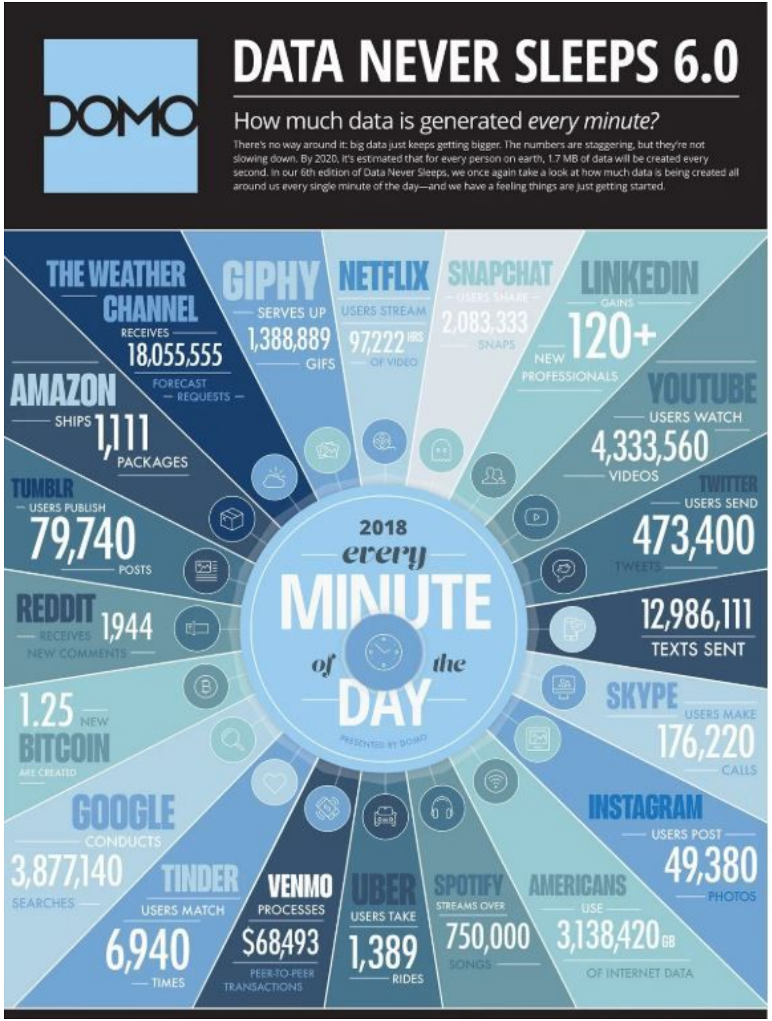
STAKEHOLDERS

The stakeholders are just about anybody.

1.7MB of data is created every second for every person on earth.

No one has time to read even an infentesimal fraction of it.

A domain specific summarier lets you scan the summary of articles that are germane to your discipline or allows you to create summarizers that you can use in your business -- either internally or client facing.



DOMO

DATA NEVER SLEEPS 6.0

How much data is generated *every minute?*

There's no way around it: big data just keeps getting bigger. The numbers are staggering, but they're not slowing down. By 2020, it's estimated that for every person on earth, 1.7 MB of data will be created every second. In our 6th edition of Data Never Sleeps, we once again take a look at how much data is being created all around us every single minute of the day—and we have a feeling things are just getting started.

2018 *every* MINUTE *of the* DAY

PRESENTED BY DOMO

THE WEATHER CHANNEL RECEIVES 18,055,555 FORECAST REQUESTS

GIPHY SERVES UP 1,388,889 GIFS

NETFLIX USERS STREAM 97,222 HRS OF VIDEO

SNAPCHAT USERS SHARE 2,083,333 SNAPS

LINKEDIN GAINS 120+ NEW PROFESSIONALS

YOUTUBE USERS WATCH 4,333,560 VIDEOS

TWITTER USERS SEND 473,400 TWEETS

12,986,111 TEXTS SENT

SKYPE USERS MAKE 176,220 CALLS

INSTAGRAM USERS POST 49,380 PHOTOS

AMERICANS USE 3,138,420 GB OF INTERNET DATA

SPOTIFY STREAMS OVER 750,000 SONGS

UBER USERS TAKE 1,389 RIDES

VENMO PROCESSES $68,493 PEER-TO-PEER TRANSACTIONS

TINDER USERS MATCH 6,940 TIMES

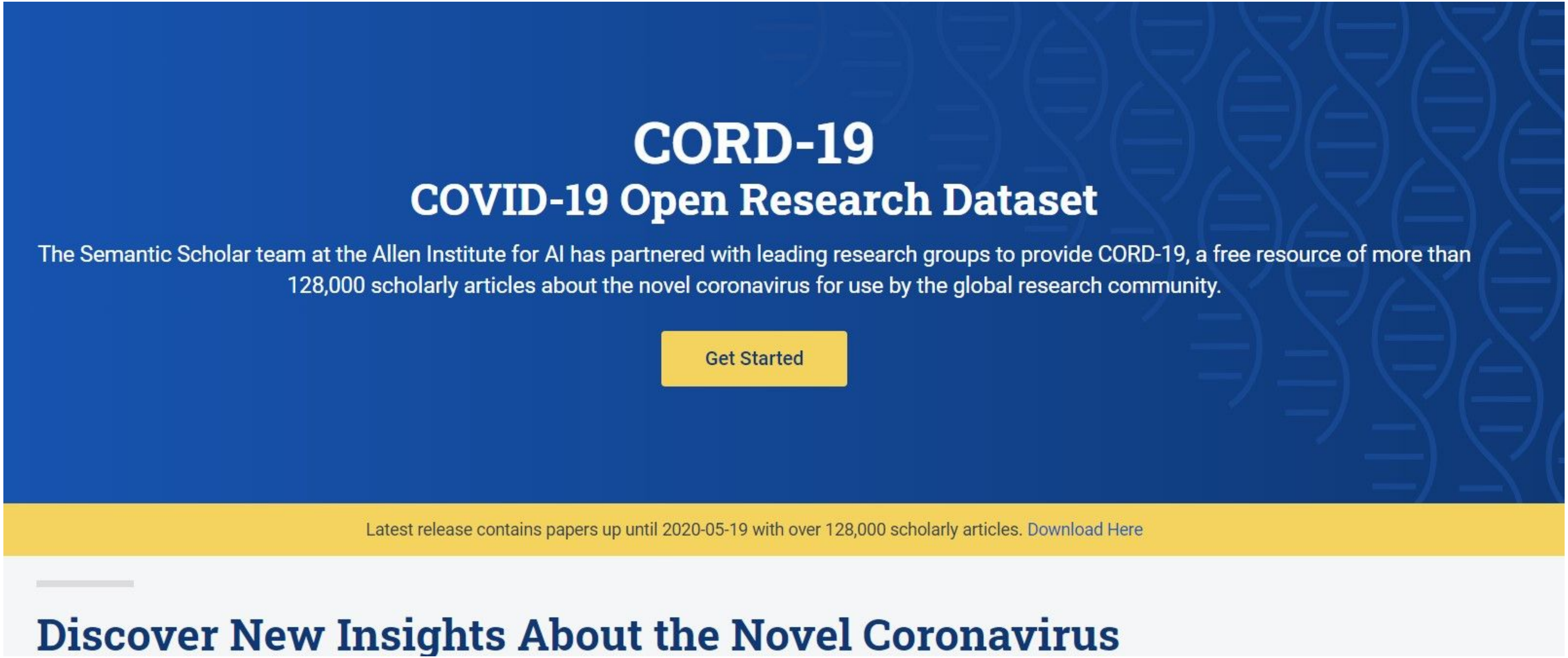
GOOGLE CONDUCTS 3,877,140 SEARCHES

1.25 NEW BITCOIN ARE CREATED

REDDIT RECEIVES 1944 NEW COMMENTS

TUMBLR USERS PUBLISH 79,740 POSTS

AMAZON SHIPS 1,111 PACKAGES



CORD-19

COVID-19 Open Research Dataset

The Semantic Scholar team at the Allen Institute for AI has partnered with leading research groups to provide CORD-19, a free resource of more than 128,000 scholarly articles about the novel coronavirus for use by the global research community.

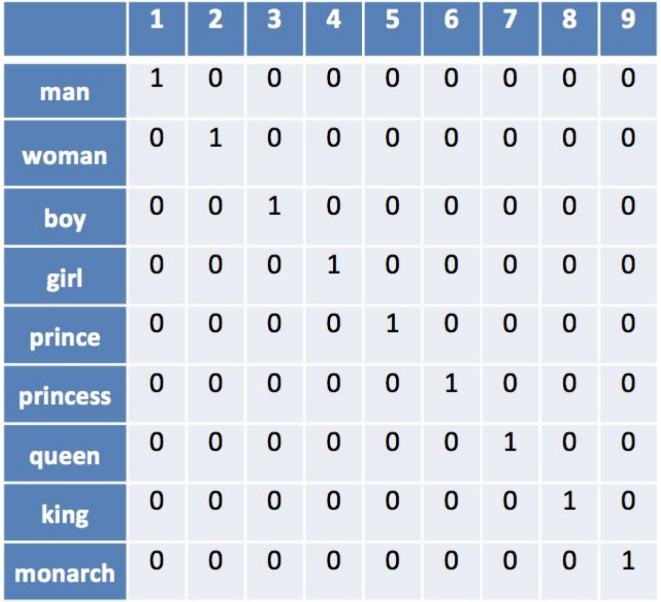
Latest release contains papers up until 2020-05-19 with over 128,000 scholarly articles. Download Here

Discover New Insights About the Novel Coronavirus

"Enzyme assays for synthesis and degradation of 2-5As and other 2\u2032-5\u2032 oligonucleotides",

WORD VECTORS

Vocabulary:
Man, woman, boy, girl, prince, princess, queen, king, monarch



	1	2	3	4	5	6	7	8	9
man	1	0	0	0	0	0	0	0	0
woman	0	1	0	0	0	0	0	0	0
boy	0	0	1	0	0	0	0	0	0
girl	0	0	0	1	0	0	0	0	0
prince	0	0	0	0	1	0	0	0	0
princess	0	0	0	0	0	1	0	0	0
queen	0	0	0	0	0	0	1	0	0
king	0	0	0	0	0	0	0	1	0
monarch	0	0	0	0	0	0	0	0	1

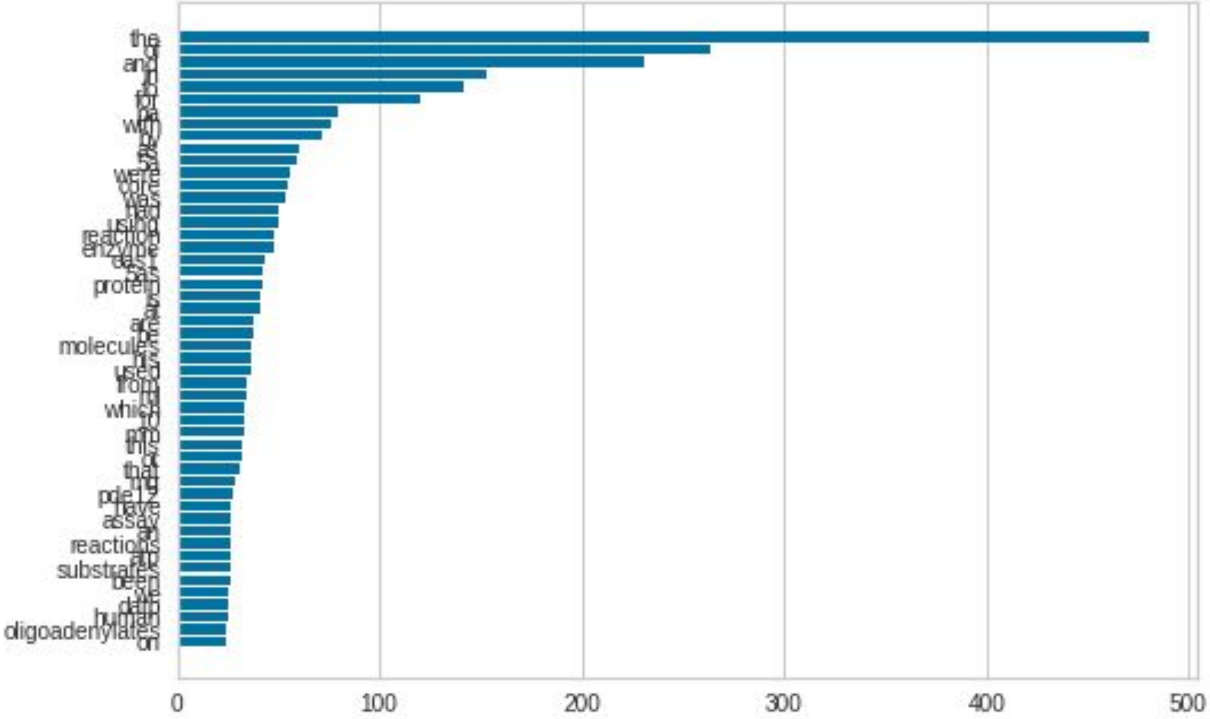
Each word gets a 1x9 vector representation

COUNT VECTORIZE SUMMARY

The chromatograms were obtained using 254 nm as the absorbance wavelength.The enzyme activity reaction catalyzed by OAS can be formulated with the general formula below, in which S1 and S2 are the substrates and P the product formed: S1 þ S2 → P E, the specific activity of the enzyme (mole P generated per second per gram protein) can be calculated from the equation:[S2] start is the start conc of S2, V is the reaction volume, M is the amount of protein added and T is the reaction time.The reaction schemes for the two kinds of OAS activities measured in this paper are depicted in reaction 1 and reaction 2, respectively.In reaction 1, the specific activity of the OAS1 enzyme (E1(OAS1) with dATP and A(pA) 3 as substrates (S1 and S2, respectively) was calculated with the formula:The integral refers to the area under the individual peaks from the substrates and products in the chromatograms.Reaction 2: dNTP + NAD + → NAD − pdN In reaction 2, dNTP (dATP, dCTP, dGTP or TTP) and NAD + were used as substrates (S1 and S2, respectively) in reactions with the OAS1 enzyme

The specific activity of PDE12 (mole 5′-AMP synthesized per second per gram of protein (mole AMP/(sec*g)) was calculated with the formula:Total 254nm refers to the sum of the integral of all peaks in a given chromatogram at 254 nm for a single reaction, [A(pA) n] start is the start concentration of the substrate A(pA) n, V is the reaction volume, T the reaction time and M the amount of protein added.As an example, the tetramer core substrate would be expected to undergo cleavage into the following products using PDE12 as the enzyme:In the case of purified protein, the specific enzyme activity of PDE12, E(PDE12) could therefore be calculated using the formula:For reactions including crude protein extracts, the specific enzyme activities were calculated in mmole AMP/ (sec*g of total protein)

FREQUENCY OF TOP 50 WORDS



TF-IDF

The chromatograms were obtained using 254 nm as the absorbance wavelength.The enzyme activity reaction catalyzed by OAS can be formulated with the general formula below, in which S1 and S2 are the substrates and P the product formed: S1 þ S2 → P E, the specific activity of the enzyme (mole P generated per second per gram protein) can be calculated from the equation:[S2] start is the start conc of S2, V is the reaction volume, M is the amount of protein added and T is the reaction time.The reaction schemes for the two kinds of OAS activities measured in this paper are depicted in reaction 1 and reaction 2, respectively.In reaction 1, the specific activity of the OAS1 enzyme (E1(OAS1) with dATP and A(pA) 3 as substrates (S1 and S2, respectively) was calculated with the formula:The integral refers to the area under the individual peaks from the substrates and products in the chromatograms.Reaction 2: dNTP + NAD + → NAD – pdN In reaction 2, dNTP (dATP, dCTP, dGTP or TTP) and NAD + were used as substrates (S1 and S2, respectively) in reactions with the OAS1 enzyme

The specific activity of PDE12 (mole 5′-AMP synthesized per second per gram of protein (mole AMP/(sec*g)) was calculated with the formula:Total 254nm refers to the sum of the integral of all peaks in a given chromatogram at 254 nm for a single reaction, [A(pA) n] start is the start concentration of the substrate A(pA) n, V is the reaction volume, T the reaction time and M the amount of protein added.As an example, the tetramer core substrate would be expected to undergo cleavage into the following products using PDE12 as the enzyme:In the case of purified protein, the specific enzyme activity of PDE12, E(PDE12) could therefore be calculated using the formula:For reactions including crude protein extracts, the specific enzyme activities were calculated in mmole AMP/ (sec*g of total protein)

GENSIM SUMMARY

'The following general reaction scheme apply to the varied enzyme capabilities of the OASs: RpA + (d)NTP → PPi + RpA-(d)NMP in which the incorporation of AMP prompts for oligoadenylate synthesis by means of multiple 2′ elongation events (the other NMPs and dNMPs make up single incorporation events only).Due to low sequence specificity, RNase L degrades cellular RNA and prolonged activation results in an antiproliferative response leading to apoptosis [9] [10] [11] 16] .',

'This assay is well-suited to assess the broader cellular role expected of the OASs, based on the very diverse substrate specificity in vitro.Human OAS1 p42 containing an N-terminal His-tag (His-OAS1) cloned in the pET9d bacterial expression vector was a kind gift from Kineta (formerly Illumigen Biosciences).The plasmid was transformed into the E.coli BL21 (DE3) strain and plated on selective LB agar (50 µg/mL ampicillin), followed by inoculation of single colonies to selective LB medium for incubation at 37°C for 16 h at 200 rpm.',

'Thus the OAS1 specific enzyme activities E2(OAS1) from these reactions were calculated with the formula:The 2-5A nuclease assayThe nuclease reactions were performed in a total volume of 20 µl containing 1.2 µg of purified recombinant PDE12ΔmTP-His or 10 µg of crude protein extract together with 0.5 mM of substrate, either A(pA) 3 , A(pA) 4 or A(pA) 5 .',

'The specific activity of PDE12 (mole 5′-AMP synthesized per second per gram of protein (mole AMP/(sec*g)) was calculated with the formula:Total 254nm refers to the sum of the integral of all peaks in a given chromatogram at 254 nm for a single reaction, [A(pA) n] start is the start concentration of

BERT SUMMARY

Background 5′-triphosphorylated, 2′-5′-linked oligoadenylate polyribonucleotides (2-5As) of the structure pppA(pA) n where n ≥ 1, are the key components of the interferon (IFN)induced antiviral 2-5A system',

'The following general reaction scheme apply to the varied enzyme capabilities of the OASs: RpA + (d)NTP → PPi + RpA-(d)NMP in which the incorporation of AMP prompts for oligoadenylate synthesis by means of multiple 2′ elongation events (the other NMPs and dNMPs make up single incorporation events only).Due to low sequence specificity, RNase L degrades cellular RNA and prolonged activation results in an antiproliferative response leading to apoptosis [9] [10] [11] 16] ',

'Brown curves: Experimental salt gradients',

"We calculated the specific enzyme activity of the human OAS1 p42 for production of A(pA) 3 pdA to be 1.1 mmole/(sec*g).We also made additional set-ups and optimized the assay using alternative 'acceptor' and 'donor' molecules as substrates for the OAS (Fig",

'The position of the reactants and product are indicated in the chromatograms',

'The 2-5A core molecules are numbered according to length with 2 specifying the ApA dimer core, 3 the A(pA) 2 trimer core and so forth',

CONCLUSIONS/RECOMMENDATIONS

Based on this limited example, BERT is by far the best summaizer.

But -- BERT is slow unless you have a lot of TPUs (Tensor Processing Units)

This is for a Textrank based extractive summarizer.

Implementing this as a production ready abstractive summarizer has not been done.

Once one is built, the true power of BERT will be unleashed.

Until then, human rating is still the gold standard.

WEAKNESSES

Extremely limited data

QUESTIONS?